Research Article

Journals
The Institution of
Engineering and Technology

Enhanced hierarchical model of object recognition based on a novel patch selection method in salient regions

ISSN 1751-9632
Received on 8th August 2014
Revised on 3rd February 2015
Accepted on 13th April 2015
doi: 10.1049/iet-cvi.2014.0249
www.ietdl.org

Yan-Feng Lu[1], *Tae-Koo Kang*[2], *Hua-Zhen Zhang*[1], *Myo-Taeg Lim*[1] ✉
[1]*School of Electrical Engineering, Korea University, Seoul, South Korea*
[2]*Dept. of Information and Communication Engineering, Sangmyung University, Cheonan, South Korea*
✉ *E-mail: mlim@korea.ac.kr*

Abstract: The biologically inspired hierarchical model for object recognition, Hierarchical Model and X (HMAX), has attracted considerable attention in recent years. HMAX is robust (i.e. shift- and scale-invariant), but its use of random-patch-selection makes it sensitive to rotational deformation, which heavily limits its performance in object recognition. The main reason is that numerous randomly chosen patches are often orientation selective, thereby leading to mismatch. To address this issue, the authors propose a novel patch selection method for HMAX called saliency and keypoint-based patch selection (SKPS), which is based on a saliency (attention) mechanism and multi-scale keypoints. In contrast to the conventional random-patch-selection-based HMAX model that involves huge amounts of redundant information in feature extraction, the SKPS-based HMAX model (S-HMAX) extracts a very few features while offering promising distinctiveness. To show the effectiveness of S-HMAX, the authors apply it to object categorisation and conduct experiments on the CalTech101, TU Darmstadt, ImageNet and GRAZ01 databases. The experimental results demonstrate that S-HMAX outperforms conventional HMAX and is very comparable with existing architectures that have a similar framework.

1 Introduction

Object recognition has been widely used in the visual navigation of robots, video surveillance and pedestrian detection [1–3]. In practical applications, the difficulties that arise in object recognition are typically caused by variations in the appearance of the objects and the background complexity of the input images. Object variability in terms of scale, rotation and illumination, especially in the case of cluttered backgrounds, seriously disrupts the recognition [4, 5]. For instance, various human postures (e.g. squatting, stooping, running or standing) in a real environment make accurate recognition a difficult task. Many algorithms have recently been proposed to address this issue.

Traditional appearance-based methods mainly employ low-level visual features such as colors, textures and edges [6, 7]. Although these methods generally take these features into account, they do not selectively address discriminative features. They are also sensitive to occlusion, deformation, scale and variations in illumination. Local feature-based methods combine local descriptors and keypoint detectors with spatial information. Representative local feature methods have been proposed, such as scale-invariant feature transform (SIFT) [8], gradient location and orientation histogram [9], histogram of gradients [10] and speeded up robust features [11]. These methods are effective in terms of describing locally discriminative features, but they lack oriented local information. Bag-of-words (BoW) [12] and bag-of-features [13] are effective for resolving this issue; however, the amount of structural information still falls short.

In recent years, significant advances have been made in the understanding of brain cognition in the biological vision field. The findings related to the primary visual cortex (area V1) are especially important. Althoughresearching V1, Hubel and Wiesel discovered that the visual system analyses patterns into multiple and independent channels that have various spatial frequencies and orientations [14]. This discovery gives biological support to early stage psychophysical theories. On the basis of these theories, Riesenhuber and Poggio [15] presented an initial computational model of object recognition, called Hierarchical Model and X (HMAX) that attempts to model the rapid object recognition mechanism of the cortex. Serre *et al.* [16] improved the original HMAX model significantly and proposed standard HMAX, demonstrating that the visual cognitive model efficiently enhances the performance of object recognition.

HMAX is an appearance-based feature descriptor that focuses on feature invariance and selectivity. It is robust to scale and shift deformations, but it shows sensitivity to rotational deformation [16, 17]. Improvement of the rotation invariance of local features is challenging, although recently some valid approaches have been proposed [18–20]. The robustness to rotation, which is improved in HMAX only by introducing rotated versions of the training images, is inadequate. Conventional HMAX uses patches that are randomly selected in the second (C1) layers, which generates a huge amount of redundant information and also prevents robustness against rotational deformation.

The stored patches in the C1 layers are the key components of the discriminative and robust abilities of HMAX. Superior features extracted by the stored patches determine the feature invariance and selectivity, preserving HMAX performance in the cases of object appearance variation and cluttered backgrounds. The majority of patches selected by the random method, however, are redundant and not discriminative for the recognition task, which results in performance degradation and high computational cost. These drawbacks seriously limit the overall performance of HMAX. We propose a solution to this issue, based on a novel patch selection method called saliency and keypoint-based patch selection (SKPS). SKPS is a patch selection method based on a saliency mechanism and multi-scale keypoints that aims to reduce the number of patches chosen but keep those with better discrimination than those chosen by random selection. Unlike standard HMAX with randomly selected patches, our method extracts patches in the C1 layers of HMAX by SKPS. We further propose a SKPS-based HMAX model (S-HMAX). We show its effectiveness, by applying it to object categorisation and by conducting experimental studies on the CalTech101, TU Darmstadt (TUD), ImageNet and GRAZ01 databases.

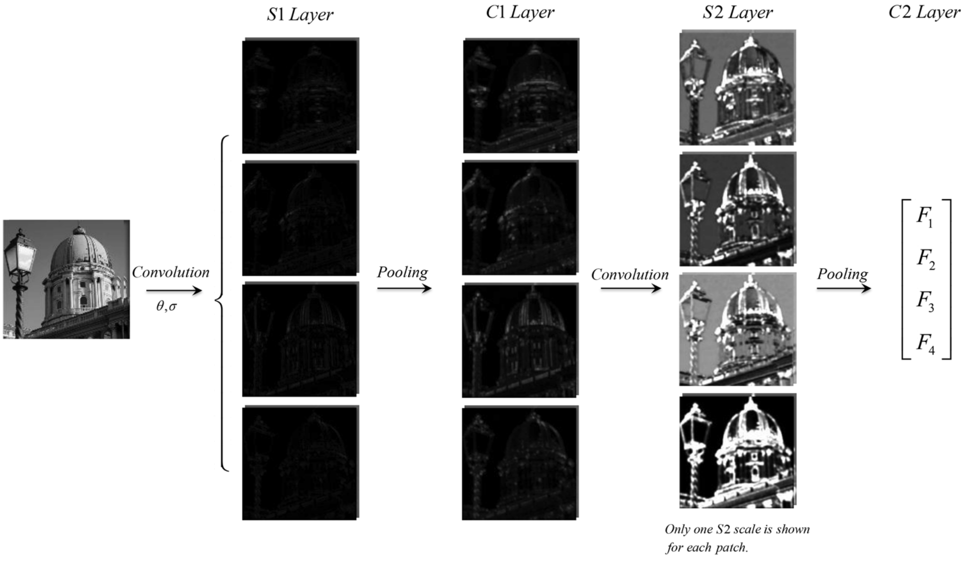

Fig. 1 *HMAX structure overview*

The rest of the paper is organised as follows. In Section 2, we briefly review the conventional HMAX model. In Section 3, we describe the SKPS method and S-HMAX model. In Section 4, we present experimental results based on four public databases. Finally, in Section 5, we give our conclusions.

2 HMAX review

The conventional HMAX [16] is a computational framework with four layers: S1, C1, S2 and C2, as shown in Fig. 1. The framework follows the mechanisms of the primary visual cortex (area V1) and builds feature representation by patch matching and maximum pooling operations. We briefly describe the operations of each layer as follows.

S1 layers: The units in the S1 layers correspond to simple cells in V1. The S1 units take the form of Gabor functions [21], that model cortical simple cell receptive fields. Gabor functions are defined as

$$G(x, y) = \exp\left(-\frac{x_o^2 + \gamma^2 y_o^2}{2\sigma^2}\right) \times \cos\left(\frac{2\pi x_o}{\lambda}\right)$$
$$\text{s.t} \quad x_o = x\cos\theta + y\sin\theta \text{ and } y_o = -x\sin\theta + y\cos\theta \qquad (1)$$

where θ represents orientation, λ is wavelength, σ is scale and γ indicates spatial aspect ratio.

Given an input image, the S1 layer with orientation θ and scale σ is calculated by

$$S1_{\sigma,\theta} = |G_{\sigma,\theta} * I| \qquad (2)$$

where * denotes convolution, I is the input image and $G_{\sigma,\theta}$ is a Gabor function with specific parameters.

C1 layers: These layers describe the complex cells in V1. The layers are the dimensionally reduced S1 layers obtained by selecting the maximum over local spatial neighbourhoods. This maximum pooling operation over local neighbourhoods increases invariance (providing some robustness to shift and scale transformations).

S2 layers: In these layers, S2 units pool over afferent C1 units from a local spatial neighbourhood across all four orientations. The S2 layers describe the similarity between the C1 layers and stored patches in a Gaussian-like way using Euclidean distance. The responses of the corresponding S2 layers are calculated by

$$S2 = \exp(-\beta \parallel C1(j, k) - P_i\|^2) \qquad (3)$$

where β is the sharpness of the exponential function, $C1(j, k)$ denotes the afferent C1 layer with scale j and orientation k and P_i is the i_{th} patch from the previous C1 layers.

C2 layers: The final set of shift- and scale-invariant C2 responses is computed by taking a global maximum of afferent S2 units across all scales and positions. The responses of the C2 layers are calculated by

$$C2 = \max_{(m,n,\sigma)}(S2(m, n, \sigma)) \qquad (4)$$

where (m, n) is the position of S2 units and σ denotes the corresponding scale. The output is a vector of N C2 values, where N corresponds to the number of patches. The vector is used as the C2 feature in the recognition task.

3 SKPS-based HMAX model

The HMAX model is an appearance-based feature descriptor that balances feature invariance and selectivity. HMAX is robust in object recognition, but it is sensitive to rotational deformation because many stored patches are not refined and discriminative with respect to rotation. We addressed this drawback by introducing a novel patch selection method, saliency and keypoint-based patch selection (SKPS) and proposing a SKPS-based HMAX model (S-HMAX). The following sections present the details of SKPS and S-HMAX.

3.1 Saliency and keypoint based patch selection

The stored patches in the C1 layers are the key components of the discriminative and robust abilities of HMAX; thus, the construction of a proper patch set is very important for the visual recognition task. A random selection of patches from the universal training set is an option, but that option is prone to bringing in huge amounts of redundant information and is sensitive to rotation. We address this by proposing a novel patch selection method, SKPS, which is based on a saliency mechanism and multi-scale keypoints. The method extracts fewer patches with better discrimination and invariance when compared with those obtained by random selection.

SKPS consists of the following five steps: (1) processing layer extraction, (2) finding salient regions, (3) keypoint candidate localisation, (4) selection of optimal keypoint candidates and (5) robustness enhancement. Fig. 2 shows the structure of SKPS in S-HMAX. The green dots are the location of keypoint candidates. Different scales of red circles indicate different strictness of the

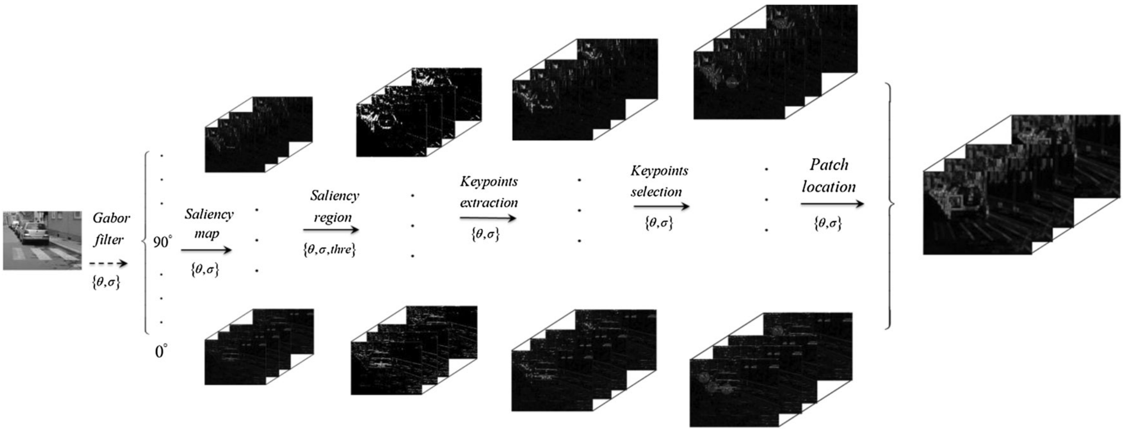

Fig. 2 *Overview of SKPS in S-HMAX*

location constraint. The red rectangle denotes the location of a selected patch.

3.1.1 Processing layers: Input images are processed by Gabor filters with different directions and scales. We obtain Gabor scale pyramids as per the method in [16]. We make the directional multi-scale information tractable, by considering four orientations and sixteen scales for further processing in HMAX. The processing layers can be calculated using (2).

3.1.2 Salient region: A huge amount of irrelevant information exists in the processing layers, which complicates locating the more discriminative regions in the whole image. Obtaining dense distinctive features requires the construction of a salient region with rich discriminative information. Based on a biological visual perception mechanism, attention is an important visual processing stage that guides the gaze towards objects of interest in a visual scene [22]. This ability to orientate towards salient objects in a cluttered visual environment is of great significance because it allows rapid and accurate detection and tracking of prey or predators by organisms in the visual world. Itti and Koch [23] first introduced a biologically inspired model to generate a saliency map. Recently, some valid saliency models were proposed [24, 25]. In our paper, the saliency map is constructed in the processing layers based on a simple saliency model in [26]. Some other saliency methods also can be employed in our framework. Given an input image I, the log spectrum L is calculated by

$$L(f) = \log\left(\left|F(I)\right|\right) \tag{5}$$

where F is the Fourier transform, f is frequency and log is a logarithm operation. The phase spectrum of the image can be computed by

$$P(f) = \text{angle}(F(I)) \tag{6}$$

The spectral residual $R(f)$ can be obtained by

$$R(f) = L(f) - \boldsymbol{h}_n * \left|F(I)\right| \tag{7}$$

where $\boldsymbol{h}_n$ is a matrix of all ones (in this study, a 3 × 3 ones matrix is used) and * denotes the convolution operation.

The saliency map of the input image can be calculated

$$\text{Sal}(I) = \left|\left(F^{-1}\left(\text{e}^{R(f)+i\cdot P(f)}\right)\right)^2\right| \tag{8}$$

We inhibit non-dominant information by adopting a simple version of the saliency map. We segment the constructed saliency map to obtain the salient region, that is, where the distinctive features and patch extraction areas are concentrated. Given the saliency map of the input image, the salient region at location (x, y) can be obtained

$$\text{SR}(x, y) = \begin{cases} 1 & \text{if } \text{Sal}(I(x, y)) > \text{threshold} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

In general, we set threshold = $M(\text{Sal}(I)) \times 2$, where $M(\text{Sal}(I))$ is the mean value of every pixel in the saliency map. (SKPS experimentally shows the best performance when threshold = $M(\text{Sal}(I)) \times 2$, therefore we chose this value). The construction of the salient region is illustrated in Fig. 3.

3.1.3 Keypoint candidate localisation: In the constructed salient regions, we locate the keypoint candidates in each layer with their corresponding direction. In the conventional HMAX model, patches are randomly extracted from the overall C1 layers to form the vocabulary of visual features. However, these visual features are neither refined nor discriminative; they include irrelevant and redundant information and degrade performance. Achieving a reasonable recognition performance with HMAX requires matching many patches, which results in high

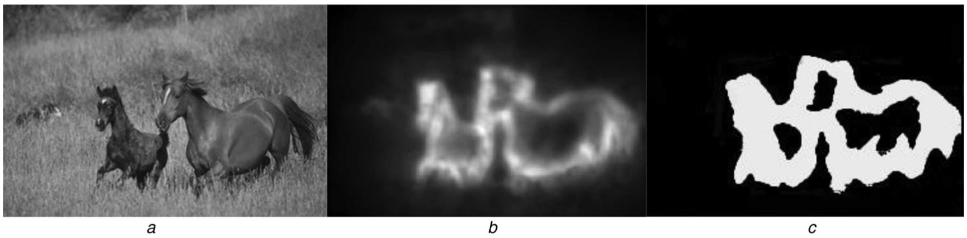

Fig. 3 *Construction of the salient region*

a Original image
b Saliency map of the input image
c Salient region

computational cost. SKPS locates the keypoint candidates within the salient region, which are identified by a keypoint detection method named FAST [27]. FAST is widely used because of its accuracy and speed; however, it does not have an orientation component nor does it produce multi-scale features. Hence, we employ processing layers that are processed by Gabor scale pyramids at certain angles and produce FAST keypoints at each layer. In this way, we extract multi-scale keypoint candidates with specific angles. The keypoint candidate position key can be localised by

$$\text{key} = \text{FAST}(P_{\theta,\sigma}(x, y)),\ (x, y) \in \text{SR} \tag{10}$$

Here, FAST is the keypoint detection method, $P_{\theta,\sigma}$ denotes the processing layer with orientation θ and scale σ, (x, y) are pixel coordinates in the layer and SR is the salient region. We preferentially extract image patches around these detected keypoint candidates.

3.1.4 Selection of keypoint candidates:

The keypoint candidates directly extracted by FAST require further preprocessing before recognition. The best keypoints are selected by weighting all the keypoint candidates. Each patch is composed of four layers that correspond to four different orientations; the keypoint candidates in the same location of each layer are weighted sums. The keypoint with the maximum weighted sum value determines the patch location.

A large number of keypoint candidates are extracted during keypoint extraction. Each patch is composed of four orientation layers; therefore determining a discriminative location depending on only one layer of the patch is unreasonable, but this easily generates repeatable candidates at near-duplicate locations. Therefore we use a balanced value based on the weighted sum over each layer of the patch. The keypoint candidates of each layer with corresponding direction are weighted summed. The weight is obtained based on non-maximal suppression, which is effective for edge and corner detection. However, it cannot be applied directly to the features. Thus, a score function W is computed for each detected keypoint. As the value of W increases, the number of detected keypoints decreases [28]. The strength of any keypoint p is defined to be the maximum value of W

$$W(p) = \max\left(\sum_{x\in R_+} \left|I_{p\to x} - I_p\right| - t,\ \sum_{x\in R_-} \left|I_p - I_{p\to x}\right| - t\right) \tag{11}$$

where

$$R_+ = \{x | I_{p\to x} \geq I_p + t\}$$
$$R_- = \{x | I_{p\to x} \leq I_p - t\}$$

The maximum of the sum of the absolute difference between the keypoint location and surrounding pixels is assigned to W. Here, t is a threshold, p is the label of keypoint, I_p is the value of the p_{th} keypoint pixel in input image, x is a pixel position around keypoint p, the pixel at position x relative to p is denoted by $p \to x$ and R is the adjacent area around the central keypoint p.

Using (12), we sum the scores at the location of the keypoint candidates for each layer. We then sort the aggregate scores of every keypoint candidate, and the top N keypoints are selected as the patch locations, where N is the number of patches used in the recognition task

$$\text{Loc} = \max \sum_{i=1}^{n} W_i \cdot Q_{\theta_i}, \quad \text{s.t.} \quad Q_\theta = \begin{cases} 1 & \text{num} > 0 \\ 0 & \text{num} = 0 \end{cases}$$

where Loc is the patch position, num is the keypoint number of the location determined by FAST, θ represents orientation, i is the index of the orientation and n is the number of orientations.

3.1.5 Robustness enhancement:

The robustness is improved by further processing of the location of the keypoints by robustness enhancement.

FAST records the location of the keypoint pixels, and imposes a strict restriction on keypoint locations. However, this strict restriction results in high discriminative power but poor robustness. Therefore we relax the restriction using keypoint expansion to determine a suitable trade-off between selectivity and invariance. One possible method for improving robustness is to loosen the location constraint in localisation measurement, that is, keypoints from adjacent locations could be also taken into account. The size of the extracted image patch corresponding to the k_{th} keypoint is defined as

$$R_k = r \cdot \text{key}_k \tag{13}$$

where key_k denotes the k_{th} keypoint location, R_k is the circular region centred at key_k and r is the radius of R_k that controls the strictness of the location constraint. We can therefore infer that r modulates the location constraint. A smaller radius corresponds to a smaller circular region and more accurate keypoint positioning, but causes poor feature invariance. A larger r corresponds to a larger circular region that contains more spatial clues, and thus improves robustness. However, too loose a constraint will degrade the accuracy of the keypoint positioning and incorrectly group nearby keypoints together.

The proposed keypoint expansion R_i is illustrated in Fig. 4. As shown in the figure, the keypoint candidate (red dot) is expanded in thickness to the neighbouring pixels in a circular region of radius r. Since other keypoint candidates are likely to exist in the circular region, the location of the patch can be calculated by

$$\text{Loc}_R = \max \sum_{j=1}^{m} \sum_{i=1}^{n} W_{i,j} \cdot Q_{\theta_{i,j}} \tag{14}$$

where m is the number of keypoint candidates in the circular region, n is the number of orientations, i is the index of the orientation, j is the index of the keypoint in the region and θ represents orientation.

3.2 SKPS-based HMAX

In contrast to conventional HMAX that selects patches randomly, our proposed method of S-HMAX uses SKPS to refine patches at the C1 layers of HMAX.

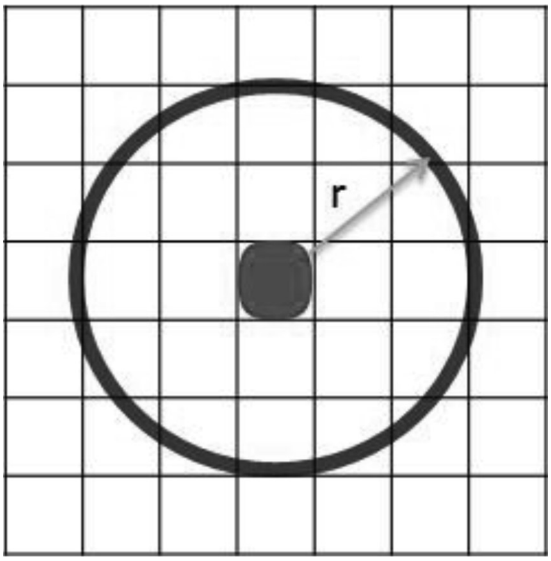

Fig. 4 *Circular region expansion of the keypoint candidate*

The random selection in HMAX results in extraction of a significant amount of redundant or irrelevant information (e.g. background or gaps in the input images) in the recognition task, which clearly degrades the performance of the HMAX model. In S-HMAX, the patches are mainly extracted from the salient region obtained by the biological attention mechanism; this salient region contains more discriminative features for recognition. We further find the locations with the most discriminative features in the salient region and extract the patches for S-HMAX from those locations. We also reserve 10% of the patches as a random choice, because of the complexity of the input image. The mechanism of random patch selection makes the proposed method adaptive and robust to practical applications. In this paper, we use 10% random patches, as this setting, gives a reasonable performance.

4 Experiments

We evaluated the performance of S-HMAX in several recognition tasks. In Section 4.1, we evaluate the S-HMAX model under rotational deformation using Caltech101. In Section 4.2, we evaluate the S-HMAX under normal circumstances using four datasets (TUD [29], Caltech101 [30], ImageNet [31] and GRAZ01 [32]). Owing to the large variations in the appearance of the input images, we utilised the scale and position-invariant C2 features [16], and passed these features to a linear classifier trained to perform the classification recognition task. (Both a linear kernel and a polynomial-kernel SVM were tested, and gave very similar results. We chose the linear Lib-SVM [33] as the classifier). Four orientations were set in advance, as for standard HMAX (e.g. 0°, 45°, 90°, and 135°). Except for the SKPS in the C1 layers, the other layers of S-HMAX are similar to those of conventional HMAX. We chose the evaluation metrics classification rate, recall and 1-precision, defined as follows

$$1 - \text{precision} = \frac{\text{number of false positives}}{\text{total number of positives}} \tag{15}$$

$$\text{recall} = \frac{\text{number of true positives}}{\text{number of true positives and false negatives}} \tag{16}$$

$$\text{classification rate} = \frac{\text{number of true positives and true negatives}}{\text{total number of positives and negatives}} \tag{17}$$

where a true positive is a correct classification of a positive (object), a true negative is a correct classification of a negative (background), a false positive is an incorrect positive classification and a false negative is an incorrect negative classification.

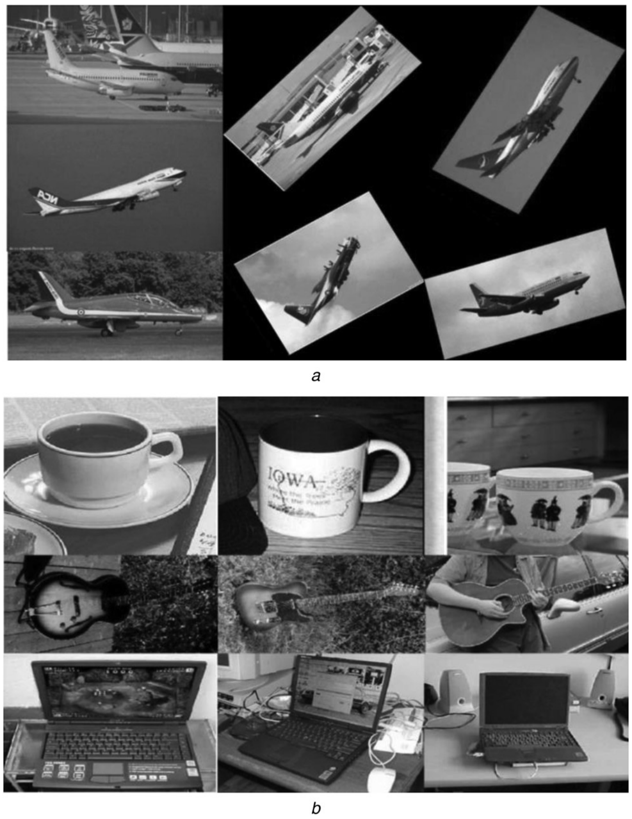

Fig. 5 *Examples of training and testing set images and sample categories for the experiment*

a Sampling images for training and testing (aeroplanes)
b Sampling images of other categories: cups, guitars and laptops

4.1 Comparison between HMAX and S-HMAX in local rotation

We evaluated the influence of local rotation on the output of the S-HMAX model by comparing the conventional HMAX model with the S-HMAX model under rotational deformation. The sampling categories of the CalTech101 database were used in this experiment. Sampling images of the training sets and testing sets are shown in Fig. 5. We trained the models with original images from the database. In the experiment, 15 original images (0°) of each sampling category were randomly selected as training sets, and 15 of their rotated versions with increasing amplitudes (15°, 30°, 45°, 60° and 90°) were used as testing sets. Twenty-five patches of HMAX were utilised in the experiment. In addition, 5, 10 and 25 patches of S-HMAX were utilised for the performance evaluation. The results of the experiment were averaged over 10 independent trials. We reported the mean and standard deviation of the classification rate across all cases.

As shown in Fig. 6, the S-HMAX models are more robust to rotational deformations. When the rotation is less than 30°, the recognition rate of HMAX and S-HMAX is comparable. However, when the rotation is over 30°, conventional HMAX shows a sharp decrease in performance in nearly all categories. In comparison, the S-HMAX models with 5, 10 and 25 patches have a more stable performance and higher classification rate. Note that HMAX with 25 patches does not have an obvious performance advantage even when compared with the S-HMAX model with five patches. On the contrary, S-HMAX with five patches has better classification results for large rotation deformations (over 30°). The performance of S-HMAX is clearly a significant improvement over that of conventional HMAX with the same number of patches. The results confirm that SKPS is an effective patch selection method for discriminative and invariant features. Furthermore, S-HMAX significantly improves performance compared with conventional HMAX in the case of rotation deformations and is a valid way to reduce the redundant information extracted by the HMAX model.

The S-HMAX model improved the performance by the use of salient regions and keypoints. To show the contribution of the two modifications, we evaluated the model by separating the two modifications into individual patch selection strategy. The conventional HMAX was used as our baseline model. 15 rotated images with the amplitude of 45° (it is most challenging for recognition in local rotation) were used as testing sets. We separately added the saliency strategy and keypoint strategy to the baseline model for the classification, and compared them with S-HMAX that combined these two together.

Table 1 shows the contribution to performance of our modifications in HMAX. The saliency strategy and keypoint strategy independently improve the performance of HMAX in local rotation. Combining the two strategies together presents a significant improvement in the local rotation.

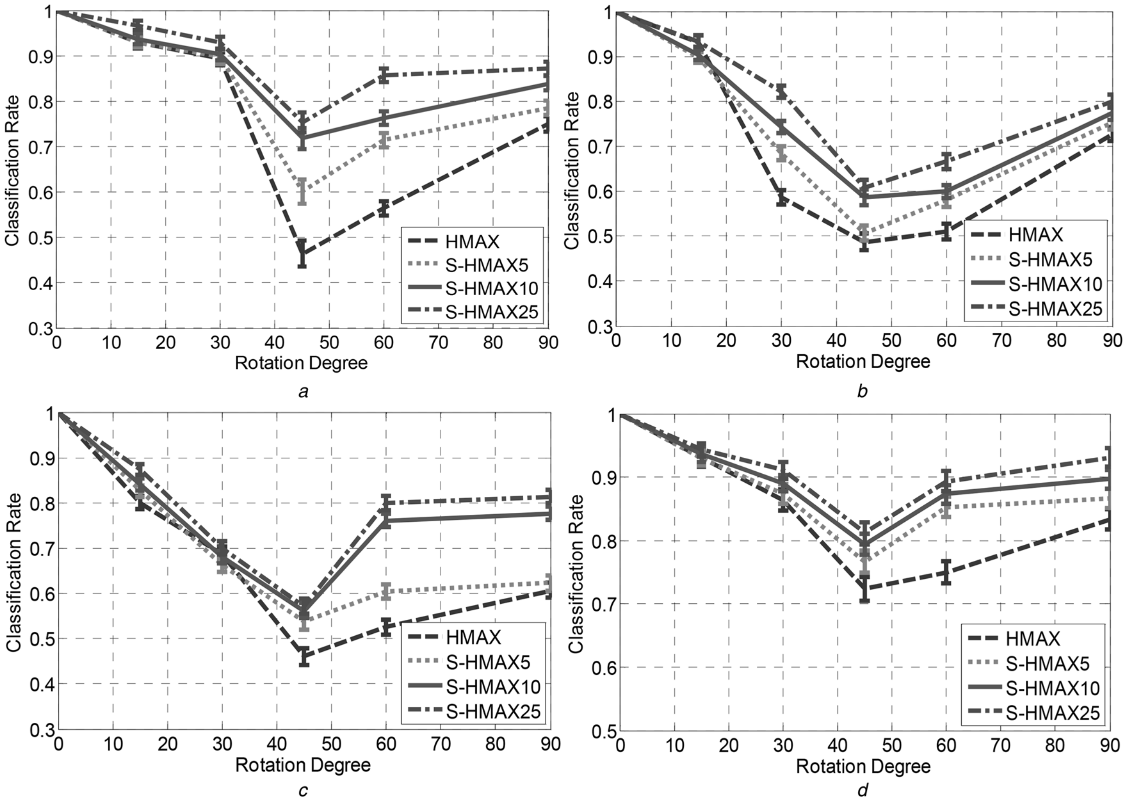

Fig. 6 *Comparisons of HMAX and S-HMAX with different number of patches in local rotation on categories from the Caltech101 database*

a Aeroplanes
b Laptops
c Guitars
d Cups

4.2 Object classification experiment

S-HMAX was shown in Section 4.1 to have a superior performance in the case of local rotation. We further evaluated S-HMAX by comparing its performance with that of other related algorithms on four public image datasets: TUD, CalTech101, ImageNet and GRAZ01.

4.2.1 TU Darmstadt: The TUD database (formerly the ETHZ database) contains side views of cars, motorcycles and cows, as shown in Fig. 7. We evaluated the S-HMAX model, the conventional HMAX that uses the random patch selection method and a modified HMAX model (M-HMAX) based on a maximum energy patch selection method [34]. In addition, we also compared SIFT [8] and spatial pyramid matching using sparse coding (Sc-SPM) [35] in the experiment.

To make the comparison at the feature level, we compared the scale- and position-invariant C2 features of HMAX models with the features produced by SIFT and Sc-SPM by passing them to a linear SVM that was trained to perform the object present/absent recognition task. We compared the classification rate for various numbers of features (5, 10 and 25). In the experiment, we randomly chose 15 images from each category of the TUD database as positive training images and 15 background images as the negative training set. For the tests, 50 images from each category of the TUD dataset and 50 images from backgrounds were randomly chosen as a test set. The results were generated from 10 independent trials. We report the mean and standard deviation of the classification across all classes.

Table 1 Contribution of our modifications to the classification in local rotation

Model version	Aeroplanes	Laptops	Guitars	Cups
Base	46.43	48.83	47.25	72.07
+saliency	65.71	54.76	53.33	75.66
+keypoint	67.29	55.12	52.17	77.85
+saliency&keypoint	78.57	61.85	57.63	82.67

Fig. 8 shows the simulation results on the TUD dataset for different numbers of features. In general, S-HMAX clearly outperforms SIFT, Sc-SPM, HMAX and M-HMAX in terms of accuracy for most of the categories in the dataset. In particular, S-HMAX significantly outperforms the other methods for the cars and cows.

4.2.2 Caltech101: The Caltech101 dataset contains 101 object categories plus a background category comprising 9144 images. Sampling images are shown in Fig. 9. The size of each image is around 300 × 200 pixels. We conducted this experiment using 1000 patches for the multi-classification procedure. We randomly selected either 15 or 30 images from each category for training and used the 50 remaining images for testing. The classifier was a multiclass linear SVM that used the all-pairs method and was trained on the 101 object and background categories. In the experiment, we compared with self-taught learning (STL) [36], invariant feature hierarchies (IFH) [37], conventional HMAX and an enhanced version of HMAX, FHLib [38]. In addition, we also compared with deep convolutional neural network architectures, for example, convolutional deep belief network (CDBN) [39], convolutional networks (Convnet) [40], caffe and vgg models [41, 42]. The results reported here are the average and standard deviation, taken over all 101 classes. The object recognition performance was obtained from 10 independent trials.

As shown in Table 2, we achieved 54% test accuracy using 15 training images per class, and 59% test accuracy using 30 training images per class. Our results were competitive with STL, IFH, HMAX and FHLib. Compared with Convet, caffe, vgg and

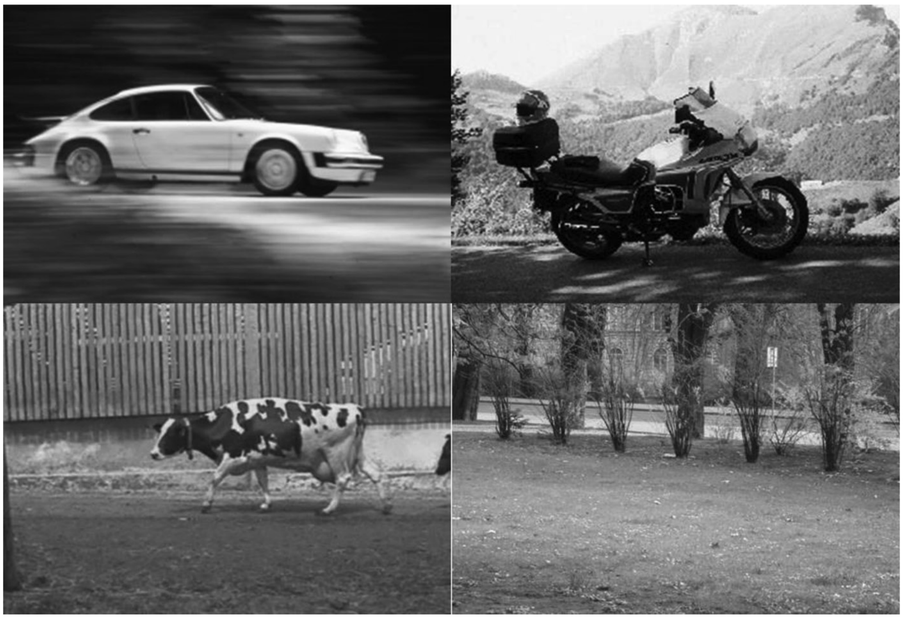

Fig. 7 *Sampling images from the TUD dataset*

Last image is a background image

CDBN, we observed that S-HMAX outperformed Convnet, caffe and vgg in both cases, while showed slightly worse performance than CDBN. Overall, S-HMAX is generally comparable with these methods.

It is worth mentioning that the Convnet, caffe and vgg architectures have gotten quite promising results for object recognition [40–42], but these results are based on the ImageNet-pretrained models [43], which are trained by millions of

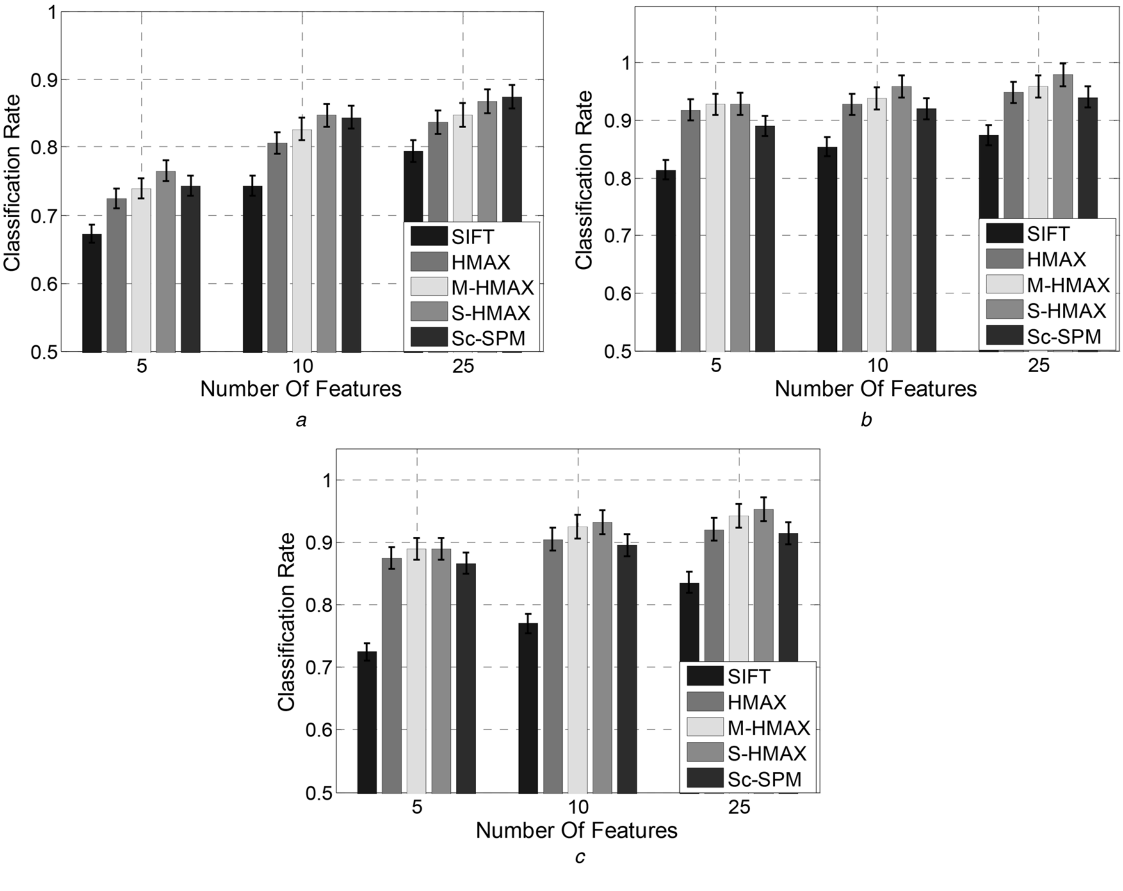

Fig. 8 *Comparison of S-HMAX with standard HMAX, M-HMAX, SIFT and Sc-SPM on the TUD database*

a Motorcycles
b Cars
c Cows

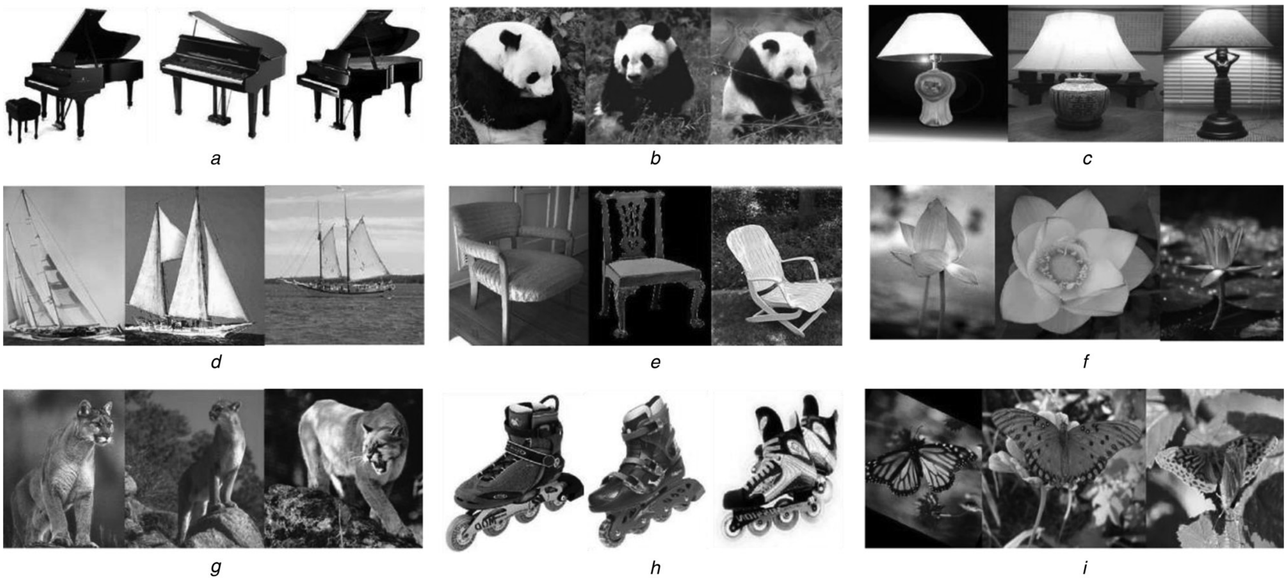

Fig. 9 *Sampling images from the Caltech101 database*

a Piano
b Panda
c Lamp
d Ketch
e Chair
f Lotus
g Cougar body
h Inline skate
i Butterfly

images. However, in the experiment, the models trained from scratch performed relatively poorly.

4.2.3 ImageNet: ImageNet is a large-scale dataset of object classes with millions of images. Using this dataset, we studied the performance of our S-HMAX model and compared it with several convolutional neural networks (CNN) models, for example, caffe [41], vgg-f, vgg-m and vgg-s [42, 44], which have generated quite promising results in object recognition. Caffe is a deep learning framework developed with cleanliness and speed in mind, which separates model representation from actual implementation. Vgg is a convolutional neural network that increases depth using convolution filters. Vgg-f, vgg-m and vgg-s, respectively, are fast, medium, and slow versions of the vgg models by different processing in conv1 layer. To perform this comparison at the feature level, we used the scale and position-invariant C2 features of S-HMAX, and the 20th-layer features of CNN models [43, 45]. We conducted this experiment utilising 2000 patches. We randomly selected 100 categories from ImageNet database and randomly chose 30, 50 and 1000 images from each category for training [The trained models of CNNs are publicly available in vlfeat [44], which were trained to perform object classification on ILSVRC12 (about 1000 images for training per category, hence we used the pretrained models in the case of 1000 training images. In the cases of 30 and 50 training images, all the models in the experiment were trained from scratch.]. 150 randomly selected images were used for testing. The classifier was a multiclass linear SVM. In the experiment, we resized the whole image to 224 × 224. Table 3 provided the classification results on ImageNet. The results reported here were the average and standard deviation. The object recognition performance was obtained from 10 independent trials.

As shown in Table 3, S-HMAX achieved 46.3% recognition accuracy using 30 training images per category, and 54.7% recognition accuracy using 50 training images per category. It outperformed the CNN models when the number of training images was relatively small, that is 30 or 50. However, when the number of training images was up to 1000 per category, CNNs showed more promising results than S-HMAX.

We note that the CNNs use high-level features (20th layer), while S-HMAX chooses C2 features that are low-level (4th layer). The high-level features of CNNs get adequate training and exhibit promising performance if the size of training sets is big (i.e. 1000 training images); if not, their performance is often indifferent. By contrast, C2 features are low-level and are prone to overfitting when using big training data [16], which limits the improvement of performance. However, when the size of training data are moderate (i.e. 30 or 50 training images), S-HMAX performs well in classification. In the CNN models, millions of parameters need to be trained for recognition tasks. In the case of sufficient training

Table 2 Multi-classification comparison of several approaches on Caltech101

Model	15 training images/cat.	30 training images/cat.
S-HMAX, %	54 ± 1.7	59 ± 1.5
STL, %	46.6	52.5
IFH, %	48	54
HMAX, %	44 ± 1.1	51 ± 1.2
FHLib, %	51	56
caffe, %	21.7 ± 1.7	43.4 ± 1.5
vgg,%	22.1 ± 1.9	44.8 ± 1.7
convnet, %	22.8 ± 1.5	46.5 ± 1.7
CDBN, %	57.7 ± 1.5	65.4 ± 0.5

Table 3 Multi-classification comparison of several approaches on ImageNet

Training size (per class)	30	50	1000
caffe, %	22.6 ± 1.5	38.3 ± 1.2	75.9 ± 0.3
vgg-f, %	20.2 ± 1.7	36.7 ± 1.5	71.3 ± 0.7
vgg-m, %	23.8 ± 1.3	38.8 ± 1.3	76.4 ± 0.5
vgg-s, %	24.6 ± 1.2	39.3 ± 1.0	**77.5 ± 0.5**
S-HMAX, %	**46.3 ± 0.5**	**54.7 ± 0.3**	67.3 ± 0.2

Best results are shown in bold

Fig. 10 *Sampling images of GRAZ-01*

From left- to right-hand side, the categories are bikes, people and backgrounds

images, these parameters are full-fledged and the CNN models can achieve promising performance for complex recognition tasks. If these parameters are inadequate training, the performance of the CNN models will sharply decrease. Therefore in the case of small sizes of training data, S-HMAX does outperform the CNN methods. Overall, we observe that the S-HMAX model achieves very competitive performance with the CNN models in the relatively small data case.

Table 4 Performance comparison of several approaches on GRAZ-01

Method	Bikes		Persons	
	EER	AUC	EER	AUC
SIFT	64.3	73.6	61.2	66.9
BoW	75.8	81.5	74.4	81.4
SPM	76.6	84.9	75.7	83.5
BoFLH	81.5	91.6	**79.3**	**88.6**
HMAX	87.8	95.9	71.4	79.7
M-HMAX	85.7	96.5	75.5	84.1
S-HMAX	**90.8**	**99.0**	76.7	86.4

Best results are shown in bold

4.2.4 GRAZ01: GRAZ-01 is a challenging database with high interclass variability on highly cluttered backgrounds, containing people, bikes and backgrounds. Sampling images are shown in Fig. 10. For this database, we followed the method presented in [32]: 100 images (bike or person) and 100 images (backgrounds) were randomly chosen as the training set and 50 other images (bike or person) and other images (backgrounds) were chosen as the testing set. One hundred patches (features) were used for the experiment. We repeated the experiment 10 times and reported the mean values of the test results. For effective evaluation of the S-HMAX model, we also tested the receiver operating characteristic (ROC) and recall-precision (RP) curves and compared the performance of the proposed model to that of related approaches (i.e. SIFT [8], BoW [12], SPM [46], bag of frequent local histograms (BoFLH) [47], HMAX and M-HMAX [34]). The results are shown in Table 4 and Fig. 11.

Table 4 provides the ROC curves results. S-HMAX achieved competitive performance for the detection equal-error-rate (EER) of the ROC curve and area under the ROC curve (AUC) tests. We observed that S-HMAX gave promising results for people, and outperformed SIFT, BoW, SPM, HMAX and M-HMAX. Our proposed model also clearly outperformed these approaches for bikes. Even compared with BoFLH, S-HMAX exhibited slightly better performance for bikes. Fig. 11 shows the RP curves on GRAZ-01. S-HMAX was competitive with the related approaches:

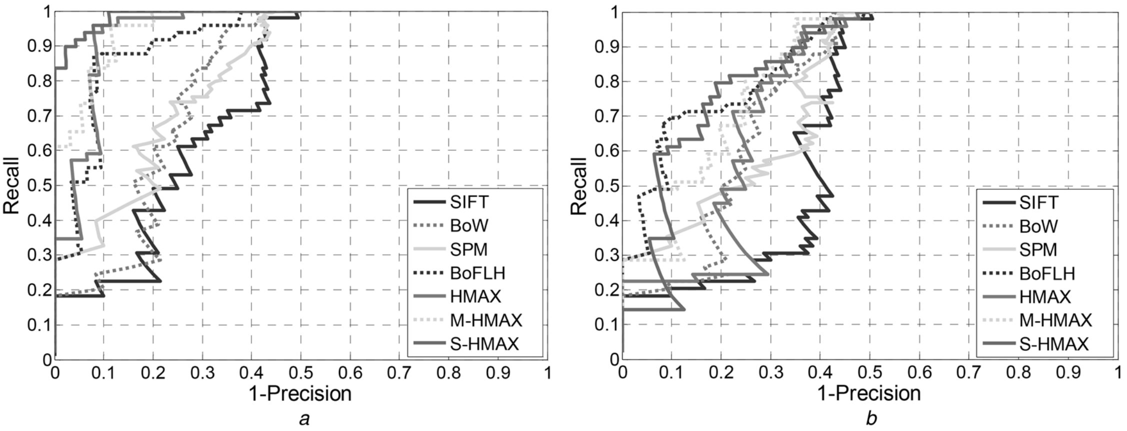

Fig. 11 *RP curves of several approaches on GRAZ-01*

a Bikes

b Persons

S-HMAX obviously outperforms all the comparison methods in the bike category. Comparing with BoFLH, S-HMAX showed a slightly lower RP result for people when the precision is high (1−precision < 0.2). However, S-HMAX outperforms BoFLH when the precision is low (1−precision > 0.2). In general, S-HMAX is competitive for the recognition of both bikes and people.

5 Conclusion

In this article, we presented SKPS, a SKPS method aimed at solving the issues of the huge amount of redundant information extracted by and the poor rotation invariance of the conventional HMAX model. We enhanced the HMAX model with the SKPS method to extract discriminative and invariant features. The SKPS-based patches are robust to image distortions, including rotation. The proposed S-HMAX model increases the rotation invariance and reduces redundant information, thereby providing a good balance between selective representation and invariance. Experiments on four different databases demonstrated that our proposed model performs well in a variety of visual recognition tasks. Our work thus far has focused mainly on the patch selection of HMAX. Obtaining higher-level features based on HMAX will constitute our future work.

6 Acknowledgment

This work was supported by the General Research Program through the National Research Foundation of Korea funded by the Ministry of Education (grant no. NRF-2013R1A1A2008698).

7 References

1 DeSouza, G.N., Kak, A.C.: ‘Vision for mobile robot navigation: A survey’, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002, **24**, (2), pp. 237–267

2 Hsieh, J.W., Yu, S.H., Chen, Y.S., *et al.*: ‘Automatic traffic surveillance system for vehicle tracking and classification’, *IEEE Trans. Intell. Transp. Syst.*, 2006, **7**, (2), pp. 175–187

3 Ye, Q., Liang, J., Jiao, J.: ‘Pedestrian detection in video images via error correcting output code classification of manifold subclasses’, *IEEE Trans. Intell. Transp. Syst.*, 2012, **13**, (1), pp. 193–202

4 Li, L., Huang, W., Gu, I., *et al.*: ‘Statistical modeling of complex backgrounds for foreground object detection’, *IEEE Trans. Image Process.*, 2004, **13**, (11), pp. 1459–1472

5 Zhang, B., Gao, Y., Zhao, S., *et al.*: ‘Kernel similarity modeling of texture pattern flow for motion detection in complex background’, *IEEE Trans. Cir. Syst. Video Tech.*, 2011, **21**, (1), pp. 29–38

6 Le Meur, O., Le Callet, P., Barba, D.: ‘Predicting visual fixations on video based on low-level visual features’, *Vis. Res.*, 2007, **47**, (19), pp. 2483–2498

7 Zhang, S., Yao, H., Sun, X., *et al.*: ‘Robust visual tracking using an effective appearance model based on sparse coding’, *ACM Trans. Intell. Syst. Tech.*, 2012, **3**, (3), p. 43

8 Lowe, D.G.: ‘Distinctive image features from scale-invariant keypoints’, *Int. J. Comp. Vis.*, 2004, **60**, (2), pp. 91–110

9 Mikolajczyk, K., Schmid, C.: ‘A performance evaluation of local descriptors’, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, **27**, (10), pp. 1615–1630

10 Dalal, N., Triggs, B.: ‘Histograms of oriented gradients for human detection’. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2005, pp. 886–893

11 Bay, H., Ess, A., Tuytelaars, T., *et al.*: ‘Speeded-up robust features (SURF)’, *Comput. Vis. Image Und.*, 2008, **110**, (3), pp. 346–359

12 Joachims, T.: ‘Text categorization with support vector machines: Learning with many relevant features’ (Springer Berlin Heidelberg, 1998), pp. 137–142

13 Leung, T., Malik, J.: ‘Representing and recognizing the visual appearance of materials using three-dimensional textons’, *Int. J. Comp. Vis.*, 2001, **43**, (1), pp. 29–44

14 Hubel, D.H., Wiesel, T.N.: ‘Receptive fields of single neurones in the cat’s striate cortex’, *J. Physiol.*, 1959, **148**, (3), p. 574

15 Riesenhuber, M., Poggio, T.: ‘Hierarchical models of object recognition in cortex’, *Nat. Neurosci*, 1999, **2**, (11), pp. 1019–1025

16 Serre, T., Wolf, L., Bileschi, S., *et al.*: ‘Robust object recognition with cortex-like mechanisms’, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007, **29**, (3), pp. 411–426

17 Serre, T., Riesenhuber, M.: ‘Realistic modeling of simple and complex cell tuning in the HMAX model, and implications for invariant object recognition in cortex’ (Massachusetts Inst of Tech Cambridge Computer Science and Artificial Intelligence Lab, 2004)

18 Li, L., Yuan, X., Lu, Z., *et al.*: ‘Rotation invariant watermark embedding based on scale-adapted characteristic regions’, *Inf. Sci.*, 2010, **180**, (15), pp. 2875–2888

19 Lu, Y.F., Zhang, H.Z., Kang, T.K., *et al.*: ‘Extended biologically inspired model for object recognition based on oriented Gaussian–Hermite moment’, *Neurocomputing*, 2014, **139**, pp. 189–201

20 Singh, C., Walia, E., Mittal, N.: ‘Rotation invariant complex Zernike moments features and their applications to human face and character recognition’, *IET Comp. Vis.*, 2011, **5**, (5), pp. 255–266

21 Jones, J.P., Palmer, L.A.: ‘An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex’, *J. Neurophysiol.*, 1987, **58**, (6), pp. 1233–1258

22 Itti, L., Koch, C.: ‘Computational modelling of visual attention’, *Nat. Rev. Neurosci.*, 2001, **2**, (3), pp. 194–203

23 Itti, L., Koch, C., Niebur, E.: ‘A model of saliency-based visual attention for rapid scene analysis’, *IEEE Trans. Pattern Anal. Mach. Intell.*, 1998, **20**, (11), pp. 1254–1259

24 Perazzi, F., Krahenbuhl, P., Pritch, Y., Hornung, A.: ‘Saliency filters: Contrast based filtering for salient region detection’. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2012, pp. 733–740

25 Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: ‘Saliency detection via graph-based manifold ranking’. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2013, pp. 3166–3173

26 Hou, X., Zhang, L.: ‘Saliency detection: A spectral residual approach’. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2007, pp. 1–8

27 Rosten, E., Porter, R., Drummond, T.: ‘Faster and better: A machine learning approach to corner detection’, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010, **32**, (1), pp. 105–119

28 Rosten, E., Drummond, T.: ‘Machine learning for high-speed corner detection’. Proc. European Conf. Computer Vision (ECCV), 2006, pp. 430–443

29 Bastian, L., Leonardis, A., Schiele, B.: ‘Combined object categorization and segmentation with an implicit shape model’. Proc. Workshop on Statistical Learning in Computer Vision (ECCV), 2004, p. 7

30 Li, F., Perona, P.: ‘A Bayesian hierarchical model for learning natural scene categories’. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2005, pp. 524–531

31 Deng, J., Dong, W., Socher, R., *et al.*: ‘Imagenet: A large-scale hierarchical image database’. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2009, pp. 248–255

32 Opelt, A., Pinz, A., Fussenegger, M., Auer, P.: ‘Generic object recognition with boosting’, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2006, **28**, (3), pp. 416–431

33 Chang, C., Lin, C.: ‘LIBSVM: A Library for Support Vector Machines’, http://www.csie.ntu.edu.tw/~cjlin/libsvm/, accessed 2014

34 Huang, K., Tao, D., Yuan, Y., *et al.*: ‘Biologically inspired features for scene classification in video surveillance’, *IEEE Trans. Syst. Man. Cybern. B.*, 2011, **41**, (1), pp. 307–313

35 Yang, J., Yu, K., Gong, Y., *et al.*: ‘Linear spatial pyramid matching using sparse coding for image classification’. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2009, pp. 1794–1801

36 Raina, R., Battle, A., Lee, H., *et al.*: ‘Self-taught learning: transfer learning from unlabeled data’. Proc. ACM Conf. on Machine Learning, 2007, pp. 759–766

37 Ranzato, M., Huang, F.J., Boureau, Y.L., *et al.*: ‘Unsupervised learning of invariant feature hierarchies with applications to object recognition’. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2007, pp. 1–8

38 Mutch, J., Lowe, D.G.: ‘Object class recognition and localization using sparse features with limited receptive fields’, *Int. J. Comput. Vis.*, 2008, **80**, (1), pp. 45–57

39 Lee, H., Grosse, R., Ranganath, R., *et al.*: ‘Unsupervised learning of hierarchical representations with convolutional deep belief networks’, *Commun. ACM.*, 2011, **54**, (10), pp. 95–103

40 Zeiler, M.D., Fergus, R.: ‘Visualizing and understanding convolutional networks’. Proc. European Conf. Computer Vision (ECCV), 2014, pp. 818–833

41 Jia, Y., Shelhamer, E., Donahue, J., *et al.*: ‘Caffe: Convolutional architecture for fast feature embedding’. Proc. ACM Conf. on Multimedia, 2014, pp. 675–678

42 Chatfield, K., Simonyan, K., Vedaldi, A., *et al.*: ‘Return of the devil in the details: delving deep into convolutional nets’, arXiv preprint arXiv: 1405.3531, 2014

43 Krizhevsky, A., Sutskever, I., Hinton, G.E.: ‘Imagenet classification with deep convolutional neural networks’. Adv. in Neur. Inf. Proc. Sys., 2012, pp. 1097–1105

44 http://www.vlfeat.org/matconvnet/, accessed 2014

45 Razavian, A.S., Azizpour, H., Sullivan, J., *et al.*: ‘CNN Features off-the-shelf: an Astounding Baseline for Recognition’, arXiv preprint arXiv: 1403.6382, 2014

46 Lazebnik, S., Schmid, C., Ponce, J.: ‘Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories’. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2006, pp. 2169–2178

47 Fernando, B., Fromont, E., Tuytelaars, T.: ‘Mining mid-level features for image classification’, *Int. J. Comp. Vis.*, 2014, **108**, (3), pp. 186–203