Neural Process Lett
DOI 10.1007/s11063-008-9084-y

Robust Handwritten Character Recognition with Features Inspired by Visual Ventral Stream

Ali Borji · Mandana Hamidi · Fariborz Mahmoudi

Abstract This paper focuses on the applicability of the features inspired by the visual ventral stream for handwritten character recognition. A set of scale and translation invariant C2 features are first extracted from all images in the dataset. Three standard classifiers kNN, ANN and SVM are then trained over a training set and then compared over a separate test set. In order to achieve higher recognition rate, a two stage classifier was designed with different preprocessing in the second stage. Experiments performed to validate the method on the well-known MNIST database, standard Farsi digits and characters, exhibit high recognition rates and compete with some of the best existing approaches. Moreover an analysis is conducted to evaluate the robustness of this approach to orientation, scale and translation distortions.

Keywords Optical character recognition · Handwritten character recognition · Visual system · Visual ventral stream · HMAX · C2 features

1 Introduction

Handwritten character recognition is still a challenging problem for many languages like Farsi, Chinese, English, etc. Developing robust optical character recognition (OCR) techniques would be very rewarding in today technology. Some of the successful applications

A. Borji (✉)
School of Cognitive Sciences, Institute for Studies in Theoretical Physics and Mathematics, Niavaran Bldg., P. O. Box 19395-5746, Tehran, Iran
e-mail: borji@ipm.ir

M. Hamidi
Computer and Information Technology Department, Azad University Branch of Zarghan, Booali Boulevard, Azad University Street, P. O. Box 73415-314, Zarghan, Iran

F. Mahmoudi
Computer, Engineering and Information Technology Department, Azad University Branch of Qazvin, Qazvin, Iran

are: mail sorting, form data entry, bank checking processing, etc. Huge amount of research in this area has also contributed to solve other open problems in pattern recognition.

Two of the most successful approaches in the literature of OCR are Neocognitron [1] and Convolutional Networks [2], which resemble biology to some extent. In [3], a successful biology inspired approach for handwritten digit recognition is reported for Indian numerals using probabilistic neural networks. In [4], a selective attention based method for visual handwritten digit recognition is proposed. A benchmark analysis for state-of-the-art handwritten digit recognition techniques is introduced in [5]. For review of other successful methods for English handwritten character recognition, the reader is referred to [6–9]. A few numbers of studies have been reported for Farsi language [10–12]. One drawback with previous studies on Farsi language is that, they have reported their results on different non-standard datasets which makes comparison of their results difficult. In this paper, experiments are carried out on standard Farsi datasets recently developed and published in [13,14].

A long term desire of researchers has been to mimic the human behavior on character recognition because of his efficiency, speed and robustness in presence of various image distortions. The recent attempts in neuroscience have led to great advances in revealing the mysteries behind function, organization and anatomy of the human and primate visual systems. Such studies have the advantage of generating ideas for developing artificial solutions for tasks in which humans have high performances.

Researchers at MIT university[1] have tried to quantitatively model the processing of the visual ventral stream during visual perception and object recognition tasks based on extensive neurophysiological, psychophysical and fMRI studies. In [15], authors have investigated the shape representation in visual area V4 using standard model of object recognition (HMAX). State of the art theory of object recognition in feedforward path of the visual ventral stream is presented in [16]. Other biophysical models of neural computation and invariant visual representation by single neurons are discussed in [17,18]. In [19], they have shown that the standard model also accounts for rapid categorization. In [20], standard model is applied to some computer vision applications like object and face recognition, scene understanding, etc.

In this paper, we investigate the application of the above mentioned model for recognition of both handwritten digits and characters based on two motivations (1) Characters are special forms of objects (Shapes) and (2) Same visual structures are involved in recognition of both characters and objects. We also analyze the sensitivity of this set of biology inspired features to various image distortions.

The rest of this paper is organized as follows. In Sect. 2 a brief literature of the visual system to the extent relevant to our work is reviewed. Standard model of visual ventral stream and feature extraction is explained in this Section too. Experiments and results are illustrated in Sects. 3 and 4 respectively. Finally Sect. 5 brings discussions and concludes the paper.

2 Biology of the Visual System

Modeling response properties of neurons in early visual areas has resulted to several applications in image processing and computer vision. Gabor wavelet filters as a model of V1 neurons [21,22] have been widely used for edge detection, texture processing, writer identification, font recognition, etc. Despite the great familiarity with processing in early visual

[1] http://cbcl.mit.edu.

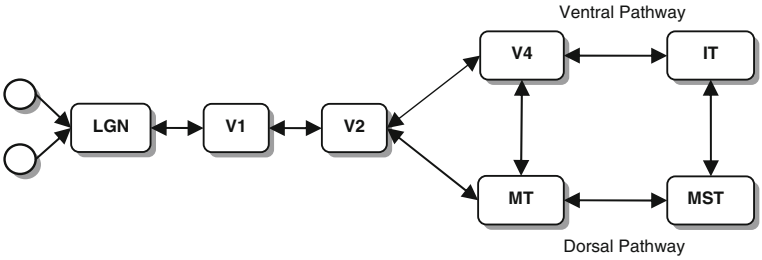

Fig. 1 Basic divisions of the human visual system. Visual input from two eyes is transmitted to the primary visual cortex (V1) via LGN. It is then sent to visual area V2 and from there is segregated in two visual pathways. Ventral stream is mainly focused on object recognition, while dorsal stream processes spatial information

areas, less abstract information is yet available on the functionality of higher visual areas which seem to be more important in higher cognitive tasks. Current work is intended to step further beyond the common inspirations from the early vision and investigate complex operations in the higher visual areas for optical character recognition.

The primary visual cortex (also referred to striate cortex or V1) is the first cortical area in visual cortex that receives information from a thalamical component called lateral geniculate nucleus (LGN). LGN in turn is a major destination area for outputs of Retina. V1 is highly specialized for processing information about static and moving objects and is excellent in pattern recognition. It then sends projections to other higher areas like V2, V3, V4 and V5 (also called MT), together called extrastriate cortex. Neurons in the primary visual cortex in contrast to neurons in higher areas are selective to simple features. For example a V1 neuron fires when a vertical bar is located in its receptive field (RF).[2] Neurons in area V2 have many properties in common with V1 neurons. They respond to orientation, spatial properties such as illusory contours and whether the stimulus is part of the figure or background. In area V3, neurons are in general sensitive to global motion and combinations of visual stimuli. Color-sensitive neurons are also more common in area V3. Neurons in area V4 like V1 neurons are responsive to orientation, spatial frequency and color, but unlike V1 neurons, they are tuned to object features of intermediate complexity like simple geometric shapes. Visual area V5 is thought to play a major role in the perception of motion, the integration of local motion signals into global percepts and the guidance of some eye movements. The last stage of ventral stream, in an area called inferotemporal cortex, IT, neurons respond to complex shapes like objects, faces and hands.

Two visual pathways segregate from the primary visual cortex: ventral and dorsal streams. Details of these streams, shown in Fig. 1, are as follows.

Ventral Stream: Starts from V1, goes through visual area V2, then through area V4 and from there to inferior temporal lobe. It is also known as "what pathway". As its name reveals, it is associated with form recognition and object representation, stating otherwise "what aspects" of an object. Neurons along this pathway answer typically to form, color, etc. From a pattern recognition point of view, this pathway is in charge of extracting features for tasks like categorization and identification. In the last stage of this pathway in area IT, neurons are sensitive to complex objects. There are projections from this area to prefrontal cortex (PFC) which is associated with storage of long term memory.

[2] A region of the visual field in which presence of a stimulus alters firing of a neuron.

Dorsal Stream: Starts from V1, goes to V2, and then goes to the dorsomedial area and MT. It then projects to inferior parietal lobe. The dorsal stream is also called "where stream" and is associated more with spatial properties like motion, representation of object location and control of the eyes and arms, especially when visual information is used to guide saccades or reaching.

Visual processing in ventral pathway is done in a hierarchical manner with the advantage of being scale and orientation invariant. Early visual areas, Retina, LGN, V1 and V2, are involved in extraction of simple features like orientation, edge, intensity, color, etc. As we move up in the visual hierarchy of ventral stream, the optimal stimulus of a neuron becomes more complex, while its receptive field gets larger. Neuronal activity recording studies on monkeys have revealed that object recognition is mostly feedforward perhaps for serving fast recognition. In higher stages of the ventral hierarchy neurons show plasticity and learning.

An overall understanding of how visual processing is performed could be attained by considering all findings from each individual component and relationships among different components. This information could be used to model a visual phenomenon. Computational modeling and experimental studies have mutual dependency on each other. Experimental studies help modeling works by preparing them more accurate data, thus ending to more biology plausible models. On the other hand computational modeling gives experimentalists a better notion of how processing might happen as wells as providing hints for orienting their experiments.

The rest of this section presents a brief introduction to standard model of object recognition, HMAX, based on generally acceptable findings from neuroscience. Details of the model could be found in the appendix. The model consists of four alternative layers of Simple (S) and Complex (C) units. Different models of neurons are considered in each layer. S and C units, model the function of simple and complex neurons in visual area V1. Units S2 and C2 mimic the behavior of neurons in areas V4 and IT. The S units combine their inputs with Gaussian-like tuning to increase object selectivity. The C units pool their inputs through a maximum operation, thereby introducing invariance to scale and translation. C2 features of the model are used here for classification. HMAX model in its simplest version [23] uses a very simple static dictionary of features. In [24] authors have extended the basic HMAX model by incorporating a learning mechanism to learn a vocabulary of features from a set of positive input patterns. Figure 2 illustrates the structure of the standard model of object recognition. In our experiments we used a MATLAB® implementation of the model.[3] Model parameters used in our experiments are listed in Table 1.

3 Experimental Setup

We tested the presented method on both standard Farsi and English datasets. For the Farsi dataset, experiments were performed on a recently developed standard Farsi digit corpus [13]. For the English dataset, we employed one of the most common datasets: the MNIST handwritten digit corpus which is widely used in the literature [25–28]. In order to further evaluate the generalization of the method we conducted an experiment for recognizing Farsi handwritten characters. Details of these datasets are explained in following sections.

[3] Implementation HMAX model could be downloaded from: http://riesenhuberlab.neuro.georgetown.edu/hmax.html.

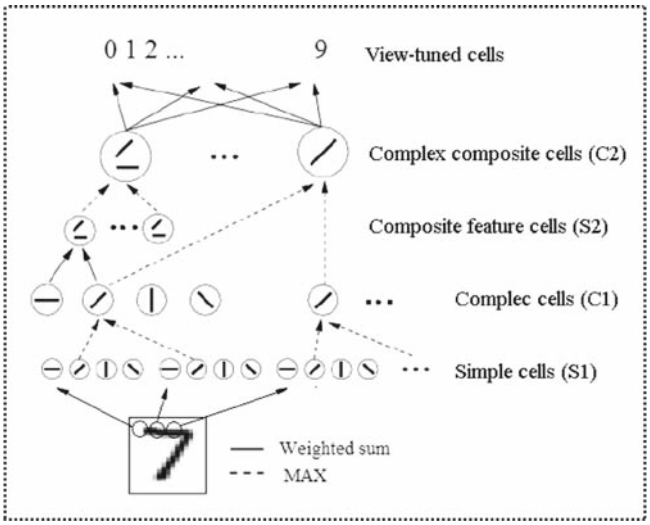

Fig. 2 Standard model of object recognition in the visual ventral stream. Visual input data are processed in a hierarchical manner like visual system. In each layer of the hierarchy, models of neurons in that layer process image. Simple extracted features in early layers successively build more complex features in top layers. In the topmost layers, nodes code the view tuned representation of the objects (here digit and character classes). Image Courtesy of Thomas Serre and Tomaso Poggio

Table 1 Model (HMAX) parameters used in the experiments

Band 1	Band 2	Band 3	Band 4	Band 5	Band 6	Band 7	Band 8
Filter size s 7 & 9	11 & 13	15 & 17	19 & 21	23 & 25	27 & 29	31 & 33	35 & 37
Grid size 8 × 8	10 × 10	12 × 12	14 × 14	16 × 16	18 × 18	20 × 20	22 × 22
Gabor σ 2.8 & 3.6	4.5 & 5.4	6.3 & 7.3	8.2 & 9.2	10.2 & 11.3	12.3 & 13.4	14.6 & 15.8	17.0 & 18.2
Gabor λ 3.5 & 4.6	5.6 & 6.8	7.9 & 9.1	10.3 & 11.5	12.7 & 14.1	15.4 & 16.8	18.2 & 19.7	21.2 & 22.8

3.1 Farsi Digit Dataset[4]

Khosravi et al. [13] have introduced a very large corpus of Farsi handwritten digits. They extracted handwritten digits from 11,942 forms filled by diploma and bachelor students registered in the Iran's nationwide university entrance exam; 5,393 forms were filled by Diploma students and 6,549 others by BS students. All forms were scanned at 200 dpi resolution in 24 bit color format. After applying a threshold, they came into 102,352 binary images from which they chose 60,000 images for train and 20,000 for test.

3.2 Modified NIST (MNIST)[5]

In the spring of 1992, the National Institute of Standards and Technology organized a classification competition for handwritten digits over original NIST dataset which included a training set of 223,000 and a test set of 59,000 samples [29]. These two sets had different

[4] Contact information for obtaining the dataset is available at http://www.modares.ac.ir/eng/kabir, where a 1000-sample subset of the database can be freely downloaded.

[5] The MNIST dataset is downloadable from: http://yann.lecun.com/exdb/mnist.

Table 2 Distribution of digits in train and test sets of both digit datasets

		0	1	2	3	4	5	6	7	8	9	Total
Farsi	Train	6,000	6,000	6,000	6,000	6,000	6,000	6,000	6,000	6,000	6,000	60,000
	Test	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	20,000
MNIST	Train	5,923	6,742	5,958	6,131	5,842	5,421	5,918	6,265	5,851	5,949	60,000
	Test	980	1,135	1,032	1,010	982	892	958	1,028	974	1,009	10,000

Table 3 Farsi character dataset class distribution (Train and Test).

Character	آ	ب	پ	ت	ث	ج	چ	ح	خ	د	ذ	ر
Train size	9994	1049	209	1421	40	559	50	1228	360	2660	40	3018
Test size	3001	316	64	427	13	169	16	370	109	799	13	1321
Character	ز	ژ	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق
Train size	1116	50	3049	719	470	360	200	65	872	140	916	470
Test size	336	16	715	167	142	109	61	16	262	43	276	142
Character	ک	گ	ل	م	ن	و	ه	ی	آ	ء	هه	
Train size	640	260	3000	4970	3829	1360	2530	4520	70	70	280	
Test size	193	79	901	1492	1149	409	760	1357	22	22	85	

distributions which affected the test results. A modified dataset was then built by merging the two sets using a 50/50 ratio into a new set with 60,000 train and 10,000 test samples. The new dataset is called the Modified NIST or simply MNIST. The main difference between this set and the previous one is the number of samples. Digits in the MNIST dataset are slightly bigger and are stored in images of 28×28 pixels. Pixel intensities are in the range of 0 and 255.

3.3 Farsi Character Dataset

Isolated Farsi/Arabic Handwritten Character Database (IFHCDB)[6] was created at the electrical engineering department of Amirkabir university of technology in 2006 [14]. It contains gray scale 300 dpi images of characters and numerals. The number of samples in each class of this database is non-uniform corresponding to their real life distributions. IFHCDB is a subset of a larger database gathered as part of a research project under sponsorship of Iranian national information and communication technology (ICT) organization. The main goal of this database is to help researchers to develop new techniques, technology and algorithms for automatic recognition of handwritten Farsi/Arabic characters. Distributions of digits of both English and Farsi datasets are shown in Tables 2 and 3. Figure 3 illustrates some example patterns from each dataset.

[6] Contact information for obtaining the dataset is available at http://ele.aut.ac.ir/imageproc/downloads/IFHCDB.htm.

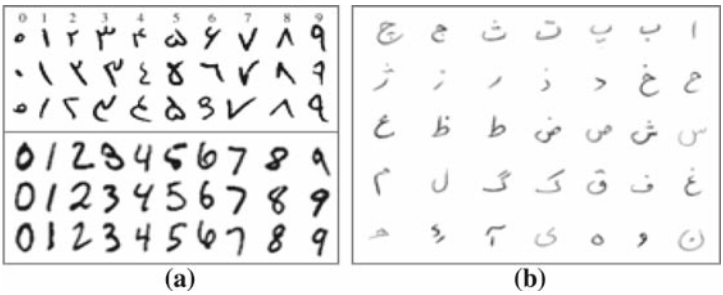

Fig. 3 (**a**) Sample digits from Farsi (up) and MNIST datasets (bottom). (**b**) Sample characters from IFHCDB dataset

4 Results

In this section, we show that the set of biology inspired features introduced in Sect. 2, are suitable for handwritten character recognition and comply well with the nature of this problem. Higher geometrical structures like corners, ridges, curves etc are more useful in human recognition of shapes than simple orientation features. Our visual system achieves high efficiency by means of neurons responsive to complex features which are themselves built upon simple orientation features in a visual hierarchy [16,20].

According to the above statement, our aim in this paper is to examine the following hypothesis: "The special organization of human visual system has an essential role in his achievement in object recognition, thus an elaborate model of it which fits highly with the experimental data, might perform well on optical character recognition".

To prove that this set of biology inspired features is appropriate for character recognition, we first show high accuracy is attainable on both digit and character datasets using three types of classifiers. Next, we explain that these features lead to a recognition method with high solidity to image distortions like orientation, scale, and translation distortions, therefore having high generalization over pattern transformations.

4.1 Classification

Train and test images are both resized to 32×32 pixel images at first. Then a real valued feature vector of length either 256 or 4,096, depending on the number of orientations in the S1 layer (4 or 8), is extracted. Two types of filters in S1 layer are considered: Gabor and Derivative of Gaussians (DOG) (see Appendix). In each feature dimension, derived features are linearly normalized in the range [0, 1].

Trained classifier over training feature vectors is later evaluated on a separate test set. Reported results are averaged over ten runs on randomly ordered training sets. Both multi layer perceptrons (MLP) [30] and support vector machines (SVM) [31] classifiers are examined to show that the proposed features are independent of classifier type. To further investigate the effectiveness C2 features, we have conducted an experiment with kNN ($k = 3$) classifier. Based on some experiments the value of k was set to 3.

To do digit classification with ANN, we used a three-layer MLP neural network with 10 log-sigmoid neurons in the hidden layer. Four linear neurons in the output layer used binary coding to code 10 digit classes. To stop training, a validation set consisting one tenth of the training set was selected in a way to preserve the relative number of patterns in classes. In

order to decrease computational load and to achieve high accuracy, dimensionality reduction was performed using principal component analysis (PCA). We found that, only the first 25 principal components using four orientations on Farsi digit dataset (32 for MNIST) were sufficient to express 90% of variance of training samples (including validation set). When using eight orientations, top 37 principal components are enough to express 90% of variance on Farsi digit dataset and 45 on MNIST. These numbers determine dimensionality of the input vectors and thus number of neurons in the input layer of ANN.

To perform classification with support vector machines, the same preprocessing and feature extraction were followed as in ANN. Three types of kernels 'Linear', 'Polynomial', and 'RBF' were investigated. Classification results over both digit datasets are shown in Table 4. Numbers in parentheses show standard deviations. There was no dimensionality reduction here.

Table 5 illustrates the confusion matrix of SVM classification on the Farsi digit dataset. As it shows, there are great overlaps in classifying patterns {2, 3, 4}, {5, 0}, and {6, 9} in the training phase. Analyzing confusion matrix over MNIST dataset, we noticed overlaps among {4, 7, 9} and {2, 3, 5, 8} patterns in SVM classification.

To reach higher accuracy, we constructed a two stage cascade classifier. In the first stage, a six class, classification problem was solved by putting each set of overlapped patterns in one class. In the next stage, different preprocessing was performed. These preprocessings

Table 4 Handwritten digit recognition rate over both digit datasets

Dataset	HMAX Parameters		Classifier				
	Filter	Orientations	KNN $K = 3$	ANN	SVM		
					Linear ($d = 0$)	Polynomial ($d = 2$)	RBF
MNIST	Gabor	4	77.1(.7)	93.3(.61)	93.16(.36)	94.02(1.08)	96(.15)
		8	81.7(1.1)	93.6(.42)	95.5(.61)	96.4(.52)	96.1(.29)
	DOG	4	76.3(1.03)	89.9(.82)	88.6(.78)	90.9(.92)	92.2(.46)
		8	77(.65)	92.11(.67)	90.8(.17)	92(1.2)	93.7(.58)
Farsi	Gabor	4	81.3(1.5)	95.30(.78)	94.89(1.2)	95.02(.54)	95.12(.8)
		8	84.7(.5)	97.64(.93)	96(.43)	95.7(.22)	96.51(.68)
	DOG	4	78(.24)	92.3(1.1)	89.1(1.1)	90.9(1.4)	92.2(1.1)
		8	79.4(.96)	93.35(.29)	91(.84)	93.44(.65)	94.3(.99)

Polynomial degree two was used. Numbers in parentheses shows standard deviations over 10 runs (%)

Table 5 Confusion matrix for SVM classifier over Farsi digit dataset

	0	1	2	3	4	5	6	7	8	9
0	7,648									
1	15	7,750								
2	7	42	7,464							
3	18	7	573	7,327						
4	63	3	57	158	7,626					
5	164	1	6	17	23	7,650				
6	97	27	40	14	15	34	7,582			
7	8	9	41	9	1	0	47	7,715		
8	1	4	0	0	0	5	0	0	7,785	
9	21	41	5	4	26	38	112	0	12	7,657

Numbers are summed for (a, b) and (b, a) permutations

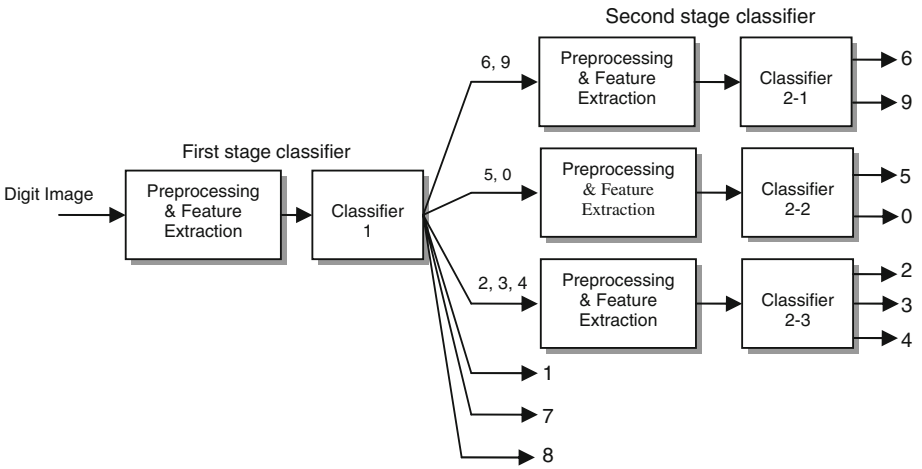

Fig. 4 Cascade classifier for digit recognition over the Farsi digit dataset. In the first stage, input data are preprocessed and features are extracted using HMAX. Then a classifier is built over these features. A second stage classifier is then used to separate overlapped classes. A different preprocessing (usually edge detection) is done in the second stage

were done to increase the separability among overlapped patterns. A schematic diagram of this two-stage classification is illustrated in Fig. 4. Results in Table 6 show the recognition rates of the cascade classifier.

As the results in Table 6 shows, recognition rate of 98.6% (96.5% over MNIST) was achieved using C2 features. The high recognition rate of 99.63% (98.9% over MNIST) was obtained using a two stage cascade classifier with Sobel edge detection preprocessing at the second stage. It could be easily verified that results with the cascade classifier are higher than the base classifiers.

For the purpose of character recognition, an MLP classifier was used with 10 log-sigmoid neurons in the second and six linear neurons in the output layer to code 35 classes. All other settings were the same except dimensionality reduction which was 24 first principal components for four orientations in S1 layer and 31 for 8 orientations. The recognition rate of 93.2% (STD = 0.89) was achieved using 8 Gaussian filters in S1 layer. Using four Gabor filters we reached 97.08% recognition rate (STD = 0.51).

4.2 Sensitivity Analysis

Like human visual system a distinguished advantage of the proposed features is their robustness to various image distortions. Classifiers built over these features act better than other methods at least in sense of being invariant to image distortions. Some experiments are performed in this section to investigate sensitivity of C2 features against orientation, scale and translation noises. For each image in two digit datasets, both train and test, a discrete random number was uniformly generated between 1 and 3. If the outcome was 1 the image was left unchanged. If it was 2, orientation noise and if it was 3, scale and translation noise was added to the image. Training and testing were done in the same manner as stated in previous section using distorted images.

In order to add orientation noise, a random number θ was generated from a uniform distribution in the range of $[-30, 30]$, then the digit image was rotated θ degrees. To add scale and

Table 6 Handwritten digit recognition rate over both digit datasets using a cascade classifier (%)

Dataset	Preprocessing	HMAX parameters		Classifier				
		Orientations	Filter	KNN	ANN $K = 3$	SVM		
						Linear ($d = 0$)	Polynomial ($d = 2$)	RBF
MNIST	No preprocessing	4	Gabor	81.3(.23)	94.65(1.2)	96.14(.73)	97.70(1.03)	98.75(.89)
	Canny edge detection			83.3(1.3)	95.1(.49)	97.8(.33)	98.1(.58)	98.15(1.1)
	Sobel edge detection			87.4(.62)	97.7(.25)	97.95(.3)	98.4(.72)	98.55(.31)
	Sobel edge detection	8		87.8(.83)	97(.4)	97.62(.34)	98.9(.61)	98.39(.43)
Farsi	No preprocessing	4	Gabor	83.3(.91)	98.65(.44)	98.14(.61)	96.6(.58)	98.75(.21)
	Canny edge detection			87(.76)	98.92(.18)	98.3(.63)	97.1(.55)	98.5(.2)
	Sobel edge detection			89(1.07)	99(.71)	98.78(.52)	99.3(.25)	99.4(.24)
	Sobel edge detection	8		91.5(1.4)	99.41(.54)	99.12(.53)	99.25(.6)	99.63(.31)

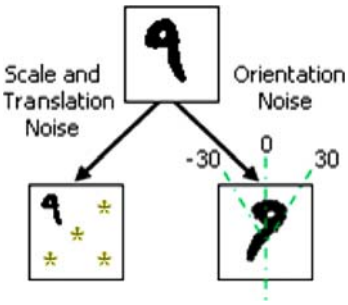

Fig. 5 Adding distortions to digit images. Two types of distortions are considered. A digit image is randomly scaled and translated (shown in left). The digit image is randomly rotated in the range [−30, 30] (shown in right)

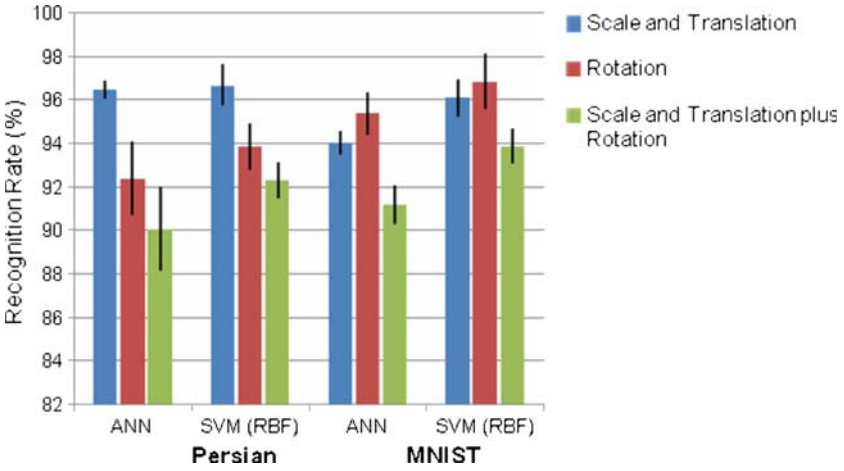

Fig. 6 Recognition rates over distorted handwritten digits using SVM, ANN and kNN classifiers

translation distortion, first a random number, $0.5 < \sigma < 1$, was uniformly generated and then the digit image was scaled accordingly. Then a second random number $\mu \in \{1, 2, 3, 4, 5\}$ representing translation in five directions {up-right, bottom-right, bottom-left, up-left, middle} was randomly selected with the same chance. After that the scaled image was translated in a way not to cross the image boundaries. Figure 5 illustrates an image under distortion operation. Star signs in this figure show the five possible positions for translation.

Figure 6 shows classification results over distorted datasets. Recognition was done with a cascade classifier with DOG filters and eight orientations in the first stage and Sobel edge detection in the second stage.

The Sensitivity analysis over both datasets demonstrates high solidity of the proposed features to the image distortions. Adding rotation noise reduces performance but less than scale and translation noise. When both distortions were added simultaneously, performance fell to the lowest magnitude.

5 Discussions and Conclusions

Classification results show high recognition rates using ANN and SVM classifiers. SVM classifier outperforms other two classifiers in average. Results rank SVM kernels as RBF,

Table 7 Comparison of C2 features for handwritten digit recognition over MNIST (%)

Method	Our method	Ranzato et al. NIPS 2006	Ranzato et al. CVPR 2007	Simard et al. ICDAR 2003	LeCun et al. 1998	LeCun et al. 1998
Features	C2 features	Large conv. net, unsupervised pretraining	Large conv. net, unsupervised features	Conv. net, cross-entropy	Conv. net LeNet-5	Conv. net LeNet-4
Recognition error (%)	1.1	0.60	0.62	0.4	0.95	1.4

polynomial and linear over both digit datasets. When shifting to a usually considered weak classifier (kNN) for recognition, performance remained still high.

We have assessed Gabor versus DOG filters, and for simple cell responses, we have shown that Gabor functions resulted in higher accuracy when compared to DOG filters. This result is in accordance with results reported in [32]. It is important to note that this does not necessarily mean that Gabors are more successful in modeling single cell responses. This could be only demonstrated by testing over neurological data.

Increasing the number of orientations also results to higher recognition rates. High recognition rate over Farsi characters further supports the idea that these features might be language independent. Results also show that recognition rate is still high in the presence of distortions.

Table 7 compares the classification error over MNIST dataset using C2 features with other methods in the literature. Although our method has higher classification error compared with newer methods (columns 3 to 5 in Table 7), its performance is near famous LeNet-5 and is higher than LeNet-4. Despite having higher classification error, using C2 features not only shows effectiveness of a biologically inspired approach for handwritten digit recognition, which has its own benefits, it also has the promise for future extensions.

The best performance reported in [33] over IFHCDB database is 96.92%. Using C2 features we were able to achieve 97.08% accuracy without cascade classifiers. In [13], authors have used a multiple classifier system consisting of four MLP classifiers for digit recognition over the same Farsi database used in this study. Using a modified gradient technique over 15,000 train and 5,000 test digits, they achieved 98.8% recognition rate which is above 96.51% but below 99.63% recognition rates we achieved with and without cascade classifier respectively both with Gabor filters.

Results also show that edge detection has not significant effect on enhancing the results. That is because Gabor filters perform edge detection in the S1 layer of the HMAX model. Therefore improvement in the results shown in the Table 6 is mainly because of the cascade classifier in the second stage.

Application of features inspired by the primate visual system for handwritten character recognition was investigated in this paper. A set of scale and translation invariant C2 features from a training set of digit images was first extracted. Then a classifier was constructed over these data. Classification results using these features prove appropriateness of these features for the mentioned problems.

Success of these features over this problem has two advantages. First, it suggests another evidence for the standard model of object recognition in ventral stream of visual cortex. Second it introduces a general framework for many pattern recognition and machine vision problems with the same structural characteristics where complex structures are built over simple basic features. This of course needs more verification both theoretically and experimentally over other practical problems.

Acknowledgements Authors would like to thank anonymous reviewers for their helpful comments on the manuscript.

Appendix: Model Implementation

S1 units

A gray-value input image is first analyzed by a multidimensional array of simple S1 units. S1 units take the form of Gabor functions [21], which have been shown to provide a good model of cortical simple cell receptive fields [22]. Filter parameters are adjusted so that the tuning properties of the corresponding S1 units match with the V1 parafoveal simple cells [18,19,23,24]. S1 filters are arranged in such a way to form a pyramid of 16 scales, spanning a range of sizes from 7×7 to 37×37 pixels in steps of two pixels. In our experiments, we considered either four or eight orientations. Starting from zero and moving counterclockwise in steps of 45 or 22.5 degrees leads to either four or eight orientations, which when is calculated over 16 scales leads to 64 or 128 filters. At each pixel of the input image, filters of each size and orientation are centered. The filters are sum-normalized to zero and square-normalized to 1, and the result of the convolution of an image patch with a filter is divided by the power (sum of squares) of the image patch. This yields an S1 activity between -1 and 1. Gabor and derivative of Gaussian (DOG) filters formulas are shown in equations one and two respectively.

$$F(x, y) = \exp\left(-\frac{x_0^2 + y^2 y_0^2}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda}\right) x_0 \cdot \mathrm{s} \cdot \mathrm{t}. \tag{1}$$

$$x_0 = x\cos\theta + y\sin\theta \quad \text{and} \quad y_0 = -x\sin\theta + y\cos\theta$$

$$G^2(x, y) = \frac{\gamma^2 - \sigma_y^2}{2\pi\sigma_x\sigma_y^2} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \tag{2}$$

C1 units

In the next step, filter bands are defined, i.e., groups of S1 filters of a certain size range (7×7 to 9×9 pixels; 11×11 to 15×15 pixels; ...; 35×35 to 37×37 pixels). Within each filter band, a pooling range is defined which determines the size of the array of neighboring S1 units of all sizes in that filter band which feed into a C1 unit (roughly corresponding to complex cells of striate cortex). Only S1 filters with the same preferred orientation feed into a given C1 unit to preserve feature specificity. We used pooling range values from 4 for the smallest filters (meaning that 4×4 neighboring S1 filters of size 7×7 pixels and 4×4 filters of size 9×9 pixels feed into a single C1 unit of the smallest filter band) over 6 and 9 for the intermediate filter bands, respectively, to 12 for the largest filter band. The pooling operation

that the C1 units use is the "MAX" operation, i.e., a C1 unit's activity is determined by the strongest input it receives.

S2 units

In the S2 layer, units pool over afferent C1 units from a local spatial within each filter band, a square of four adjacent, non-overlapping C1 units is then grouped to provide input to a S2 unit. There are 256 different types of S2 units in each filter band, corresponding to the 4^4 possible arrangements of four C1 units of each of four types (i.e., preferred bar orientation). The S2 unit response function is a Gaussian with mean 1 (i.e., {1; 1; 1; 1}) and standard deviation 1, i.e., an S2 unit has a maximal firing rate of 1 which is attained if each of its four afferents fires at a rate of 1 as well. S2 units provide the feature dictionary of HMAX, in this case all combinations of 2×2 arrangements of "bars" (more precisely, C1 cells) at four possible orientations.

C2 units

To finally achieve size invariance over all filter sizes in the four filter bands and position invariance over the whole visual field, the S2 units are again pooled by a MAX operation to yield C2 units, the output units of the HMAX core system, designed to correspond to neurons in extrastriate visual area V4 or posterior IT (PIT). There are 256 C2 units, each of which pools over all S2 units of one type at all positions and scales. Consequently, a C2 unit will fire at the same rate as the most active S2 unit that is selective for the same combination of four bars, but regardless of its scale or position.

References

1. Fukushima K (1980) Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern 36(4):193–202. doi:10.1007/BF00344251
2. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W et al. (1990) Handwritten digit recognition with a back-propagation network. In: Touretzky D (ed) Advances in Neural Information Processing Systems 2 (NIPS 89)
3. Al-Omari FA, Al-Jarrah O (2004) Handwritten Indian numerals recognition system using probabilistic neural networks. Adv Eng Inform 18(1):9–16. doi:10.1016/j.aei.2004.02.001
4. Salah AA, Alpaydin E, Akarun L (2002) A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition. IEEE Trans Pattern Anal Mach Intell 24(3):420–425. doi:10.1109/34.990146
5. Liu CL, Nakashima K, Sako H, Fujisawa H (2003) Handwritten digit recognition: Benchmarking of state-of-the-art techniques. Pattern Recognit 36(10):2271–2285. doi:10.1016/S0031-3203(03)00085-2
6. Shi M, Fujisawa Y, Wakabayashi T, Kimura F (2002) Handwritten numeral recognition using gradient and curvature of gray scale image. Pattern Recognit 35(10):2051–2059. doi:10.1016/S0031-3203(01)00203-5
7. Teow LN, Loe KF (2002) Robust vision-based features and classification schemes for off-line handwritten digit recognition. Pattern Recognit 35(11):2355–2364. doi:10.1016/S0031-3203(01)00228-X
8. Cheung K, Yeung D, Chin RT (1998) A Bayesian framework for deformable pattern recognition with application to handwritten character recognition. IEEE Trans Pattern Anal Mach Intell 29(12):1382–1388. doi:10.1109/34.735813
9. Tsang IJ, Tsang IR, Dyck DV (1998). Handwritten character recognition based on moment features derived from image partition. In: International conference on image processing, vol 2, pp 939–942
10. Soltanzadeh H, Rahmati M (2004) Recognition of Persian handwritten digits using image profiles of multiple orientations. Pattern Recognit Lett 25(14):1569–1576. doi:10.1016/j.patrec.2004.05.014
11. Said FN, Yacoub RA, Suen CY (1999). Recognition of English and Arabic numerals using a dynamic number of hidden neurons. In: Proceedings of the fifth international conference on document analysis and recognition, pp 237–240

12. Sadri J, Suen CY, Bui TD (2003) Application of support vector machines for recognition of handwritten Arabic/Persian digits. In: Second Iranian conference on machine vision and image processing, vol 1, pp 300–307
13. Khosravi H, Kabir E (2007) Introducing a very large dataset of handwritten Farsi digits and a study on their varieties. Pattern Recognit Lett 28(10):1133–1141. doi:10.1016/j.patrec.2006.12.022
14. Dehghan M, Faez K, Ahmadi M, Shridhar M (2001) Handwritten Farsi (Arabic) word recognition: a holistic approach using discrete HMM. Pattern Recognit 34(5):1057–1063. doi:10.1016/S0031-3203(00)00051-0
15. Kouh M, Riesenhuber M (2003) Investigating Shape Representation in Area V4 with HMAX: orientation and grating selectivities. CBCL Paper #231/AIM #2003-021, Massachusetts Institute of Technology, Cambridge, MA
16. Serre T, Kouh M, Cadieu C, Knoblich U, Kreiman G, Poggio T (2005). A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. AI Memo 2005-036/CBCL Memo 259, Massachusetts Institute of Technology, Cambridge, MA
17. Knoblich U, Bouvrie J, Poggio T (2007) Biophysical models of neural computation: max and tuning circuits. CBCL paper, Cambridge, MA
18. Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I (2005) Invariant visual representation by single neurons in the human brain. Nature 435:1102–1107. doi:10.1038/nature03687
19. Serre T, Oliva A, Poggio T (2007) A feedforward architecture accounts for rapid categorization. Proc Natl Acad Sci USA 104(15):6424–6429. PNAS. doi:10.1073/pnas.0700622104
20. Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T (2007) Object recognition with cortex like mechanisms. IEEE Trans Pattern Anal Mach Intell 29(3):411–426. doi:10.1109/TPAMI.2007.56
21. Gabor D (1946) Theory of communication. J Inst Electr Eng 93(26):429–457
22. Hubel D, Wiesel T (1965) Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. J Neurophysiol 28:229–289
23. Serre T, Riesenhuber M (2004) Realistic modeling of simple and complex cell tuning in the hmax model, and implications for invariant object recognition in cortex. Technical Report CBCL Paper 239/AI Memo 2004- 017, Massachusetts Institute of Technology, Cambridge, MA
24. Serre T, Wolf L, Poggio T (2004) A new biologically motivated framework for robust object recognition. Technical Report CBCL Paper 243/AI Memo 2004- 026, Massachusetts Institute of Technology, Cambridge, MA
25. Keysers D, Deselaers T, Gollan C, Ney H (2007) Deformation models for image recognition. IEEE Trans Pattern Anal Mach Intell 29(8):1422–1435. doi:10.1109/TPAMI.2007.1153
26. Zhang P, Bui TD, Suen CY (2007) A novel hierarchical ensemble classifier system with a high recognition performance on handwritten digits. Pattern Recognit 40(12):3415–3429. doi:10.1016/j.patcog.2007.03.022
27. Marc'Aurelio R, Poultney C, Chopra C, LeCun Y (2006) Efficient learning of sparse representations with an energy-based model. In: Platt J et al (eds) Advances in Neural Information Processing Systems (NIPS 2006). MIT Press
28. Kussul EM, Baidyk TN, Wunsch DC II, Makeyev O, Martin A (2006) Permutation coding technique for image recognition systems. IEEE Trans Neural Netw 17(6):1566–1579. doi:10.1109/TNN.2006.880676
29. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324. doi:10.1109/5.726791
30. Rumelhart DE, McClelland JL (1986) Parallel distributed processing, vol 1 & 2. MIT, Cambridge
31. Vapnik VN (1995) The nature of statistical learning theory. Springer-Verlag, NewYork
32. Serre T, Wolf L, Poggio T (2005) Object recognition with features inspired by visual cortex. In: Proceedings of IEEE conference computer vision and pattern recognition, Massachusetts Institute of Technology
33. Dehghan M, Faez K (1997) Farsi handwritten character recognition with moment invariants. In: Proceedings of the 13th international conference of digital signal processing, vol 2, issues 2–4, pp 507–510