A Selective Attention-Based Method for Visual Pattern Recognition with Application to Handwritten Digit Recognition and Face Recognition

Albert Ali Salah, Ethem Alpaydin, and Lale Akarun

Abstract—Parallel pattern recognition requires great computational resources; it is NP-complete. From an engineering point of view it is desirable to achieve good performance with limited resources. For this purpose, we develop a serial model for visual pattern recognition based on the primate selective attention mechanism. The idea in selective attention is that not all parts of an image give us information. If we can attend only to the relevant parts, we can recognize the image more quickly and using less resources. We simulate the primitive, bottom-up attentive level of the human visual system with a saliency scheme and the more complex, top-down, temporally sequential associative level with observable Markov models. In between, there is a neural network that analyses image parts and generates posterior probabilities as observations to the Markov model. We test our model first on a handwritten numeral recognition problem and then apply it to a more complex face recognition problem. Our results indicate the promise of this approach in complicated vision applications.

Index Terms—Selective attention, Markov models, feature integration, face recognition, handwritten digit recognition.

1 Introduction

Primates solve the problem of visual object recognition and scene analysis in a serial fashion [1], which is slower but less costly than parallel recognition, which is NP-complete [2]. The idea in visual selective attention is that not all parts of an image give us information and analyzing only the relevant parts of the image in detail is sufficient for recognition and classification.

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. These sharp, directed movements of the fovea are not random. The periphery provides low-resolution information, which is processed to reveal salient points as targets for the fovea [3], and those are inspected with the fovea. The eye movements are a part of overt attention, as opposed to covert attention, which is the process of moving an attentional "spotlight" around the perceived image without moving the eye [4].

There are two broad categories of computational models of selective attention. In the first category, we find models that are motivated by the need to account for a large body of experimental data collected from neurological and psychophysical experiments. These are mainly descriptive models, with possible applications to real-world problems. The second category consists of prescriptive models that are inspired by the biological processes, but have some computational problem in focus. We place the model we propose into the latter category.

In a recent survey [5], Itti and Koch describe the essential features of computational models of selective attention. These models depend on the idea that massively parallel computation is not the most efficient way of visual recognition. The idea that a fast and broad parallel search should guide a slower and more elaborate serial search comes from Koch and Ullman [3]. The serial recognition process gathers two types of information from the image that are described by Grossberg [6] as being complementary: the contents of the fovea window and the location to which the fovea is directed. Following Ungerleider and Mishkin [7], we call these *"what"* and *"where"* information, respectively. The object is thus represented as a temporal sequence, where, at each time step, the content of the fovea window and the fovea position are observed. Usually, the parallel part is implemented with a saliency map that indicates the informative spots on the image. We use a set of problem-specific feature maps to compute a saliency map.

The saliency map is primarily bottom-up; the presence of various salient features like oriented lines, edges, and corners indicate salient spots. The selection and integration of features have been explored by Treisman and Gelade in detail [8]. The contribution of each feature is computed separately. Various ways of combining features for the saliency map is inspected in Itti and Koch [9].

An alternative approach is to use a competitive scheme, where the features compete for attention and one location emerges as the winner through inhibition. Culhane and Tsotsos [10] use a Winner-Take-All (WTA) network [3] instead of a saliency map, where a parallel pruning of the processing hierarchy cuts the computational cost down. The WTA is commonly used in competitive schemes [2], [11]. Another distributed approach to bottom-up saliency calculation without a saliency map is presented in Desimone and Duncan [12]. In their work, top-down influences promote individual feature maps.

One approach to the target selection process involves using information theoretic measures to minimize uncertainty. This involves a saliency calculation based on the expected information gain [13]. Legge et al. [14] use visual, lexical, and oculomotor information in order to guide a saccadic search that minimizes uncertainty about hypotheses in a reading task. Their computational models show some of the phenomena that can be observed in visually impaired people. Schill et al. [15] propose a scheme in which each eye movement can support one or more hypotheses about the scene. The expected information gain from a new saccade is computed from the belief distribution derived from the previous saccades.

The saccadic movements, called the scanpath by Noton and Stark [1], have been used in Didday and Arbib [16] as a distinct feature in classification. The scanpath for the memorized item is stored in memory, and compared with novel items for recognition. Hacisalihzade et al. [17] inspect how scanpath sequences can be analysed for their similarity. They use a Markov model for the sequences and inspect the sequences with the help of string editing methods. Markov models were used in a human behavior model for surveillance systems that is controlled by top-down and bottom-up attention [18]. Rimey and Brown [19] have used a left-to-right Augmented HMM to do classification based on the scanpath information. Another important study that makes use of the scanpath information is Rao et al. [20], where bottom-up and top-down influences are considered to plan saccades. In a recent study, Schill et al. [15] use the eye-movement vectors that indicate relative position changes as features in a scene analysis task.

We present a selective attention model that combines fovea contents with scanpath information for recognition. Alpaydin [21] used recurrent multilayer perceptrons to learn both the fovea features and the scanpath. In another neural network approach, Fukushima incorporated selective attention into the Neocognitron model [22]. Although good performance is reported on binary recognition tasks, his approach is sensitive to parameters, requires extensive computation, and very large networks.

In a study with aspects similar to ours, Rybak et al. [23] used neural network classifiers to recognize patterns and combined the information in a higher level. In their work, both the *"what"* and *"where"* information is stored as feature vectors in memory and combined in a behavioral recognition program. The information flow has a bottom-up characteristic. A grid of context points serves to determine the potential targets of attentional window shift. One

• *The authors are with the Perceptual Intelligence Laboratory, Department of Computer Engineering, Bogaziçi University, 80815, Bebek, Istanbul, Turkey. E-mail: {salah, alpaydin, akarun}@boun.edu.tr.*

Manuscript received 15 Nov. 2000; revised 27 Apr. 2001; accepted 19 July 2001.
Recommended for acceptance by D. Jacobs.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 113146.

0162-8828/02/$17.00 © 2002 IEEE

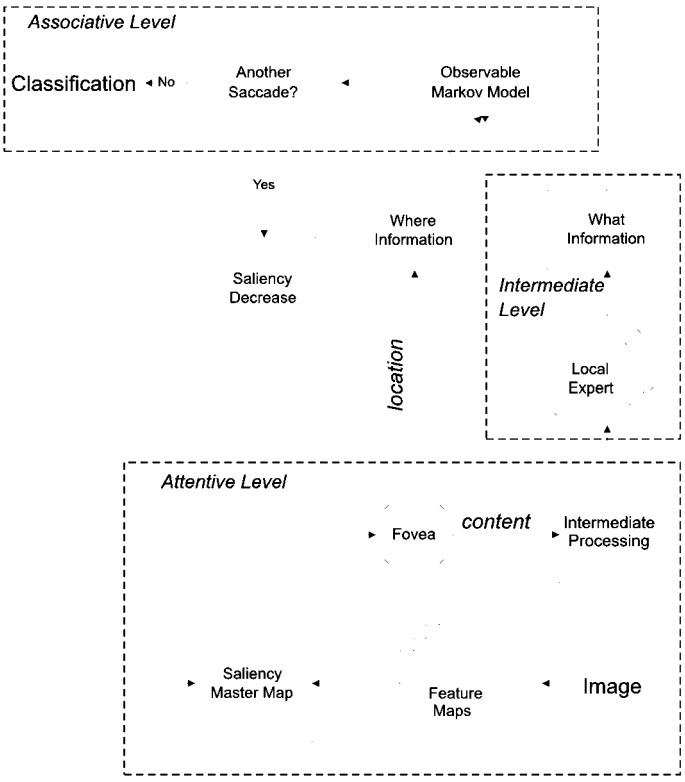

Fig. 1. The selective attention model for visual recognition.

major drawback of their model is that the scanning of the attentional window at each fixation involves a serial search among all the stored retinal images in the sensory memory. However, this increased computational cost has the benefit of allowing for shift, rotation, and scale invariant recognition of patterns. Another drawback is the use of experimentally tuned semantic significance functions calculated in advance for each image, that decrease the flexibility of the model.

Our aim is to design a scalable system based on selective attention which is applicable to problems where the input data is high-dimensional (e.g., face recognition), or not of fixed size. Implementing a parallel scheme within reasonable complexity and with good performance is not trivial in such cases.

The first task we use to test our scheme is handwritten digit recognition. In our database [24], there are 10 classes (numerals from zero to nine) with $1,934$ training and $1,797$ test cases. Each sample is a 32×32 binary image which is normalized to fit the bounding box. There are parallel architectures to solve this problem in the literature [25] and they have good performance, but they do not scale well.

To justify our approach, we test our model on a face recognition task. We use the Olivetti Research Laboratory (ORL) database [26]. There are 400 grayscale images, divided into 40 classes. A class corresponds to a person and there are 10 different images of each face. The images are 92×112 bitmaps with a plain background, positioned on the centre of the bounding box with up to 10 percent scale variation. Lightning conditions, facial expressions, and details (like glasses) are the allowed variations. The maximum tilt and rotation variation is 20 degrees.

This paper is organized as follows: In Section 2, we describe our model and its three levels. We report our simulation results on the handwritten digit recognition in Section 3. In Section 4, we comment on the application of our model to the face recognition problem. In Section 5, we present a comparative complexity analysis. In the last section, we conclude and discuss future work.

2 The Model

The block diagram of the system we propose is given in Fig. 1. It is composed of the *attentive level* that decides on where to look, the *intermediate level* that analyzes the content of the fovea, and the *associative level* that integrates saccades in time.

2.1 Attentive Level

In the first step of the model, the bottom-up part of the visual system is simulated. In constructing the saliency map, we use a simple set of features to decrease the computational cost, namely, line orientations. Line orientations are detected by different primitive mechanisms in the visual cortex, operating in coarse, intermediate and fine scales [27].

The feature maps are combined in a saliency map, which indicates the interesting spots on the image. We simulate the fovea by moving a window over the image. The saliency values of the visited spots and their periphery are decreased and these spots are not visited again. This process has a biological counterpart that is termed *the inhibition of return* [28]. The saliency decrease of the attended location is a commonly used method in computational models of selective attention [5].

2.2 Intermediate Level

The simulation of saccades should provide us with *"what"* and *"where"* information, but we want them to be sufficiently quantized to be used in the associative level. We divide the image space into uniform regions, in effect, performing a quantization on the location information. We obtain a time-ordered sequence of visited regions after the simulation of saccades. This constitutes the *"where"* stream for the particular sample.

In order to efficiently analyze the content of the fovea, we train neural network experts at each region of the image. The experts are single-layer perceptrons that are trained in a supervised manner. Their input is the fovea content vector. The output of the experts are class posterior probability vectors, which are treated as the *"what"* information stream.

We select single-layer perceptrons over multilayer perceptrons for a number of reasons. Multilayer perceptrons overlearn the training data quickly and perform worse on the cross-validation set. The number of parameters we need to store for the multilayer perceptron is larger and the training time is significantly longer. These properties make the single-layer perceptron the better choice of expert in the final model.

2.3 Associative Level

In the associative level, the two types of quantized information are combined with a discrete, observable Markov model (OMM). We treat the regions visited by the fovea as the states of a Markov model and the quantized output of the local neural network experts as the observations from each state. We simulate a number of saccades for each sample in the training set, obtain the *"where"* and *"what"* streams, and adjust the probabilities of the single Markov chain of the corresponding class to maximize the likelihood of the training data. The number of saccades depends on the application. Recognition is achieved by selecting the class that maximizes the posterior probability.

Training an observable Markov model is much faster than training a Hidden Markov Model [29]. In the observable model, the model parameters are directly observed from the data. Since we know the states, we can count the state transitions and normalize the count to find the state transition probabilities a_{ij}, as well as the initial state distribution probabilities π_i. Similarly, we count the occurrences of the observation symbols (quantized outputs of the local neural networks) at each state and normalize them to find the observation symbol probability distribution $b_j(k)$. Finding the probability of the observation sequence is much simpler in the observable Markov model since the states are visible. We just multiply the corresponding state transition probabilities and the observation probabilities:

$$P(O, S|\lambda) = \pi_{S_1} b_{S_1}(O_1) \prod_{i=2}^{n} a_{S_{i-1}S_i} b_{S_i}(O_i), \tag{1}$$

where S is the state sequence, O is the observation sequence, and $\lambda = \{\pi_i, a_{ij}, b_j(k)\}$ stands for the parameters of the Markov model. $i, j = 1 \ldots N$ are indices for states, $k = 1 \ldots M$ is the index for the observation symbols. There are as many Markov models as there are classes and, in classification, we choose the class C_c that has the highest observation probability:

$$P(O, S|\lambda_c) = \max_j P(O, S|\lambda_j). \tag{2}$$

2.4 Dynamic Fovea

One important advantage of using a Markov model is the ease with which we can control the number of saccades necessary for recognition. In the training period, our model simulates a predetermined number of saccades, which represents an upper bound for the particular application. After each saccade, the Markov model has enough information to give a posterior probability for each class. We may calculate the probability $\alpha_t(c)$ of the partial sequence in the Markov model, which reflects the probability of the sample belonging to a particular class, given the *"where"* and *"what"* information observed so far. Using (1), we have

$$\alpha_t(c) = P(O_1, \ldots, O_t, S_1, \ldots, S_t|\lambda_c), \tag{3}$$

where $O_1, \ldots, O_t$ is the observation sequence up to time t, $S_1, \ldots, S_t$ is the state sequence, and λ_c are the parameters of the Markov model for class C_c with $c = 1 \ldots K$.

We can use this probability to stop our saccades whenever we reach a sufficient level of confidence in our decision. Let us define $\alpha_t^*(c)$, the posterior probability for class C_c at time t:

$$P(C_c|O_1, \ldots, O_t, S_1, \ldots, S_t) \equiv \alpha_t^*(c) = \frac{\alpha_t(c)}{\sum_{j=1}^{K} \alpha_t(j)}. \tag{4}$$

Let τ be the threshold we use as our stopping criterion:

$$\alpha_t^*(c) \geq \tau, \tag{5}$$

where the value of τ is in the range [0,1]. If we assume that absolute certainty is not reached anywhere in the model and $\alpha_t^*(i)$ is always below 1.0, selecting $\tau = 1.0$ is equivalent to treating all samples as equally difficult and looking at all possible locations. Conversely, selecting $\tau = 0$ is equivalent to looking at the first salient spot and classifying the sample. Selecting a large value for τ trades speed for accuracy. With a well selected value, we devote more time for difficult samples, but recognize a trivial sample in a few saccades.

3 Handwritten Digit Recognition

3.1 Methodology

The digits are 32×32 binary images, but we work on 12×12 downsampled images to simulate a low-resolution resource. This slightly decreases the classification accuracy, but speeds up the computation considerably. Convolving the digit image with 3×3 line orientation kernels produces four line orientation maps in 0°, 45°, 90°, and 135° angles. With a Gaussian low-pass filter, these are combined in a saliency master map [30], which indicates the presence of aligned lines on the image.

Our experiments showed that adding other feature detectors like corner maps, Canny edge detector, and further line orientation maps in higher resolutions increased the classification accuracy only slightly (i.e., less than one percent), whereas the increase in the computational cost was significant. This results from the simple nature of the problem, more complex tasks like scene analysis would require more complex bottom-up feature detectors [31], [32].

We use a 4×4 fovea window over the 12×12 downsampled image. We divide the image into nine regions for the experts. Since the images are small, we use a second set of overlapping windows to reduce the effect of window boundaries. We obtain a time-ordered sequence of visited regions after the simulation of saccades. This constitutes the *"where"* stream for the particular sample.

As fovea contents, we extract 64-dimensional real-valued vectors. These vectors are produced by concatenating the corresponding 4×4 windows on four oriented line maps. We prefer using the concatenated line maps to inspecting the original bitmap image, because the line maps indicate the presence of features more precisely. Furthermore, since they were constructed in the attentive level, they come at no additional cost but because of their high dimension, they cannot be used directly in the Markov model.

In order to efficiently quantize this information, we train neural network experts at each region of the image. The experts are single-layer perceptrons that are trained in a supervised manner. Their input is the 64-dimensional fovea content vector. The local experts provide the associative level with 10-dimensional class posterior probabilities. Since we need a small and discrete number of observation symbols for the Markov model, we use k-means clustering and, from the 10-dimensional vector, obtain a single observation value between 1 and 30 corresponding to the fovea content.

We simulate eight saccades to train the system. The Markov model is trained with a limited training set and, if the number of states and observation symbols is large, there will be connections that are not visited at all. Since the model is probabilistic, having a transition or observation probability of zero is not desired. Instead, the transitions that have not occurred within the training set are set to a low probability (0.0001) in the model. Then, we normalize the probabilities once more.

We have also tried fully connected, discrete Hidden Markov Models where the states are not visible and where the concatenated *where-what* information is the observation, but this structure performed worse than the observable Markov model because of the larger number of free parameters.

TABLE 1
Handwritten Digit Recognition Results

Method	Accuracy	
	Training	Test
Voting over regions	93.85	73.89
Markov model (OMM)	94.37	87.37
Dynamic fovea	91.41	84.63
All-parallel MLP	99.92	94.54

3.2 Simulation Results

We summarize the results we obtain in Table 1. The first column of the table shows the method employed. The successive columns indicate the classification accuracy on the training and test sets.

In voting, we do eight saccades, generate the posterior probabilities of classes by the local neural network experts, and take a vote without treating them as a sequence. Comparing this result with the OMM result where eight saccades are done, shows that the order information which is lost during voting but used in OMM is useful. The dynamic fovea, where the number of saccades is not fixed but depends on the certainty of the Markov model (4) and (5), has a lower classification accuracy, but it only needs 3.2 saccades on the average, instead of the previous eight. Finally, the last row indicates the accuracy of an all-parallel scheme, where we use a multilayer perceptron (MLP) with 32×32 binary input and 10 hidden units. Although the MLP has good accuracy in this problem, it is not scalable due to the curse of dimensionality.

When we simulate the dynamic fovea with a fixed threshold of $\tau = 0.95$, the classification accuracy on the test set is 84.63 percent, and the average number of saccades is 3.37, which corresponds roughly to seeing one third of the image in detail. This justifies our claim that analysing only a small part of the image is enough to recognize it.

As the threshold is increased, the accuracy of classification increases because a higher threshold means making more saccades to get a more confident answer. A lower threshold means that a quick response is accepted. What happens is that the average number of saccades increase sharply if the threshold is set to a value very close to 1.0. In this case, the classifier cannot exceed the threshold probability with eight saccades and selects the highest probability class without doing any more saccades.

4 Face Recognition

4.1 Methodology

Applying the model to face recognition necessitates a number of adaptations related to the size of the problem. We can summarize these changes as the selection of new feature maps, a bigger fovea window, more local experts, and the elimination of the k-means clustering from the intermediate level.

In the attentive level, we use Gabor wavelet filters as our feature maps [33]. These are frequently used in face recognition tasks [34], [35] and they resemble the biological Impulse Response Functions [9]. We use filters with spatial frequency $k_v = \frac{\pi}{8}$ in four angle orientations (0, $\frac{\pi}{4}$, $\frac{\pi}{2}$, and $\frac{3\pi}{4}$) to generate a 32×32 saliency master map. We have trained classifiers for a range of frequency and filter size parameters, and selected the most informative combination.

We use the saliency map to select locations, and sample a 10×10 fovea window from each location. As in the previous section, we concatenate the feature map contents to cut down the computational cost of the process. Since the images are larger, we use 36 local neural network experts on a 6×6 grid. Single-layer perceptrons are used as experts.

The input vectors fed to the experts are Z-normalized by subtracting the mean and dividing by the standard deviation, which are obtained during the training phase. Since the number of classes (people) is large, we have eliminated the k-means clustering stage in the intermediate level, and calculated an observation symbol probability distribution at each state by summing up and normalizing the corresponding local perceptron outputs for each class.

4.2 Simulation Results

We present our results in Table 2. Since we had a restricted data set, all the results are obtained by 10-fold cross-validation. The training set is classified with perfect accuracy. The first row is the accuracy of the model after the maximum number of saccades were performed. The performance of our model on the face recognition data is on the average 92 percent, which shows that the model scales to this problem. Our results are comparable to the results previously published on the ORL database. Turk and Pentland [36] report 90 percent accuracy with the eigenface method they employ. A higher classification accuracy was reported by Lawrence et al. [37]. They use convolutional neural networks and obtain 96.2 percent accuracy.

The second row of Table 2 is the dynamic fovea simulation, where an average of 4.81 (± 1.71) saccades were made for the test samples, as opposed to 15 saccades in the previous case. We have selected $\tau = 0.999$ in the dynamic fovea simulation. Note that there are 36 possible locations and doing 4.81 saccades corresponds to making a decision after analysing less than one-seventh of the image. Accuracy increases by increasing the confidence threshold and making more saccades.

In the MLP scheme (Row 3 of Table 2), we concatenate four 64×64 vectors obtained by Gabor filters with the same frequency and four directions. Thus, the system is trained with $16,384$ input dimensions and 150 hidden units were used.

5 Time and Space Complexity Analysis

We compare the time and space complexity of our model with the all-parallel scheme of a multilayer perceptron. The processing in our model starts with creating the saliency map. After the

TABLE 2
Face Recognition Results

Method	Accuracy on Test Set	Time Requirement	Space Requirement
OMM	92.00($\pm$ 7.98)	102,400+240,000+21,600=364,000	576,000+110,880=686,880
Dynamic fovea	88.50($\pm$ 12.37)	102,400+76,960+6,926=186,286	576,000+110,880=686,880
MLP	88.25($\pm$ 7.27)	409,600+2,463,600=2,873,200	2,463,600

determination of salient locations, the fovea contents are processed by the local expert and the expert output is used to generate a partial observation sequence probability for each class using the OMM. If the selected confidence level is not reached, the process is repeated. Assuming that on the average t^* fovea shifts are made, with a fovea size of v^2 and the saliency map size D^2, the time complexity is

$$\underbrace{D^2 \times f \times G^2}_{saliency\ map} + \underbrace{t^* \times v^2 \times C}_{local\ experts} + \underbrace{t^* \times C \times M}_{OMM}, \quad (6)$$

where f denotes the number of features used and G^2 denotes the Gabor filter size. The time complexity of the MLP is

$$\underbrace{I \times g \times G^2}_{Gabor\ preprocessing} + \underbrace{(I \times g) + C) \times h}_{perceptron}, \quad (7)$$

where I $(\text{width} \times \text{height})$ is the image size, g is the number of orientations used, and h is the number of hidden units.

The space complexity of our model is the combined parameters stored for the local expert weights and the parameters of the OMM:

$$\underbrace{M \times v^2 \times C}_{local\ experts} + \underbrace{C \times M \times (M + N + 1)}_{OMM}, \quad (8)$$

whereas the space complexity of the MLP is

$$((I \times g) + C) \times h. \quad (9)$$

We typically expect the number of input dimensions to the MLP to be much greater than the local experts we use. Table 2 gives the calculated values for the face recognition problem where we see that, for comparable accuracy, our proposed method has a space complexity $\frac{1}{4}$ and time complexity $\frac{1}{15}$ of the all-parallel multilayer perceptron.

6 Discussion and Future Work

The selective attention mechanism exploits the fact that real images often contain vast areas of data that are insignificant from the perspective of recognition. A low-resolution, downsampled 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. By keeping the parallel part of the method simple, we can speed-up the recognition process considerably.

The first application we chose for our system, namely, digit classification, does not show the full benefits of the model since the ratio of the fovea area to the image is not high enough. Although the accuracy is lower than the state-of-the-art parallel approaches in the literature (e.g., the MLP result in Table 1), the selective attention mechanism is much more appropriate for applications where parallel processing is too cumbersome to use and the number of input dimensions is high.

To justify our approach, we apply our model to a face recognition problem and show that the resulting model is more efficient compared to an all-parallel classifier. In a face, there is much more redundancy and small regions of the face like eyes, nose, and mouth give us information. The saliency scheme is modified for this purpose, as facial features necessitate different and more complex feature detectors. Details of our simulation on the two applications can be found in [38]. Our proposed method based on selective attention is as accurate as the all-parallel multilayer perceptron, but is four times simpler and 15 times faster.

The associative level in our model allows the calculation of an expected information gain [15] for future saccades that can be incorporated into the saliency map as a top-down influence. We consider this as an interesting future direction for top-down and bottom-up integration in our model. We are also considering to apply our model in a real-time scene analysis application.

Acknowledgments

The authors thank Berk Gökberk for his help on the face recognition application. They would also like to thank the editor and reviewers for constructive comments, which greatly improved the content and the presentation. A preliminary version of this work is presented at the 23rd Annual Conference of the Cognitive Science Society, August 2001, Edinburgh. This work is supported by Bouaziçi University Research Funds 00A101D. E. Alpaydin is a Distinguished Young Scientist of the Turkish Academy of Sciences, supported under the TÜBA/GEBIP program.

References

[1] D. Noton and L. Stark, "Eye Movements and Visual Perception," *Scientific Am.*, vol. 224, pp. 34-43, 1971.
[2] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling Visual Attention via Selective Tuning," *Artificial Intelligence*, vol. 78, pp. 507-545, 1995.
[3] C. Koch and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," *Human Neurobiology*, vol. 4, pp. 219-227, 1985.
[4] F. Crick and C. Koch, "Towards a Neurobiological Theory of Consciousness," *Seminars in the Neurosciences*, vol. 2, pp. 263-275, 1990.
[5] L. Itti and C. Koch, "Computational Modeling of Visual Attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194-203, Mar. 2001.
[6] S. Grossberg, "The Complementary Brain: Unifying Brain Dynamics and Modularity," *Trends in Cognitive Sciences*, vol. 4, pp. 233-246, 2000.
[7] L.G. Ungerleider and M. Mishkin, "Two Cortical Visual Systems," *Analysis of Visual Behaviour*, D.J. Ingle, M.A. Goodale and R.J.W. Mansfield eds., 1982.
[8] A.M. Treisman and G. Gelade, "A Feature Integration Theory of Attention," *Cognitive Pscychology*, vol. 12, no. 1, pp. 97-136, Jan. 1980.
[9] L. Itti and C. Koch, "Feature Combination Strategies for Saliency-Based Visual Attention Systems," *J. Electronic Imaging*, vol. 10, pp. 161-169, 2001.
[10] S.M. Culhane and J.K. Tsotsos, "A Prototype for Data-Driven Visual Attention," *Proc. 11th Int'l Conf. Pattern Recognition*, vol. 1, pp. 36-40, 1992.
[11] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, Nov. 1998.
[12] R. Desimone and J. Duncan, "Neural Mechanisms of Selective Visual Attention," *Ann. Rev. of Neuroscience*, vol. 18, pp. 193-222, 1995.
[13] M. Jägersand, "Saliency Maps and Attention Selection in Scale and Spatial Coordinates: An Information Thoretic Approach," *Proc. Fifth Int'l Conf. Computer Vision*, pp. 195-202, 1995.
[14] G.E. Legge, T.S. Klitz, and B.S. Tjan, "Mr. Chips: An Ideal-Observer Model of Reading," *Psychological Rev.*, vol. 104, no. 3, pp. 524-553, 1997.
[15] K. Schill, E. Umkehrer, S. Beinlich, G. Krieger, and C. Zetzsche, "Scene Analysis with Saccadic Eye Movements: Top-Down and Bottom-Up Modeling," *J. Electronic Imaging*, vol. 10, no. 1, pp. 152-160, Jan. 2001.
[16] R.L. Didday and M.A. Arbib, "Eye Movements and Visual Perception: A 'Two Visual System' Model," *Int'l J. Man-Machine Studies*, vol. 7, pp. 547-569, 1975.
[17] S.S. Hacisalihzade, L.W. Stark, and J.S. Allen, "Visual Perception and Sequences of Eye Movement Fixations: A Stochastic Modeling Approach," *IEEE Trans. System, Man, and Cybernetics*, vol. 22, no. 3, pp. 474-481, 1992.
[18] N. Oliver, B. Rosario, and A. Pentland, "A Bayesian Computer Vision System for Modeling Human Interactions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 2, no. 8, pp. 831-843, Aug. 2000.
[19] R.D. Rimey and C.M. Brown, "Selective Attention as Sequential Behavior: Modeling Eye Movements with an Augmented Hidden Markov Model," TR-327, Computer Science, Univ. of Rochester, Feb. 1990.
[20] R.P. Rao, G.J. Zelinsky, M.M. Hayhoe, and D.H. Ballard, "Eye Movements in Visual Cognition: A Computational Study," Technical Report, 97.1, Univ. of Rochester, Computer Science Dept., 1997.
[21] E. Alpaydin, "Selective Attention for Handwritten Digit Recognition," *Advances in Neural Information Processing Systems 8*, D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo eds., pp. 771-777, 1996.
[22] K. Fukushima, "Neural Network Model for Selective Attention in Visual Pattern Recognition and Associative Recall," *Applied Optics*, vol. 26, no. 23, pp. 4985-4992, Dec. 1987.
[23] I.A. Rybak, V.I. Gusakova, A.V. Golovan, L.N. Podladchikova, and N.A. Shevtsova, "A Model of Attention-Guided Visual Perception and Recognition," *Vision Research*, vol. 38, pp. 2387-2400, 1998.
[24] C.L. Blake and C.J. Mertz, *UCI Repository of Machine Learning Databases*, Univ. of California, Irine, http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.

[25] Y. Le Cun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation*, vol. 1, no. 4, pp. 541-551, 1989.
[26] *The Olivetti Research Laboratory Database of Faces*, http://www.cam-orl.co.uk/facedatabase.html. 1994.
[27] D.H. Foster and S. Westland, "Multiple Groups of Orientation-Selective Visual Mechanisms Underlying Rapid Oriented-line Detection," *Proc. Royal Soc. London*, vol. 265, pp. 1605-1613, 1998.
[28] R.M. Klein, "Inhibition of Return," *Trends in Cognitive Sciences*, vol. 4, no. 4, pp. 138-147, Apr. 2000.
[29] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition" *Proc. IEEE*, vol. 17, no. 2, Feb. 1989.
[30] L. Itti and C. Koch, "A Comparison of Feature Combination Staregies for Saliency-Based Visual Attention Systems," *SPIE Human Vision and Electronic Imaging IV*, vol. 3644, pp. 373-382, Jan. 1999.
[31] D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Context-Free Attentional Operators: The Generalized Symmetry Transform," *Int'l J. Computer Vision*, vol. 14, pp. 119-130, 1995.
[32] C.M. Privitera and L.W. Stark, "Algorithms for Defining Visual Regions-of-Interest: Comparison with Eye Fixations," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 9, pp. 970-982, Sept. 2000.
[33] J.G. Daugman, "Complete Discrete 2D Gabor Transform by Neural Networks for Image Analysis and Compression," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 36, pp. 1169-1179, July 1988.
[34] F. Smeraldi and J. Bigün, "Facial Feature Detection by Saccadic Exploration of the Gabor Decomposition," *Proc. Int'l Conf. Image Processing*, vol. 3, pp. 163-167, 1998.
[35] J.G. Keller, S.K. Rogers, M. Kabrisky, and M.E. Oxley, "Object Recognition Based on Human Saccadic Behaviour," *Pattern Analysis & Applications*, vol. 2, pp. 251-263, 1999.
[36] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Science*, vol. 3, no. 1, pp. 71-96, 1991.
[37] S. Lawrence, C.L. Giles, A.C. Tsoi, and A.D. Back, "Face Recognition: A Convolutional Neural-Network Approach," *IEEE Trans. Neural Networks*, vol. 8, pp. 98-113, 1997.
[38] A.A. Salah, E. Alpaydin, and L. Akarun, "Selective Attention Based Visual Pattern Recognition," Technical Report, FBE/CMPE-03/2001-12, Bogaziçi Univ., Dept. of Computer Eng., http://www.cmpe.boun.edu.tr/~salah/tr_salah.zip, 2001

▷ **For more information on this or any other computing topic, please visit our Digital Library at** http://computer.org/publications/dlib.

Real-Time Epipolar Geometry Estimation of Binocular Stereo Heads

Mårten Björkman and Jan-Olof Eklundh, *Member*, *IEEE Computer Society*

Abstract—Stereo is an important cue for visually guided robots. While moving around in the world, such a robot can use dynamic fixation to overcome limitations in image resolution and field of view. In this paper, a binocular stereo system capable of dynamic fixation is presented. The external calibration is performed continuously taking temporal consistency into consideration, greatly simplifying the process. The essential matrix, which is estimated in real-time, is used to describe the epipolar geometry. It will be shown, how outliers can be identified and excluded from the calculations. An iterative approach based on a differential model of the optical flow, commonly used in structure from motion, is also presented and tested towards the essential matrix. The iterative method will be shown to be superior in terms of both computational speed and robustness, when the vergence angles are less than about 15°. For larger angles, the differential model is insufficient and the essential matrix is preferably used instead.

Index Terms—Epipolar geometry, active vision, real-time stereo, dynamic vergence.

1 Background

STEREO vision can be used for several tasks in the area of robotics. For applications, such as visually guided manipulation, stereo is a valuable cue in order to recognize shapes and pose. Stereopsis has also been used for visual navigation and obstacle detection [11], [18]. Such systems typically consist of two or more small cameras located in parallel with relatively short baselines, since they are expected to work for various distances. Objects located close to the observer might otherwise only be seen by one of cameras, greatly complicating depth estimation and making the system useless for behaviors such as obstacle avoidance. However, for manipulation, accuracy is crucial and a wider baseline is desired.

This discussion leads us to believe that a mobile platform would benefit from an active stereo vision system with dynamic vergence. Based on the activity at hand, an active system can control its gaze in such a way that the visual system is always used at its full potential. Ballard [1] gave a number of reasons how active vision can be used to overcome limitations in resolution and field of view, making the visual computations much less expensive than that of a passive system. Using fixation, many visual operations, such as visual navigation, 3D motion estimation, and figure-ground segmentation can further be simplified. Fermüller and Aloimonos [6] used normal-flow in conjunction with fixation to solve problems such as ego-motion recovery, 3D motion, and time-to-impact estimation. Daniilidis [5] decoupled the optical flow and found the translation of an observer through two one-dimensional searches, radically simplifying the ego-motion estimation process.

One of the reasons why dynamic vergence has seldom been used in practical robotics is the problem of external calibration. The relative orientation of the cameras has to be known, in order for the

• *The authors are with the Computational Vision and Active Perception Laboratory (CVAP), Department of Numerical Analysis and Computing Science, Royal Institute of Technology (KTH), S-100 44 Stockholm, Sweden. Email: {celle, joe}@nada.kth.se.*

Manuscript received 03 May 2000; revised 04 Jan. 2001; accepted 24 Feb. 2001.
Recommended for acceptance by A. Shashua.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 112044.

0162-8828/02/$17.00 © 2002 IEEE