ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

A sequence-encoding neural network for face recognition

Marek Barwiński[1,2] and Rolf P. Würtz[1,2]

1- Ruhr University Bochum – International Graduate School of Neuroscience
Universitätstrasse 150, D-44801 Bochum, Germany

2- Ruhr University Bochum – Institut für Neuroinformatik
Universitätstrasse 150, D-44801 Bochum, Germany

Abstract. We propose a feature-based system for face recognition using contextual information to improve the recognition rate. A small (6 memory blocks, 3 cells each) recurrent neural network with internal memory cell states (LSTM) is trained on single images of 49 different identities randomly picked from the FERET database and tested on images with different facial expressions using a predefined saccade path. We show that the system presents an improvement of recognition rate and an outlook to the future development of the system including autonomous saccade generation, evidence accumulation and novelty detection.

1 Introduction

Recurrent neural networks Recurrent neural architecture facilitates operating on a sequential data stream, binding information distant in time. A recall of sequences of almost arbitrary length is realized through a cascade mechanism, where each step in the sequence facilitates the recall of the consecutive portion of information. This mechanism is present in various forms of perception – music encoding and decoding, manual skills, following a remembered path, etc. Prediction of perception in the immediate future is matched against actual experience. Such a recall mechanism must have its appropriate encoding counterpart. See Jensen and Lisman [3] for a model of sequential recall.

Context dependency Both feature and configurational information play a role in face processing. Recognition of a face is context dependent, as similar features in several distinct identities do not lead to multiple identity recognition, but rather a request of the neural system for additional data to resolve the ambiguity. Successful technical methods of combining configurational and feature information include Elastic Graph Matching [5] and Gabor pyramid matching [15], which take into account spatial relations between features. A psychophysically convincing strategy of encoding and recall of these spatial relations should take into account a multi-saccade approach.

Facial properties Facial features are not restricted to semantically defined ones like eyes, brows and mouth, which have a name in common language. In the language of visual analysis, or the code of the primary visual cortex, an equally

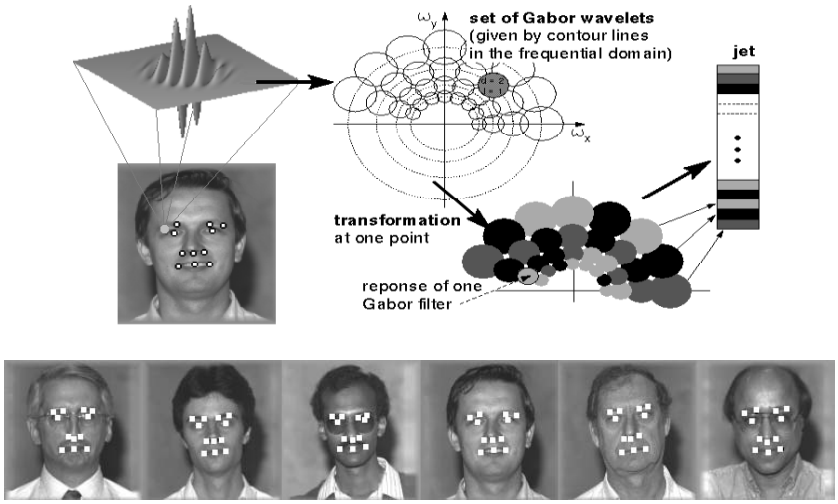

Fig. 1: Creating a multidimensional Gabor filter response set called a jet (top) on manually labeled fixation points (bottom).

important distinction between features may be based on different responses to filters of different spatial frequency or preferred orientation. A set of forty Gabor functions – eight different orientations and five different spatial frequencies – called a jet will be the filters of choice.

Involuntary visual stimuli sequencing Experiments by Thorpe and van Rullen [13, 14] have pointed to a natural mechanism of visual input sequencing. Based on the contrast of particular features and top-down attention facilitating the response of particular neurons, the brain automatically adapts to process the stimuli it considers important more quickly than others. This sequencing begins already in LGN and goes on further as the stream of spikes propagates into higher cortical levels. This is a natural sequencing mechanism within a single fixation.

2 Methods

Manual labeling and data preprocessing 49 identities with two facial expression (neutral – non neutral) were chosen randomly from the FERET database [7]. For a successful method of normalizing facial expressions see [11],[12]. Twelve fiducial points were labeled manually as points of fixation, and the response vectors of 40 Gabor filters have been extracted at these points. Manually selected points are known to be areas carrying a large amount of information in inner facial features. These jet values are normalized absolute values of complex numbers representing a scalar product of a complex Gabor function with the image. This is a model for V1 complex cell responses [4, 8].

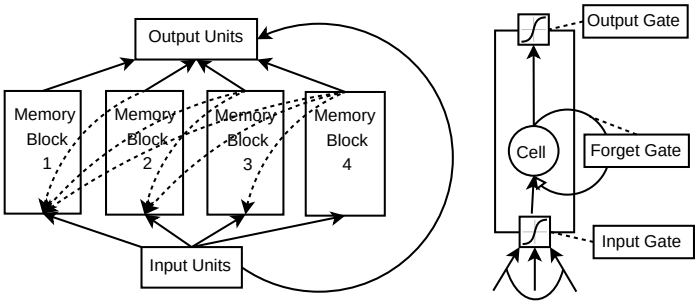

Fig. 2: In LSTM each memory block output connects to all previous memory block inputs. Each memory block has one or more cells and three multiplicative gates, whose activity depend on input. Forget gate sustains or resets cell state. Input and output gates modify the strength of information flowing in and out of a memory block.

LSTM The "Long Short Term Memory" (LSTM) recurrent neural network architecture [2, 1] is a successful tool for sequential data analysis. Neurons in this network have an internal state, which serves as a short term memory. Equipped with powerful multiplicative gates it shields the neuron states from unwanted input, when it is clear that in this context the input data is not relevant. Forget gates allow for a quick reset of the cell state when new data needs to be memorized. Long term memory is coded within synaptic weights. The sequential architecture of memory blocks allows the network to switch on consecutive blocks as time goes by, allowing for information back propagation and using the full capacity of the network only for long lags between important cues.

An LSTM network with forget gates [1] was employed to process this data. 6 memory blocks containing 3 cells each were chosen, the bias weights for consecutive memory blocks were set to $0.3 \cdot \beta$, where β is the memory block index, and the learning rate α was set to 0.25.

The network's input layer consists of $n+m+3$ input units and $n+m$ output units, where n represents the size of the jet and m the number of identities. The input vector is enhanced by two units coding for start and end of a sequence, as well as a bias neuron, always set to 1 to adapt the synaptic weights to the mean of the input data. Identity coding m output units serve as input units as well. In this way the network sustains its decision regarding identity unless further fixations prove it wrong. The network is asked not only to predict one of m identities but also the most probable jet in the next fixation (this will serve as input for saccade generation in a future version). The memory blocks have been connected in a unidirectional manner, thus serving as a kind of stack for jets encountered at earlier fixation points.

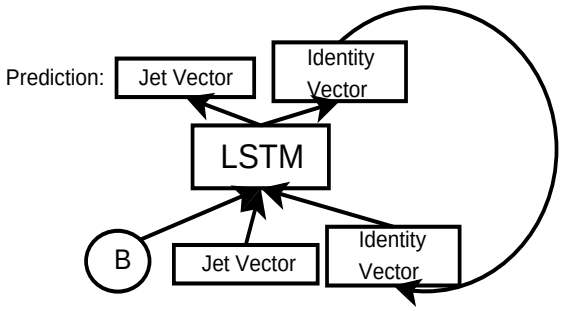

Fig. 3: Modified LSTM architecture with a bias neuron clamped to 1, and the recurrently connected part of an output layer.

Training and recall During training the jets of each face are presented in a fixed order of the saccade points, while one identity cell is clamped to one and the others to zero. In the recall phase, the first jet is presented with equal activity on the identity cells, and their activity evolves over the saccade path. The identity with the maximal response after the last fixation is taken as the recognized one.

3 Results

Using only twelve fiducial points and a small neural architecture we could improve the recognition rate using context dependent decision making (see fig. 4). The simple identity prediction feedback is the first step in improving the overall abilities of the system. This already allows for a very good prediction of the next, expected jet. The mean scalar product of expected and encountered jets is $\mu = 0.93 \pm 0.06$, while a scalar product of two random jets reaches $\mu = 0.78 \pm 0.07$. We have shown that even without extensive parameter tuning the system is capable of properly recognizing identity based on a short sequence of partly ambiguous input.

4 Discussion and outlook

Evidence accumulation Despite the improvement in recognition rate the presented system is only a first step, and the most important improvement will be a a more robust evidence accumulation mechanism. Simple identity feedback is not enough to allow for more cautious decisions by the network. The strong influence of the jet part of the input layer forces network to make fast and sometimes incorrect decisions about identity, which are then forced back to the input layer.

Sequential memory for face encoding and recall Our results show a large potential of the LSTM architecture for processing floating point data of

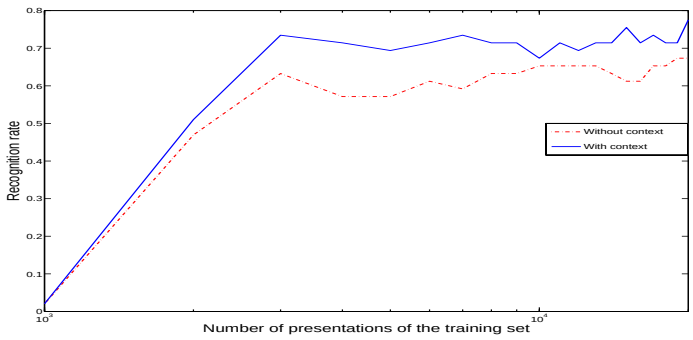

Fig. 4: The architecture including the context of the presented jet performs better than the one using merely current jet information. Recognition rate increases relatively by $12\% \pm 3\%(0.08 \pm 0.02)$.

large dimension. The network is flexible in adapting to newly acquired data and there are no theoretical obstacles in allowing the growth of this architecture. We plan to develop a system which will autonomously decide on the saccade path based on the quality of information it has gathered so far. The consecutive fixations may be related to anything from a different area of the face to a different spatial frequency of interest. This will require an "artificial central executive" to calculate the current state and probable point of interest in the next step communicating with the "artificial saccade generator".

Bayesian saccade generator Such a manually chosen saccade path is definitely not stable enough to be applied to general problems of face recognition. The results of the proposed method will however provide useful cues to limitations and possibilities of a system which autonomously chooses fixation points. The sequential memorization needs to be successful for a fixed path in order to withstand the challenge of independently chosen saccades. Human saccade behavior during search in the visual scene follows a model of an ideal Bayesian observer. Evolved strategy make much use from processing information during a single fixation rather than integration across fixations. Using a large retinal view and Bayesian priors, a system is capable of making a correct saccade decision based on low resolution marginal information [6].

Novelty detection An important goal for further development is automatic novelty detection based on information about the current fixation combined with prediction about the next location and what is expected there. Violations of predictions are relevant cues for a decision whether unpredicted encountered data implies a new identity or merely a fluctuation in volatile facial features. An ideal system would recognize a new face from the old (a familiarity signal), properly recognize a difference in a single feature e.g. new haircut, facial

hair, glasses from a completely new identity. Theory of novelty detection using information theory is gaining more support and acceptation as a predictor of hippocampal activity signaling human experienced novelty [10]. Mechanisms of cortical inference using Bayesian statistics and the role of neuromodulators are described in [9, 16].

Summary We have presented first results of a neural network to recognize faces as sequences of V1 response vectors over a saccade path. The context introduced by the sequencing could improve the recognition rate significantly over decisions based on a single jet. This shows that the data format is suitable for further development of a neuronally realistic recognition model.

References

[1] F. Gers. *Long Short-Term Memory in Recurrent Neural Networks*. PhD thesis, Département d'informatique, École Polytechnique Fédérale de Lausanne, 2001.

[2] S. Hochreiter and J. Schmidhuber. Long short term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[3] O. Jensen and J. E. Lisman. Hippocampal sequence-encoding driven by a cortical multi-item working memory buffer. *Trends in Cognitive Neuroscience*, 28(2):67–72, 2005.

[4] J. Jones and L. Palmer. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258, 1987.

[5] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993.

[6] J. Najemnik and W. S. Geisler. Optimal eye movement strategies in visual search. *Nature*, 434:387–341, 2004.

[7] P. J. Philips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[8] D. A. Pollen and S. F. Ronner. Spatial computation performed by simple and complex cells in the visual cortex of the cat. *Vision Research*, 22:101–118, 1982.

[9] R. P. Rao. Bayesian inference and attentional modulation in the visual cortex. *NeuroReport*, 16(16):1843–1848, 2005.

[10] B. A. Strange, A. Duggins, W. Pennya, R. J. Dolana, and K. J. Friston. Information theory, novelty and hippocampal responses - unpredicted or unpredictable? *Neural Networks*, 18:225–230, 2005.

[11] A. Tewes. *A Flexible Object Model for Encoding and Matching Human Faces*. PhD thesis, Physics Dept., Univ. of Bochum, Germany, Jan. 2006.

[12] A. Tewes, R. P. Würtz, and C. von der Malsburg. A flexible object model for recognising and synthesising facial expressions. In T. Kanade, N. Ratha, and A. Jain, editors, *Proceedings of the International Conference on Audio- and Video-based Biometric Person Authentication*, LNCS, pages 81–90. Springer, 2005.

[13] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, 1996.

[14] R. Van Rullen and S. J. Thorpe. Surfing a spike wave down the ventral stream. *Vision Research*, 42(23):2593–2615, 2002.

[15] R. P. Würtz. Object recognition robust under translations, deformations and changes in background. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):769–775, 1997.

[16] A. J. Yu and P. Dayan. Uncertainty, neuromodulation, and attention. *Neuron*, 46:681–692, 2005.