Machine Vision and Applications
DOI 10.1007/s00138-009-0216-9

SHORT PAPER

Invariance analysis of modified C2 features: case study—handwritten digit recognition

Mandana Hamidi · Ali Borji

Received: 29 November 2007 / Revised: 8 June 2009 / Accepted: 19 July 2009

Abstract Humans are very efficient in recognizing alphanumeric characters, even in the presence of significant image distortions. Recent advances in visual neuroscience have led to a solid model of object and shape recognition in the visual ventral stream which competes with the state-of-the-art computer vision systems on some standard recognition tasks. A modification of this model is also proposed by adding more biologically inspired properties such as sparsification of features, lateral inhibition and feature localization to enhance its performance. In this study, we show that using features proposed by the modified model results in higher handwritten digit recognition rates compared with the original model over English and Farsi handwritten digit datasets. Our analyses also demonstrate higher invariance of the modified model to various image distortions.

Keywords Handwritten digit recognition · Optical character recognition · Visual ventral stream · Sparsification of features · Lateral inhibition · Feature localization · C2 features · HMAX

M. Hamidi (✉)
Telerobotics and Applications Department, Italian Institute of Technology (IIT), Via Morego 30, 16163 Genoa, Italy
e-mail: mandana.hamidi@iit.it

A. Borji
School of Cognitive Sciences, Institute for Research in Fundamental Sciences, Niavaran, P.O. Box 19395-5746, Tehran, Iran

A. Borji
Institute of Computer Science III, Rheinische Friedrich-Wilhems-Universität, 53117 Bonn, Germany
e-mail: borji@iai.uni-bonn.de

1 Introduction

Object recognition is one of the advanced tasks of the visual cortex and is very crucial for primates. While it has attracted a lot of attention in computer vision, its underlying neural mechanisms and computational processes in the visual cortex are not yet completely understood.

Based on a large number of experimental studies on monkey and human visual system, it is thought that object recognition is mediated by the ventral visual pathway [1]. This pathway starts from primary visual cortex, V1, and goes to cortical areas V2 and V4. It then goes to inferotemporal cortex (ITC)—a center for face and object processing—and prefrontal cortex (PFC) which plays an important role in linking perception to memory. Visual ventral stream has a hierarchical structure which shows sensitivity to increasing complexity of the preferred stimuli from simple cells in V1 to complex cells in V4 and IT.

A model of object recognition based on experimental findings from neuroscience known as Hierarchical Model and X (HMAX)[1] has been proposed by Reisenhuber and Poggio [2]. This model has provided some valuable explanations for neuroscientific phenomena in the visual cortex [3]. A version of HMAX exists which reflects physiological data and performs at the level of humans on some visual tasks [4]. It has been claimed that HMAX has desirable properties such as significant degree of translation and scale invariance [2] and competes with some of the state-of-the-art computer vision systems for applications such as object and face recognition, scene interpretation, etc., [5].

The high efficiency of HMAX is due to considering structural features and their configurations for recognition. Characters and digits are special forms of objects, so there is a hope

[1] http://riesenhuberlab.neuro.georgetown.edu/hmax.html.

Published online: 12 August 2009

that features employed by this model may also show high performance over digit and character recognition. We aim to investigate the applicability of these features for handwritten digit recognition. We also systematically analyze invariance properties of features mediated by HMAX which has not been studied before over this problem.

The study of handwritten recognition covers a broad field dealing with numerous aspects of this very complex task. It involves research concepts from several disciplines such as experimental psychology, neuroscience, computer science, anthropology and education. Handwritten digit recognition is a subclass of handwritten recognition problem and has been very successful in the recent years. Applications such as postal mail sorting, recognizing address blocks on tax forms, reading general text, bank checking processing and form data entry have been widely adopted [6,7]. Due to various practical conditions, robust handwritten digit recognition techniques in terms of stability to common types of image noise and distortions are still desirable. Before handwritten digits can be recognized, they have to be processed in many steps: scanning to grayscale image, converting to binary image, feature extraction, classification and then post-decision making. Selection of a feature extraction method is critical in getting high handwritten recognition rate. Several methods have been proposed such as: direct matching [8], zoning [9], geometric moment invariants [10], Fourier descriptors [11], boundary-based features [12] and Zernike moments [13]. Each of these methods has its own advantages in one or more applications. For a comprehensive review of the recent handwritten digit recognition approaches over English and Farsi languages, the interested reader is referred to [14–19] and [20–23], respectively.

Gabor filters as models of V1 neurons [24] have received considerable attention in image processing for applications such as texture analysis [25], iris [26] and fingerprint recognition [27]. They have also been used for handwritten digit recognition [15]. While Gabor functions extract simple orientation features, HMAX and models based on it derive complex structural features built upon simple Gabor filters. Therefore, these features might have better performance than Gabor filters for handwritten digit recognition which we study in this paper over standard English and Farsi datasets. We also compare invariance properties of features derived by two variations of the HMAX model, Serre et al. [28] and Mutch et al. [29] over translation, rotation and scaling distortions. Robustness of these models to Gaussian and Salt & Pepper noises is also investigated.

The rest of this paper is organized as follows. A brief literature of object recognition models based on HMAX is reviewed in Sect. 2. Modified C2 feature model is explained in Sect. 3. It then describes experiments carried out to test and compare these models under different image transformations and distortions in Sect. 4. Results are discussed and concluded in Sect. 5.

2 Object recognition models based on HMAX

Modeling response properties of neurons in early visual areas has resulted to several applications in image processing and computer vision. For example, Gabor filters [24] have been widely used for edge detection and texture processing. In spite of the great familiarity with processes in the early visual areas, less abstract information is yet available on the functionality of higher vision which seems to be more useful for doing higher cognitive tasks. This work is intended to investigate the applicability of complex operations in higher visual areas for handwritten digit recognition.

Biologically inspired object recognition models mimic similar organization of the visual cortex, originally discovered by Hubel and Wiesel in the 1950s [24,30]. They found separate populations of simple and complex cells in primary visual cortex (V1), of which some cells (classified as "simple") exhibited strong phase dependence, whereas others (classified as "complex") did not. Hubel and Wiesel proposed that the invariance of complex cells can be formed by pooling simple cells with similar selectivities but with translated receptive fields. Perrett and Oram in 1993 [31] proposed a similar mechanism within IT to achieve invariance to any transformation by pooling afferents tuned to transformed versions of the same stimuli.

Based on these hypotheses, Fukushima presented the Neocognitron [32] which consists of a series of S and C layers mimicking simple and complex cell types, with shared weights for a set of local receptive fields and a competitive Hebbian learning rule. While it was originally invented for handwritten character recognition, it has also been used for other two-dimensional (2D) pattern classification tasks [33].

In 1998, LeCun et al. [34] introduced Convolutional Networks that generates local feature descriptors through back-propagation. Convolutional Networks has the same structure as Neoconitrons, but its basic operation in S layers is convolution. An S layer is generated by convolving the previous layer with d local filters, while each feature map of a "C" layer is generated from the corresponding map in the previous "S" layer via convolution with a fixed local filter. It has been applied to commercial-level character recognition, speech recognition, and face/object recognition.

In 1999, Riesenhuber and Poggio [2] formulated the "standard model" of object recognition in cortex which has been a basis for subsequent models. This model is composed of a hierarchy of feed-forward layers of neuron-like units, performing either a tuning computation, to increase feature complexity or a nonlinear pooling operation based on a

maximum operation to achieve invariance to translation and scaling. They also created a quantitative model [2], later called "HMAX", which was designed to account for tuning and invariance properties of neurons in IT cortex.

In its simplest form, the standard HMAX model consists of four layers of computational units, labeled as S1, C1, S2, and C2, where simple "S" units alternate with complex "C" units. The S units use convolution with local filters to compute higher-order features by combining different types of units in the previous layer (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 scaling and translation. Complex "C" layers increase invariance by pooling units of the same type in the previous layer over limited ranges. At the same time, the number of units is reduced by sub-sampling. Thus, the HMAX model performs a series of weighted-sum template-matching operations in the S layers and maximum-pooling operations in C layers to progressively build up feature complexity and invariance to scaling and translation.

The HMAX model, in its original form, generates a fixed set of features at the C2 layer which are insufficiently complex or distinct for object recognition tasks. No learning mechanism is employed in layers of this model. Later, Serre et al. [28] revised the model and applied it to a number of challenging recognition tasks. They proposed a new model for image categorization by adding to the HMAX a mechanism for learning intermediate-level shared features and changing the original Gaussian derivative filter bank by a Gabor filter bank. They argue that the Gabor filter is much more suitable in order to detect local features. In a recent study, Mutch et al. [29] proposed a modified model based on Serre's model by adding more biologically plausible operations to it such as sparsification of features, lateral inhibition and localization which are explained in detail in the next section.

3 Modified C2 feature model

The overall structure of a classifier for handwritten character recognition using modified C2 features is illustrated in Fig. 1. Input images are reduced to feature vectors, which are then used to train and test a classifier.

For building modified C2 feature model at first a base model, which performs at the level of the model described in [28], is created and then some improvements using sparsification of features and a form of lateral inhibition is added to the base model. We first discuss the base model and its differences from Serre's model. Then several changes to the base model which improve its performance are described.

3.1 Base model

Base model is constructed from one initial image layer and four alternating simple and complex cell layers. These four layers are built from alternating template matching and maximum pooling operations. Structure of the base model is shown in Fig. 2. Details of each layer of the model are briefly discussed as follows.

3.1.1 Image layer

In this layer an input image is transformed into an image-pyramid with ten scales. For creating the pyramid, the input image is converted to grayscale and the shorter edge is scaled to 140 pixels while maintaining the aspect ratio, then using bicubic interpolation, an image pyramid of ten scales, each a factor of $2^{1/4}$ smaller than the previous is created.

One of the differences between the base model and Serre's model [28] is that in base model an image pyramid is created and image width is always scaled to 140, but in Serre's model a pyramid approach is not used.

3.1.2 S1 layer

This layer corresponds to V1 simple cells in visual cortex and is computed from the image layer by centering 2D Gabor filters with full range of orientations at each possible position and scale. Like the image layer, S1 layer has three-dimensional (3D) pyramid shape, but each position and scale has multiple orientation units, so S1 layer is a four-dimensional structure. Each unit represents the activation of a particular Gabor filter centered at that position/scale. Gabor filter is described by Eq. (1):

$$G(x, y) = \exp\left(\frac{x^2 + \gamma^2 Y^2}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda} X\right) \tag{1}$$

where $X = x\cos\theta + y\sin\theta$ and $y = -x\sin\theta + y\cos\theta$. Parameters γ (aspect ratio), σ (effective width), and λ

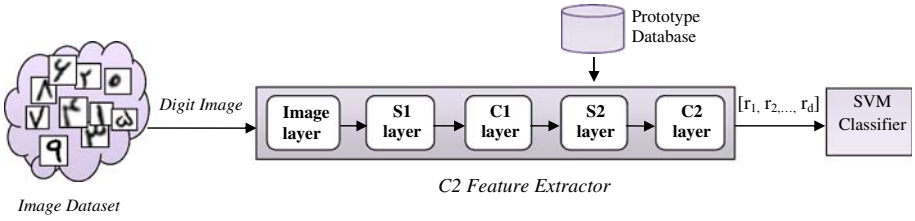

Fig. 1 Schematic view of the proposed handwritten digit recognition system. Test and train images are reduced to C2 feature vectors and are then classified by a SVM classifier

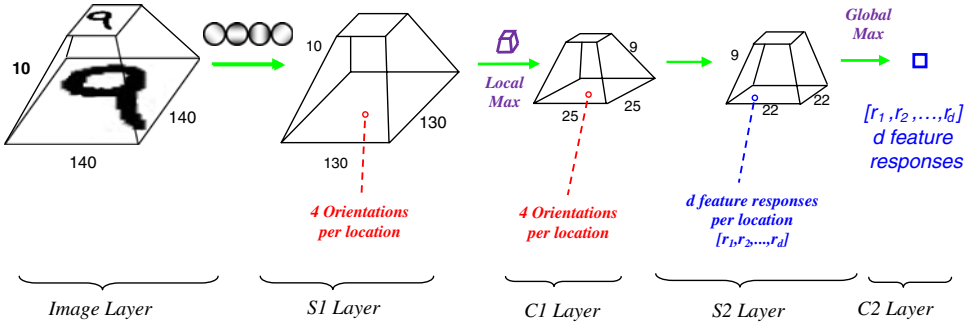

Fig. 2 Architecture of the base model. It contains five layers. Each layer is derived from its previous layer by applying template matching or max pooling filters and has 3D spatial dimensions at each of the locations. In image layer there is only one value at each location, but other layers have more than one value at each location

(wavelength) are all taken from [29] and are set to 0.3, 4.5, and 5.6, respectively. Components of each filter are normalized to zero mean. Response of a patch of pixels X to a particular S1 filter G is given by Eq. (2):

$$R(X, G) = \left| \sum X_i G_i \times \sqrt{\sum X_i^2}^{-1} \right| \tag{2}$$

where X_i and G_i are ith pixels of patches X and G, respectively and their product implement a pixel-wise multiplication.

In spite of Serre's model which applies different-sized S1 Gabor filters to full-scale image, base model applies only one Gabor filter of size (11×11) for all scales. Therefore, it results in lower computational complexity.

3.1.3 C1 layer

This layer provides a model for V1 complex cells in visual cortex. In this layer, a max-like pooling is used which builds position and scale tolerant C1 units and reduces number of units by sub-sampling. For each orientation, S1 pyramid is convolved with a 3D max filter, $m \times m$ units across in position and n units deep in scale, with m and n being parameters of the model. The max filter is moved around the S1 pyramid in steps of 5 pixels in position and value of C1 unit is simply obtained by taking the maximum of values of S1 units that fall within the max filter. Resulting pyramid is spatially smaller than the pyramid in S1 layer but number of features per position is still the same.

In spite of the Serre's model which C1 sub-sampling ranges do not overlap in scale, in the base model C1 sub-sampling overlaps with factor 2 in both position and scale.

3.1.4 Feature learning stage

The learning process corresponds to selecting d prototypes for next layer (S2) units. This is done using a sampling process, such that only during training, a large pool of d patches of various sizes at random positions and scales are extracted from the C1 layers of random training images. Values of all C1 units within that patch are read out and stored as a dense prototype. For a $n \times n$ patch, this means n^2 different positions, but for each position, there are units representing each of four orientations.

3.1.5 Intermediate feature (S2) layer

This layer is intended to correspond to cortical areas V4 or posterior IT. Values of S2 units are calculated by performing template matching using normalized RBF functions between patch of C1 units centered at that position/scale and d prototype patches. As mentioned in the learning stage, each prototype patch is sampled from the C1 layer of a training image at random positions and scales. S2 pyramid has the same number of positions/scales as C1 pyramid, but has d types of units at each position/scale, each one representing the response of the corresponding C1 patch to a specific prototype patch (Fig. 2).

A normalized Gaussian kernel (radial basis) function is used for measuring the response of a patch of C1 units X to a particular S2 prototype P:

$$R(X, P) = \exp\left(-\frac{\|X - P\|^2}{2\sigma^2 \alpha_\upsilon}\right) \tag{3}$$

The standard deviation σ is set to 1 in all experiments. Both X and P have dimensionality $n_\upsilon \times n_\upsilon \times 4$ where $n = 4, 8, 12, 16$ and $\upsilon = 1, 2, 3, 4$. Parameter $\alpha_\upsilon = (\frac{n_\upsilon}{\min(n_\upsilon)})^2$ is a normalizing factor which is used for different patch sizes to reduce weighs of extra dimensions [29].

3.1.6 Global invariance (C2) layer

The final layer of the model corresponds to the IT area in visual cortex. In this layer a "bag of features" with d dimensions is generated by taking global maximum over all scales and positions for each S2 map.

3.1.7 SVM classification

After C2 feature vectors of test and train images are obtained, first they are normalized to zero for the mean and one for the variance and then they are classified using an all-pairs

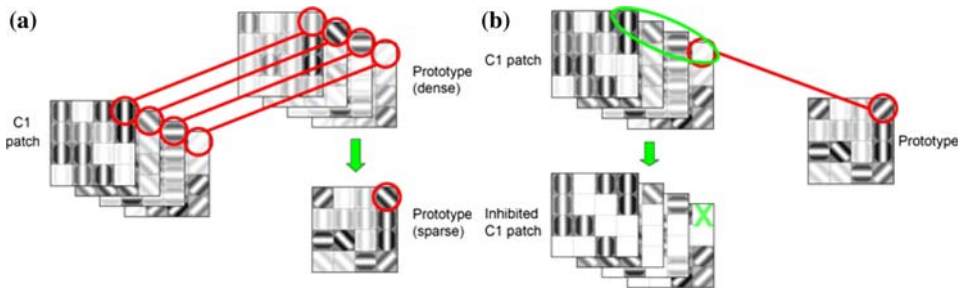

Fig. 3 **a** An illustration of the difference between dense and sparse S2 features. Sparse S2 prototypes are sensitive only to a particular orientation at each position, whereas dense S2 prototypes are all orientations of C1 units at each position. **b** The schematic view of inhibition in S1/C1. Before inhibition, the *circled unit* in the prototype patch is getting some response to its desired orientation; despite the fact other orientations dominate. Inhibition increases the distance to prototype patches looking for non-dominant orientations. A 4×4 S2 feature for a four-orientation model is shown here with stronger unit responses shown darker

multi-class linear SVM. Training images are first used to build the SVM and then the test images are assigned to digit classes using the majority-voting method.

3.2 Improvements to the base model

Modified C2 feature model has incorporated several enhancements to the base model to improve its performance. They are briefly described in the following section.

3.2.1 Sparsification

To increase the sparsity among inputs of a S2 unit, Mutch et al., selected fewer number of features from C1 layer by choosing the most dominant response in each position instead of using all four (number of orientations) responses. This is in accordance with the fact that neurons in visual cortex are more selective to a subset of their inputs. This results in reducing number of features from $4n^2$ to n^2. Since number of features is reduced, information containing combinations of different filter responses are lost. To recover the loss, number of orientations is increased from 4 to 12, having finer gradient over specificity (see Fig. 3a).

3.2.2 Lateral inhibition

Lateral inhibition means that cells in visual cortex inhibit their less-active neighbors in a winner-take-all competition. This modification is intended to correspond to this concept. It ignores non-dominant orientations and focuses on suppressing S1 and C1 outputs. Different orientations at the same position and scale in S1/C1units are encoded. These units compete to describe the dominant orientation at their location. To model this, at each location minimum and maximum responses, $R_{\min}$ and $R_{\max}$, over all orientations are computed and for each response R, if $R < R_{\min} + h(R_{\max} - R_{\min})$ then it is set to zero. Variable h is a global parameter to define the inhibition level, which represents the fraction of the response range that is suppressed (see Fig. 3b).

3.2.3 Limited C2 position/scale invariance

Serre's model suffers from co-occurrence of features from different objects and/or background clutter. Modified model, retains some geometric information above the S2 level. This is inspired by the finding that neurons in V4 and IT do not exhibit full invariance and are known to have receptive fields limited to only a portion of the visual field and range of scales [35]. Similarly, this model restricts the region of the visual field in which a given S2 feature can be found, relative to its location in the image from which it was originally sampled, to $\pm t_p\%$ of image size and $\pm t_s\%$ scales, where t_p and t_s are model parameters [29].

3.2.4 SVM weighting

This modification eliminates useless S2 features that contain background and convey no information about a digit in order to improve classification performance. In the training phase, all S2 vectors of all images are classified by SVM classifier. During constructing all-pairs multi-class linear SVM, the S2 features with low weight are dropped.

4 Experiments and results

To assess the performance of the modified C2 feature model for isolated handwritten digit recognition, in this section we apply it to standard English and Farsi datasets. For Farsi, we conducted our experiments over a recently developed standard Farsi corpus [36]. For English, we used the most interesting dataset: the MNIST handwritten digit corpus which is widely used in the literature [34].

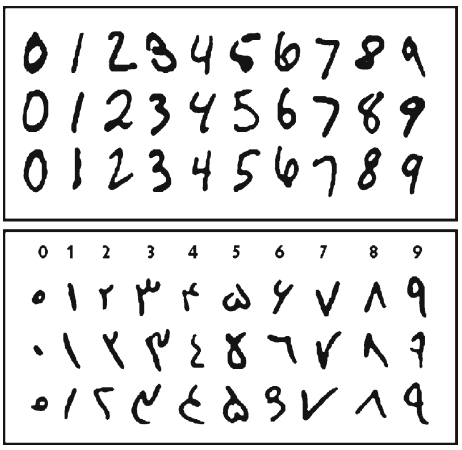

Fig. 4 Sample handwritten digits from MNIST (*top*) and Farsi (*bottom*) digit datasets

Table 1 Distribution of digits in MNIST and Farsi handwritten digit datasets

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

In all the experiments, we compared the performance of modified C2 feature model with the HMAX model of Serre et al. [28]. We used MATLAB® implementations of HMAX[2] and modified C2 features.[3]

4.1 Datasets

Khosravi et al. [36] have introduced a very large corpus of Persian handwritten digits.[4] This dataset contains 102,352 binary image digits which are scanned at 200 dpi resolution in 24-bit color format. From the whole number of digit images, 60,000 images were selected for train and 20,000 for test. Modified NIST (MNIST) dataset [34], contains 60,000 train and 10,000 test patterns.[5] Digits in the MNIST dataset are stored in images of 28 × 28 pixels and have intensities between 0 and 255. Figure 4 shows sample handwritten digits from MNIST and Farsi datasets. Distributions of digits in train and test sets of both datasets are shown in Table 1.

4.2 Classification results

In preprocessing step, all images of the Farsi dataset were converted to 64 × 64 pixels which is the size of the largest digit in this dataset. To resize an image, it was placed in center of a 64 × 64 black frame without scaling. Images of MNIST dataset are of size 28 × 28. To use the largest Gabor filter size of HMAX model which is 39, all images of this dataset were converted to 40 × 40 images. In the next preprocessing step, all images from both datasets were filtered using a 3 × 3 median filter in order to remove high-frequency noises. Then they were converted to binary images by replacing all pixels in the input image with luminance greater than *threshold level* (0.2) to 255 (white) and replacing all other pixels with 0 (black). Thus, each image in the dataset has a uniform background with a high-contrast digit. Using both models, for each input image a feature vector of 4,075 was derived. In both models Gabor filters in 12 orientations were used.

As in [29], parameters h, t_s and t_p were set to 0.5, ±5 and ±1%, respectively. Patch sizes in S2 layer were 4 × 4, 8 × 8, 12 × 12, and 16 × 16. Size of the 3D max filter was 10 × 10 units across in position and 2 units deep in scale.

After feature vectors were derived with HMAX, they were linearly normalized to zero mean and standard deviation one in each feature dimension, and all-pairs multi-class linear SVM classifier was used for classification. In modified C2 feature model, normalized feature vectors were classified again by all-pairs multi-class linear SVM. The reason we used the same classifier is because we want to make a fair comparison between two feature types. Results of SVM classification over both datasets are shown in Fig. 5. Results are for test sets and are averaged over five runs. Since training data is always fixed, the variability in results is due to random selection of patches in C1 layer.

As shown in Fig. 5, modified C2 features result in higher handwritten recognition over both datasets. Results over Farsi dataset are higher than MNIST using both feature types. This might be because of rich structural representations of Farsi digits.

[2] Implementation of HMAX model could be downloaded from http://resionhuberlab.neuro.georgetown.edu/hmax.html.

[3] Implementation of the modified C2 feature model could be downloaded from http://www.mit.edu/~jmutch/flib.

[4] Contact information for getting this dataset is available at http://www.modares.ac.ir/eng/kabir.

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

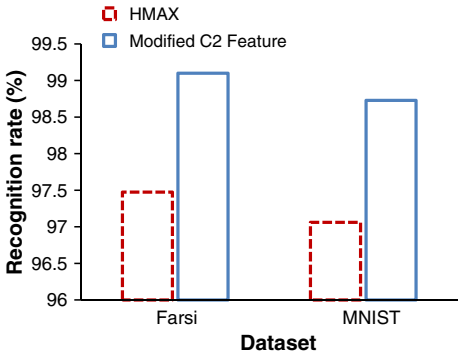

Fig. 5 Handwritten digit recognition results over Farsi and MNIST datasets with SVM classification

4.3 Invariance analysis

We conducted some experiments to investigate the stability of modified C2 features against rotation, scale and translation distortions. Performance of the model was also analyzed over Gaussian and Salt & Pepper noises. In experiments of this section, we used Gabor filters and 12 orientations in S1 layer of the modified C2 feature model. Results are averaged over five runs with random distortions over Farsi dataset.

4.3.1 Rotation invariance

In this experiment, training was done with original undistorted images and then the trained classifier was evaluated over distorted images. Twelve different degrees of rotation were tested: -90, -75, -60, -45, -30, -15, 15, 30, 45, 60, 75 and 90. For making a test set, at first each digit image was rotated and then was placed at the center of 64×64 black image. Each time, all test images were rotated the same degree. Classification results over rotated test datasets are illustrated in Fig. 6.

Comparing two feature extraction methods, it can be seen that they both are sensitive to rotations, but in most cases modified C2 feature model achieves higher performance than HMAX. As magnitude of rotation changes from 0 to +90 or -90 degrees, classification performance decreases. Since in modified C2 feature model a small number of units were part of the S1 column, a small rotation of the image does not change the winning unit. This to some extent explains higher tolerance of this model than HMAX.

4.3.2 Translation invariance

To examine the translation invariance, at first we generated 8 translated test datasets with different positions: -20, -15, -10, -5, 5, 10, 15, and 20. In each test dataset, every digit was shifted with the same number of pixels along the horizontal line at the center of the image. Then the generated datasets were tested using both HMAX and modified C2 feature models. Modified model was used with four different t_p values: 0.05, 0.3, 0.5, and 1%. Results of these tests are shown in Fig. 7.

Performance of the HMAX model does not change with translated images. This distortion tolerance has two reasons. First, in each train/test image digits are surrounded by a uniform background and second, in C2 layer the maximum response to each S2 feature is taken simply over all positions and scales, so results of the MAX operation were almost the same everywhere in the image.

In modified C2 feature model, performance decreases by increasing the magnitude of translation. Since in this model intermediate features (S2) are localized to small regions of an image, this results in limitation of position and scale invariance. This feature localization is suitable for datasets which contain background clutter or multiple objects. Images in the digit datasets contain uniform background, so it is better to decrease the feature localization. It can be done by setting

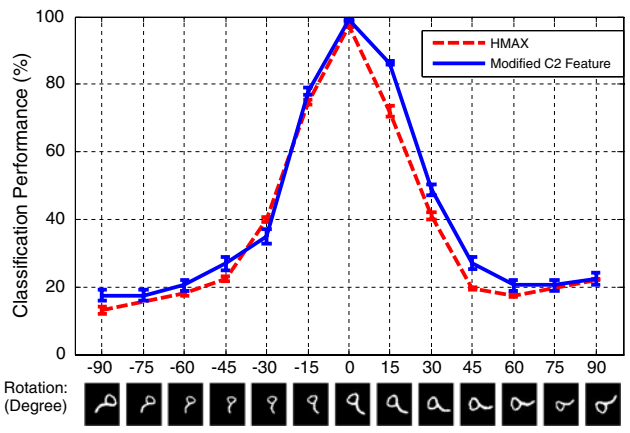

Fig. 6 Comparison between HMAX and modified model over different degrees of rotation

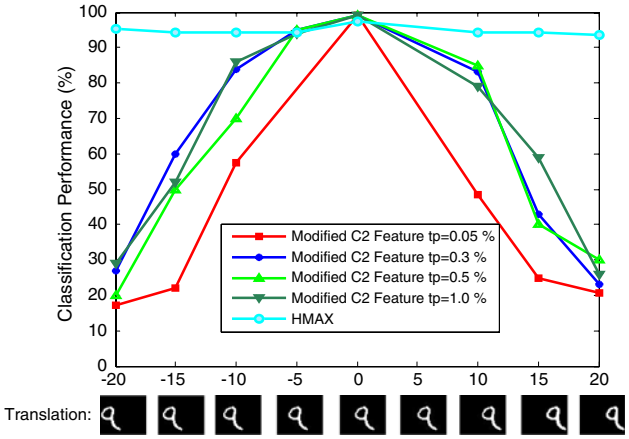

Fig. 7 Comparing HMAX and modified model over translation distortion

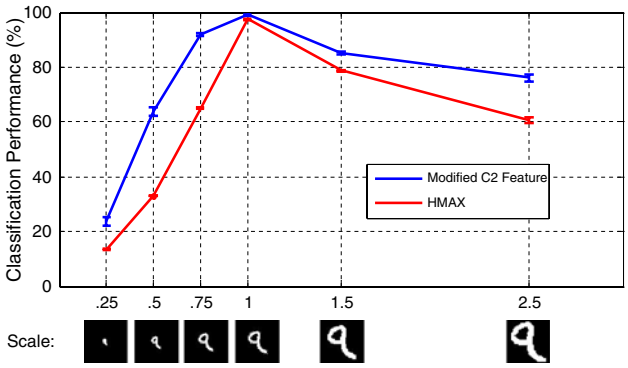

Fig. 8 Scale invariance analysis using both feature types

parameter t_p. Parameter t_p shows the percentage of positions used for taking maximum response to each S2 feature. In general, t_p shows the amount of the limitation of translation invariance. When the value of t_p is low, more geometric information retained and feature localization is increased. On the other hand when t_p has a high value, modified C2 feature model has more translation invariance.

Since in HMAX information about the exact spatial origin of the features is discarded in C layers it is not sensitive to translations. In contrast, modified C2 feature model is sensitive to rotations. As shown in Fig. 7, the t_p which indicates the amount of ignoring spatial locations increases, modified C2 features show increasing tolerance to translations. Therefore, it is appropriate to set a higher value (1%) for t_p for both feature types. This behavior depends on the background of the image dataset. If the images in the dataset have a cluttered background, it will be helpful to have more localization.

4.3.3 Scale invariance

In this section, scale invariance of both models is analyzed over five different scaled test datasets with scales as 0.25, 0.5, 0.75, 1.5 and 2.5. Comparison of results of two models is illustrated in Fig. 8. As it shows, performances of both models decrease by decreasing or increasing the scale. In all scales modified C2 features achieve higher performance than HMAX.

4.3.4 Analysis of robustness to noise

To evaluate the robustness to Gaussian and Salt & pepper noises, each time one type of noise was added to all images in the test dataset and like the previous experiments the train set was left unchanged.

For adding Salt & Pepper noise to images, seven noise densities were used: 0.001, 0.01, 0.05, 0.1, 0.3, 0.5, and 0.7. Results of both HMAX and modified C2 feature models over the dataset corrupted with Salt & Pepper noise are illustrated in Fig. 9. Modified C2 feature model has high recognition rate with small densities of Salt & Pepper noise, but loses its performance when noise density increases.

We also tested the robustness of HMAX and modified C2 feature models over 25 different Gaussian noises. These noises were generated at five different means (0, 0.05, 0.1, 0.5, and 0.7) and five different variances (0, 0.1, 0.25, 0.5, and 0.7). Figure 10 shows a sample image corrupted with Gaussian noises.

Classification rates of both models to Gaussian noise are shown in Fig. 11. As shown in this figure, classification performances decrease with increasing the mean and variance of Gaussian noise.

Comparing performances both feature types over Salt & Pepper and Gaussian noises show that modified C2 feature model is more robust than HMAX in almost all cases.

5 Discussions and conclusions

In this study, modified C2 feature model and HMAX were used for English and Farsi handwritten digit recognition. A set of scale—and translation—invariant C2 features was first extracted from a training set of digit images. A classifier was then constructed over these data and was evaluated over a separate test set. High handwritten digit recognition proves appropriateness of these features for the mentioned

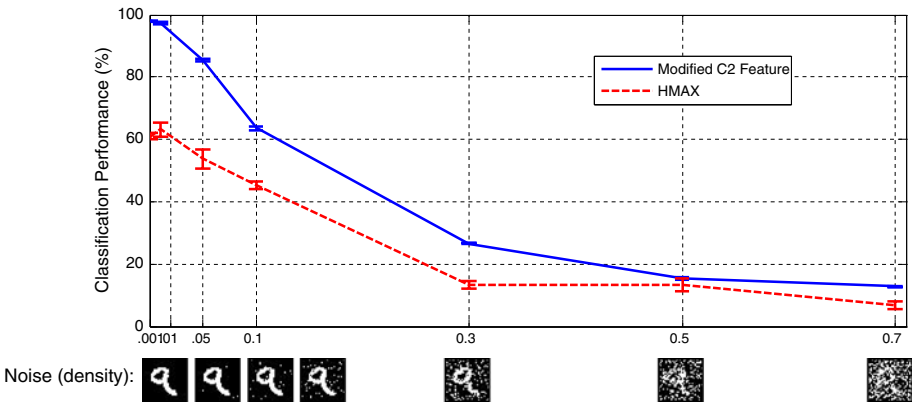

Fig. 9 Comparison between HMAX and modified model over images corrupted with Salt & Pepper noise

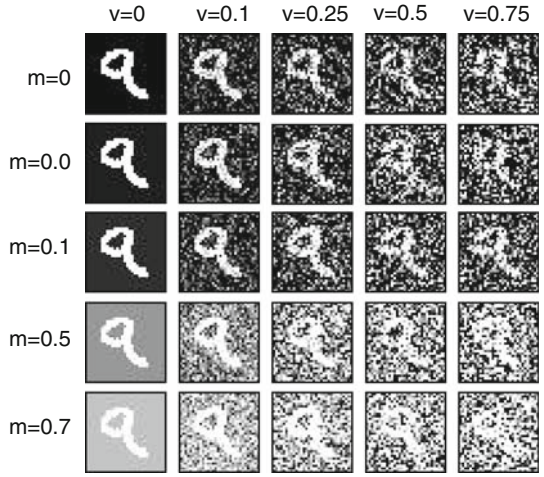

Fig. 10 A sample digit image corrupted with Gaussian noise. *Rows* show means and *columns* show variances of the noise

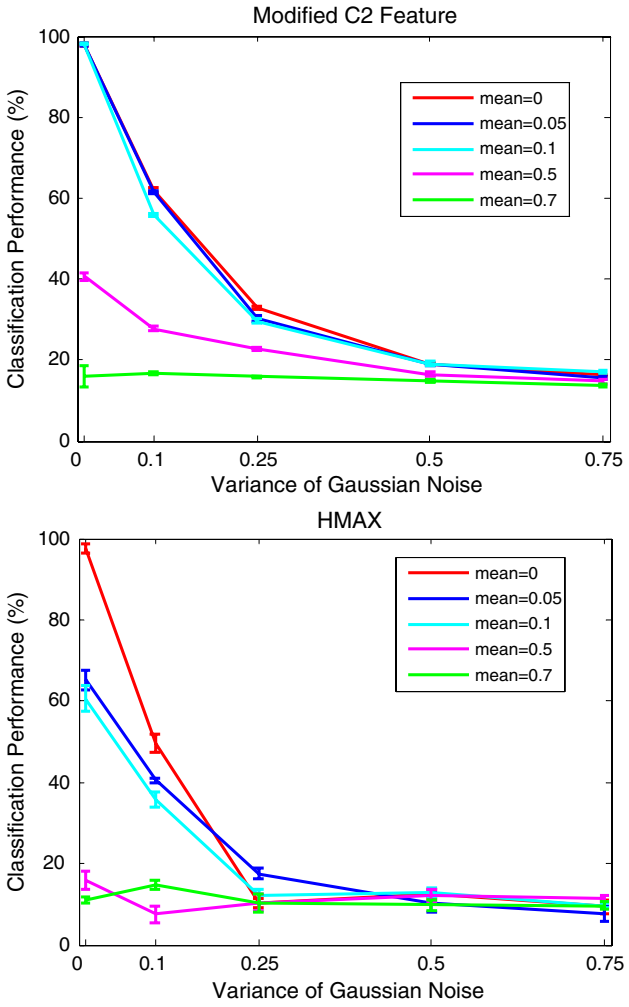

Fig. 11 Comparison of HMAX and modified model over images corrupted with Gaussian noise

task. Robustness analysis over Farsi dataset suggests high solidity of these features to common image distortions and noises. In most of the experiments, modified C2 features achieved better performance than HMAX.

Table 2 compares classification errors over MNIST dataset using modified C2 features and some leading previous methods in the literature. Our method has higher classification error compared with later methods but higher than famous LeNet-4. Compared with our previous study [23], here, we achieved lower recognition error using a single classifier over MNIST dataset.

In a similar study [37], authors have proposed a model with the same basics as the HMAX model and have incorporated a learning mechanism into it. By evaluating their method over MNIST dataset, they have achieved 94.2% recognition rate which is lower than our results. In [36], authors have used a multiple classifier system consisting of four MLP classifiers for digit recognition over the Farsi dataset used in this work. Using a modified gradient technique over 15,000 train and 5,000 test digits, they achieved 98.8% recognition rate which is below the 99.1% recognition rate we achieved using modified C2 features.

Table 2 Comparison on different methods for handwritten digit recognition over MNIST dataset

Method	Current study	Borji et al. [23] C2 features		Ranzato et al. [38]	Keysers et al. [39]	LeCun et al. [34]	Belongie et al. [40]
Features	Modified C2 features (all-pair multiclass SVM classifier)	Single classifier (SVM polynomial kernel)	Cascade classifier (SVM polynomial kernel)	Large conv. net (random features)	Non-linear deformation (kNN)	Conv. net LeNet-4 (local learning in last layer)	Shape context matching (kNN)
Recognition error (%)	1.27	3.5	1.1	0.89	0.54	1.4	0.63

Application of C2 features for character and digit recognition over other languages such as Chinese and Japanese is an interesting research area. Characters in these languages contain complex structural features which make these models appropriate for classifying them. One might also consider use of these features for biometric applications such as person identification based on palm-print, iris or finger-print biometrics.

References

1. Logothetis, N.K., Pauls, J., Poggio, T.: Shape representation in the inferior temporal cortex of monkeys. Curr. Biol. **5**, 552–563 (1995)
2. Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nat. Neurosci. **2**(11), 1019–1025 (1999)
3. Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., Poggio, T.: A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. Technical Report CBCL Paper #259/AI Memo #2005-036, Massachusetts Institute of Technology, Cambridge, MA, October 2005
4. Serre, T., Oliva, A., Poggio, T.A.: A feedforward architecture accounts for rapid categorization. Proc. Natl. Acad. Sci. USA **104**(15), 6424–6642 (2007)
5. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Object recognition with cortex like mechanisms. IEEE Trans. PAMI **29**(3), 411–426 (2007)
6. Govindaraju, V., Shekhawat, A., Srihari, S.N.: Interpretation of handwritten addresses in US mail stream. In: The Third International Workshop on Frontiers in Handwriting Recognition, Buffalo, New York, pp. 197–206 (1994)
7. Belaid, Y., Belaid, A., Turolla, E.: Item searching in forms: application to french tax form. In: International Conference on Document Analysis and Recognition, Montreal, Canada, pp. 744–747 (1995)
8. Gader, P.D., Forester, B., Ganzberger, M., Gillies, A., Mitchell, B., Whalen, M., Yocum, T.: Recognition of handwritten digits using template and model matching. Pattern Recognit. **24**(5), 421–431 (1991)
9. Freitas, C.O.A., Oliveira, L.E.S., Bortolozzi, F., Aires, S.B.K.: Handwritten character recognition using non-symmetrical perceptual zoning. Int. J. Pattern Recognit. Artif. Intell. **21**(1), 135–155 (2007)
10. Wang, D., Xie, W.: Invariant image recognition by neural networks and modified moment invariants. In: Proceedings of SPIE'96 (1996)
11. Kauppinen, H., et al.: An experimental comparison of autoregressive and Fourier-based descriptors in 2D shape classification. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, no. 2 (1995)
12. Di Zenzo, S. et al.: Optical recognition of hand-printed characters of any size, position, and orientation. IBM J. Res. Dev. **36**(3), 487–501 (1992)
13. Belkasim, S.O. et al.: Shape recognition using Zernike moment invariants. Asilomar Conf. Circuits **1**, 161–171 (1989)
14. Shi, M., Fujisawa, Y., Wakabayashi, T., Kimura, F.: Handwritten numeral recognition using gradient and curvature of gray scale image. Pattern Recognit. **35**, 2051–2059 (2002)
15. Wang, X., Ding, X., Liu, C.: Gabor filters-based feature extraction for character recognition. Pattern Recognit. **38**(3), 369–379 (2005)
16. Liu, C.L. et al.: Handwritten digit recognition: benchmarking of state-of-the-art techniques. Pattern Recognit. **36**, 2271–2285 (2003)
17. Teow, L.N., Loe, K.F.: Robust vision-based features and classification schemes for off-line handwritten digit recognition. Pattern Recognit. **35**, 2355–2364 (2002)
18. Cheung, K., Yeung, D., Chin, R.T.: A Bayesian framework for deformable pattern recognition with application to handwritten character recognition. IEEE Trans. Pattern Anal. Mach. Intell. **29**(12), 1382–1388 (1998)
19. Tsang, I.J., Tsang, I.R., Dyck D.V.: Handwritten character recognition based on moment features derived from image partition. In: International Conference on Image Processing 2, pp. 939–942 (1998)
20. Soltanzadeh, H., Rahmati, M.: Recognition of Persian handwritten digits using image profiles of multiple orientations. Pattern Recognit. Lett. **25**(14), 1569–1576 (2004)
21. Said, F.N., Yacoub, R.A., Suen, C.Y.: 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 (1999)
22. Sadri, J., Suen, C.Y., Bui, T.D.: Application of support vector machines for recognition of handwritten Arabic/Persian digits. In: Proceedings of Second Iranian Conference on Machine Vision and Image Processing 1, pp. 300S–307S (2003)
23. Borji, A., Hamidi, M., Mahmoudi, F.: Robust handwritten character recognition with features inspired by visual ventral stream. Neural Process. Lett. **28**(2), 97–111 (2008)
24. Hubel, D.H., Wiesel, T.N.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. **160**, 106–154 (1962)
25. Grigorescu, S.E., Petkov, N., Kruizinga, P.: Comparison of texture features based on Gabor filters. IEEE Trans. Image Process **11**(10), 1160–1167 (2002)
26. Horapong, K., Thainimit, S., Areekul, V.: Iris texture analysis with polar based filtering: preliminary results. In: Proceedings of the First Electrical Engineering/Electronics, Computer, Telecommunications, and Information Technology (ECTI) Annual Conference, 13–14 May, pp. 250–253 (2004)
27. Lee, C.J., Wang, S.D.: Fingerprint feature reduction by principal Gabor basis function. Pattern Recognit. **34**(11), 2245–2248 (2001)
28. Serre, T., Wolf, L., Poggio, T.: Object recognition with features inspired by visual cortex. In: CVPR, San Diego, June 2005
29. Mutch, J., Lowe, D.G.: Multiclass object recognition with sparse, localized features. In: CVPR, New York, June, pp. 11–18 (2006)
30. Hubel, D.H., Wiesel, T.: Receptive fields of single neurons in the cat's striate cortex. J. Physiol. **148**, 574–591 (1959)
31. Perrett, D.I., Hietanen, J.K., Oram, M.W., Benson, P.J.: Organization and functions of cells responsive to faces in the temporal cortex. Philos. Trans. R. Soc. Lond. **335**, 23–30 (1992)
32. Fukushima, K.: Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern. **36**(4), 193–202 (1980)
33. Fukushima, K., Wake, N.: Handwritten alphanumeric character recognition by neocognition. IEEE Trans. Neural Netw. **2**(3), 355–365 (1991)
34. LeCun, Y., Bottou, L., Bengio, Y., Haffner, Y.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
35. Rolls, E.T., Deco, G.: The computational neuroscience of vision. Oxford University Press, New York (2001)
36. Khosravi, H., Kabir, E.: Introducing a very large dataset of handwritten Farsi digits and a study on their varieties. Pattern Recognit. Lett. **28**(10), 1133–1141 (2007)
37. Oberhoff, D., Kolesnik, M.: Unsupervised shape learning in a neuromorphic hierarchy. Pattern Recognit. Image Anal. **18**(2), 314–322 (2008)
38. Ranzato, M., Huang, F.J., Boureau Y.L., LeCun, Y.: Unsupervised learning of invariant feature hierarchies with applications to object

recognition. In: Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR'07) (2007)
39. Keysers, D., Deselaers, T., Gollan, C., Ney, H.: Deformation models for image recognition. IEEE Trans. Pattern Anal. Mach. Intell. **29**(8), 1422–1435 (2007)
40. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape context. IEEE Trans. Pattern Anal. Mach. Intell. **24**(24), 509–522 (2002)