Enhanced object recognition in cortex-like machine vision

Aristeidis Tsitiridis[1] , Peter WT Yuen[1], Izzati Ibrahim[1], Umar Soori[1], Tong Chen[1], Kan Hong[1], Zhengjie Wang[2], David James[1] and Mark Richardson[1]

[1] Cranfield University,
Department of Informatics and Systems Engineering,
Defence College of Management and Technology,
Shrivenham, Swindon, SN6 8LA, United Kingdom

[2] Electrical Engineering Dept, Beijing Institute of Tech, Beijing, P.R. China

Abstract. This paper reports an extension of the previous MIT and Caltech's cortex-like machine vision models of Graph-Based Visual Saliency (GBVS) and Feature Hierarchy Library (FHLIB), to remedy some of the undesirable drawbacks in these early models which improve object recognition efficiency. Enhancements in three areas, a) extraction of features from the most salient region of interest (ROI) and their rearrangement in a ranked manner, rather than random extraction over the whole image as in the previous models, b) exploitation of larger patches in the C1 and S2 layers to improve spatial resolutions, c) a more versatile template matching mechanism without the need of 'pre-storing' physical locations of features as in previous models, have been the main contributions of the present work. The improved model is validated using 3 different types of datasets which shows an average of ~7% better recognition accuracy over the original FHLIB model.

Keywords: Computer vision, Human vision models, Generic Object recognition, Machine vision, Biological-like vision algorithms

1 Introduction

After millions of years of evolution visual perception in primates is capable of recognising objects independent of their sizes, positions, orientations, illumination conditions and space projections. Cortex-like machine vision [1] [2] [3] [4] [5] [6], attempts to process image information in a similar manner to that of biological visual perception. This work is different from other studies on human vision models such as [6] [7] [8], in which the features are extracted with a statistical distance based descriptor methodology rather than a biologically inspired saliency approach [9]. In Itti's recent biological visual research [10] [11], which utilised a 44-class data set for the scene classification work, the authors reported a slightly inferior performance of the biological inspired C2 feature than the statistical SIFT. However, it is difficult to draw conclusions based on a single dataset and more work is needed to confirm their results. Other models such as Graph-Based Visual Saliency (GBVS) [12] and Feature

Hierarchy Library (FHLib) [13], have implemented biological vision in hierarchical layers of processing channels similar to the ventral and dorsal streams of the visual cortex [14]. In these models the dorsal stream process has been implemented by means of a saliency algorithm to locate the positions of the most "prominent" regions of interest for visual attention. To mimic the simple and complex cell operations in the primary visual cortex for object recognitions, the cortex ventral stream process has been commonly presented in alternating layers of simple (S-layers) and complex (C-layers) operations. In the centre of these cortex-like models is the centre-surround operation over different spatial scales (so called image pyramids) of the image in both the dorsal-like and ventral-like processing. Although previous models have achieved rather impressive results using the Caltech 101 data as the test set [13], additional improvements are needed. Firstly, the previous work [13] considered datasets where all objects in the images are in the centre of the scene, like that in the Caltech 101 dataset. Secondly, FHLIB has been testified using a single dataset and there is real need to extend the tests using different datasets. Thirdly, the existing model extracts features for the template library in a random manner, which may reduce performance due to the inclusion of non-prominent and/or repeated features. Fourthly, the template matching process itself in FHLIB utilises a pre-stored location of the templates from the training data and then "searches" around the vicinity of this location to perform matching. This "blind" searching mechanism is neither efficient nor adaptive.

This paper is largely based on MIT and Caltech's work and it addresses the above four points. All enhancements are implemented within the cortex-like FHLIB framework. All codes utilised in this work have been implemented in MATLAB and results are compared with that of the "original" FHLIB algorithm which has been re-coded in this work, according to paper [13] and it is referred as MATLAB-FHLIB [MFHLIB]. One main contribution in this work is the incorporation of saliency within the template feature extraction process and this is termed as saliency FHLIB [SFHLIB]. Other enhancements have been the substitution of the computationally expensive Gabor filters for multiple orientations with a single circular Gabor filter, the improvement of the feature representation using larger patches as well as the addition of an extra layer to refine the feature library thereby eliminating redundant patches while at the same time ranking features in the order of significance. The classification in the present paper is achieved through a Support Vector Machine (SVM) classifier using the features extracted in a cortex-like manner similar to the previous work.

2 Cortex-like vision algorithms

2.1 Graph-Based Visual Saliency (GBVS)

In GBVS the saliency of intrinsic features is obtained without any intentional influence while incorporating a very important characteristic of biological vision, the centre-surround operation. Initially for a digital RGB input image, the algorithm extracts the fundamental features i.e. average intensity over RGB bands, double opponency colour and Gabor filter orientations, and activation maps are formed by using image pyramids across different scales under centre surround operations. The final saliency map that highlights the most prominent regions of interest (ROI) in the

image is then constructed from the normalised Markov chains of these activation maps. For more information on GBVS please refer to [12].

2.2 Feature Hierarchy Library (FHLIB)

FHLIB is an unsupervised generic object recognition model which consists of five layers. Their operations follow the early discoveries of simple and complex cells in the primary visual cortex [15] and alternate in order to simulate the unsupervised behaviour of the ventral stream as it propagates visual data up to higher cortical layers. FHLIB's layers are the input image layer, Gabor filter (S1) layer, Local invariance (C1) layer, Intermediate feature (S2) layer, Global invariance (C2) layer. S2 patches are stored in a common featurebook during the training phase and used as templates to be matched against testing images. The stored C2 vectors from the training phase can be compared against the C2 vectors of a given test image, for example by means of a linear classifier such as a Support Vector Machine (SVM). In addition, FHLIB introduced further modifications with respect to previous biologically inspired vision models to improve performance such as the sparsification of S2 units, inhibition of S1/C1 units and limitation of position/scale invariance in S2 units. For more information on FHLIB please refer to [13].

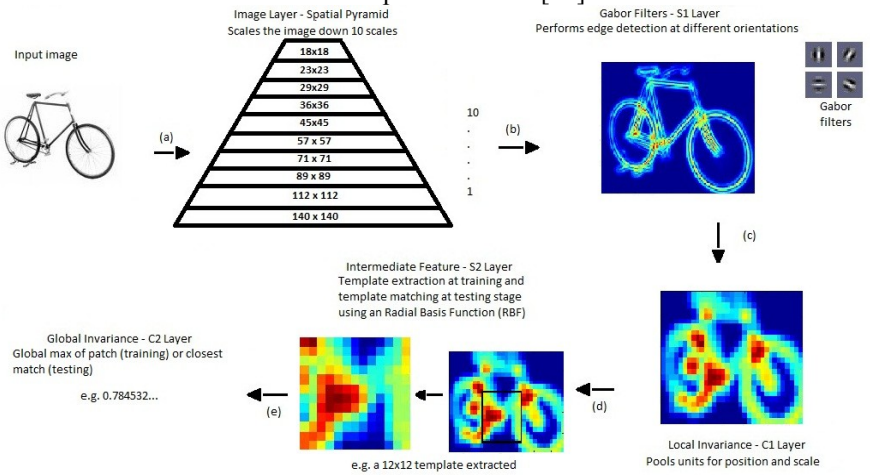

Fig.1. FHLIB's architecture by forming a pyramid of various resolutions of the image, followed by tuning the Gabor features in the S layers and max-pooling across the adjacent C layers in the pyramid, then brings spatial information down to feature vectors for classification.

3 The Cranfield University algorithm (SFHLIB)

3.1 Feature extractions from a salient Region Of Interest (ROI)

Unless there is a task in which even the most refined features are required to distinguish subtle differences or similarities between objects (often of the same category) then retaining all visual information is computationally expensive and unnecessary. Currently in FHLIB, there is no specific pattern by which features are

extracted and the selection process of both feature size and locations occurs randomly across input images. Moreover, it becomes difficult to estimate the number of features or feature sizes ideally required. Solving this problem by introducing a geometric memory in the algorithm (section 2.2) i.e. storing the location coordinates from which an S2 template was found so that a respective area in a testing image is compared, led to the conclusion that such a system becomes specialised in recognising objects in similar locations [16]. This however is impractical for real-world situations since objects may appear at other locations or may become differently orientated and so the algorithm must generically overcome this problem.

By applying the GBVS model on a particular object points to salient areas and evaluates an activation map according to priority of attention. For objects of the same category the most prominent areas are nearly the same and thus condensation of structured objects can be achieved (figure 2). We use the orientation feature only and rank salient areas of a certain circumference around the highest values. These areas effectively represent the local features which can be used along with more global shape areas (i.e. larger types of features extracted freely from any point in the image) and can be combined for the recognition stage.

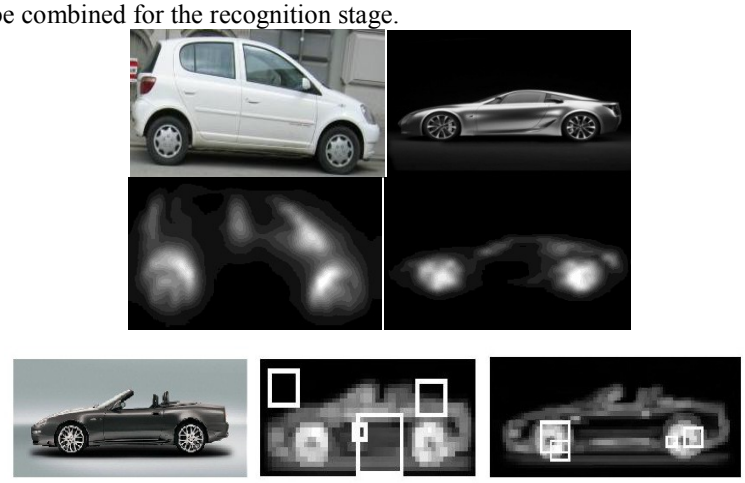

Fig. 2. Two images of cars. Top row shows the original images and second row their saliency maps using GBVS (12 Gabor angles). Highest attention accumulates on the areas of the wheels which is a common saliency feature and it is evident that saliency can effectively ignore background information. Third row shows the effect of accurate feature extraction via salience in a C1 layer map. Rectangular boxes illustrate the feature templates of varying sizes. Extraction occurs in FHLIB (centre) "blindly" while in C1 map from SFHLIB (right) patches are extracted from the salient ROI.

3.2 Higher resolution, patch sizes and Circular Gabor filters

Salient areas can be very specific to small regions of an image. At low resolutions spatial information is also low and therefore extractions yield to incoherent representations of the object. To overcome this problem and to improve spatial representation, the resolution of images has to be increased. At the same time, in order to maintain the spectrum of patch sizes required to store suitable features the patch sizes have to be enlarged. To tackle these issues, we increase the size of the short

edge of an image to 240 pixels (thus preserving the aspect ratio) and include S2 feature patches of size 20x20 and 24x24.

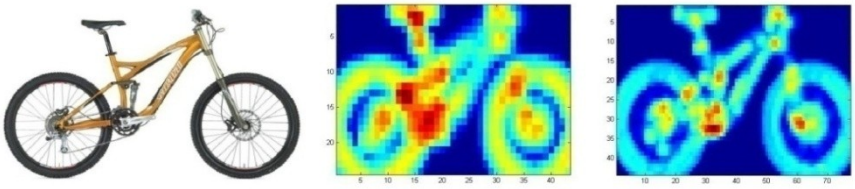

Fig. 2. An example of an original image (left) which at a lower resolution (input image 140 pixels) at the C1 layer (middle, first –finest scale) has retained little of the object's structure. At a higher resolution (input image 240 pixels) the C1 layer shows a more detailed representation (right).

The use of Gabor filter banks in object recognition simulates the tuning of V1 simple cell at different orientations (θ) well and highlights their role in bottom-up mechanisms of the brain. However, constructing S1 responses for different orientations requires the creation of an equal amount of Gabor pyramids for each orientation which is computationally expensive and time consuming as the number of orientations increases to improve an object's description. To eliminate this, we generalise the S1 responses by varying the sinusoid across all orientations which then becomes circular symmetric [17]. Using this single circular Gabor filter, one S1 pyramid is obtained and at the same time FHLib's sparsifying step over orientations (section 2.2) become redundant and are removed. The circular Gabor filter is given below:

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

Note that in equation 1, θ and γ are no longer parameters for this equation and this equation now only depends on σ and λ.

3.4 Adding S3 and C3 layers

At the object recognition part, when training template patches are extracted randomly from salient ROI, it is inevitable to have patches extracted more than once from the same location and scale, especially as the required total number of training patches is increased. Furthermore, there is no refinement mechanism currently in FHLIB that evaluates the extracted patches' performance and as such the algorithm may store patches that do not explicitly and accurately represent each class. In FHLIB, a refinement was made at the classification stage [13], however it is a post-processing, non-biological and time consuming remedy.

We address both aforementioned issues by introducing two more layers, namely S3 and C3. In the S3 layer, all patches of a particular class are directly grouped together from the S2 featurebook (section 2.2) and are organised according to their extraction sequence. The algorithm continues by extracting the training C2 vectors (as in

FHLIB) which are again grouped so that the responses of every patch from each class across all images now exist together. By examining the C2 responses of each patch for every class on objects of the same class, e.g. if the class was 'bikes' and a patch was extracted from one of its images then C2 responses for this patch from all images portraying bikes are grouped together. Patches that have yielded identical C2 responses (in practice C2 vectors are float numbers and identical responses can only be obtained from identical patches) are dropped and only one unique patch is retained therefore eliminating co-occurrences. We remember the origin of the retained C2 vectors and refine the S3 featurebook accordingly.

Additionally, the performance of each patch can be measured for every class against objects of the same category to deduce to sampled patches that best describe that class. By summing the C2 responses for every patch we rank the S3 patches from high to low (high showing patches that are most commonly found for a particular object, low showing less generalisation and thus uncommon patches that do not exist across all images). At this point, a percentage number is introduced i.e. the amount of patches to be retained and for example, setting it to a certain value means that the featurebook is reduced by a percentage and the patches retained maximally express the trained classes. The final version of the significantly reduced S3 featurebook refined from co-occurences and uncommon patches, is used over the training images to create C3 vectors which in turn are used to train the SVM classifier. Similarly, at the testing phase, the stored S3 featurebook is used over the testing images and their C3 responses are compared against the trained C3 vectors.

4 Experiments

4.1 Image Datasets

Three image datasets were used, the Cranfield University Uncluttered Dataset (CUUD), the Cranfield University Cluttered Dataset (CUCD) and the Caltech 101. CUUD consists of four categories of vehicle images that were collected from the internet and are namely airplanes, bikes, cars and tanks. Each image contains only a particular vehicle without any clutter or obscurances (figure 3a). The images are of varying aspect ratios and their resolutions are always higher than a minimum of 240 pixels for their shortest edge. Objects are in varying directions and portray some variation in spatial position. Naturally, we have separated the dataset into different training and testing images.

CUCD has also been partly assembled from the internet and in part from our own image database. All images in the dataset contain background clutter and belong to four categories background, bikes, cars, and people. The background category shows a great variability of information, i.e. buildings, roads, trees etc. The people's category is the only category of non-rigid objects and therefore within this category pose and position vary significantly. Another difference with respect to CUUD is that in an image there may be more than one object (of the same category) present. Similarly to CUUD, the images are of varying aspect ratios and their resolutions are always higher than a minimum of 240 pixels for their shortest edge (figure 3b).

(a) **(b)**

Fig. 3. (a) Four example images (airplanes, bikes, cars, tanks) from the CUUD vehicle classes. No background clutter is present and images contain only one object for classification, **(b)** four example images (background, bikes, cars, people) from the CUCD classes. Background clutter is present and images in some cases contain more than one object in each image.

The multiclass image dataset Caltech 101 [18] consists of 101 different object categories including one for backgrounds. A total of 9197 images on various spatial poses include unobstructed objects mostly centred in the foreground with both cluttered and uncluttered background environments. All images have been taken at different aspect ratios and are always higher than a minimum of 200 pixels for their shortest edge.

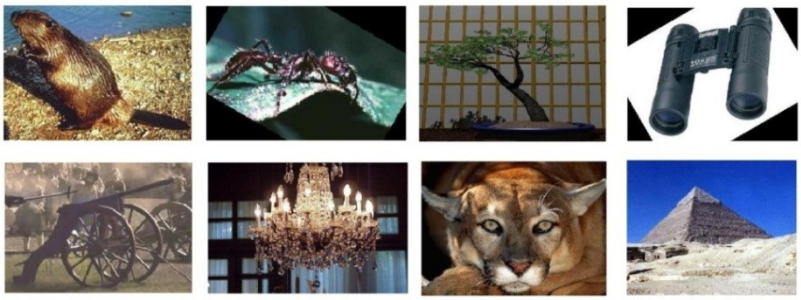

Fig. 4. Some examples of classes from the 101 Caltech dataset.

4.2 Experiments Setup

In this work, we directly use GBVS MATLAB code with some modifications while all code regarding the recognition part of the algorithm has been inspired from [13] but otherwise created from the authors .

The algorithm is first tested with FHLIB-like parameterisation, 140 pixels for the images' short edge and 4 different size patches (4x4, 8x8, 12x12, 16x16), 11x11 Gabor filter banks while a sliding window approach was used to extract the maximum C2 responses across the entire image. At this point, the Gabor filters consist of 12 banks i.e. one per orientation angle. Subsequently, we enhance this algorithm with our improvements gradually by introducing a higher resolution for each image (240 pixels, short edge) and adding two more patch sizes 20x20 and 24x24. We then attach

our feature extraction method using saliency and also substitute the 12 Gabor filters with one circular Gabor filter. Finally, the S3 and C3 layers are in turn embedded.

Efficient and fast biological-like detection and object recognition requires parallel execution. For our experiments, we concentrate on the results and use a sequential approach in order to prepare the saliency maps of both training and testing images of our dataset beforehand.

Each saliency map from GBVS matches the size of the original image that is later used in object recognition exactly i.e. 240 pixels for the shortest edge, and the only feature used is orientation at 12 Gabor angles spanning from 0 to π. For all experiments during training, we choose an abundant number of features (10000) to avoid underrepresentation of objects and Gabor filter parameters γ, σ and λ are all fixed according to [13]. For datasets CUUD and CUCD, 50 different images for each class were chosen for training and another 50 per class for training. For the Caltech dataset, we use 15 per class for training and 15 per class for testing. Classification accuracies are obtained as an average of 3 independent runs for all experiments. Finally, all classification experiments were conducted using an SVM classifier using one-against-one decomposition.

5 Results – Discussion

MFHLIB is the foundation upon which all improvements of this work are established. It extracts the maximum responses by using a sliding window approach which is overall inefficient, time-consuming and as table 1 shows the least accurate. Salience FHLIB (SFHLIB) on other hand illustrates higher performance and robust behaviour.

Method	Dataset CUUD – Classification Accuracy (%)	Dataset CUCD– Classification Accuracy (%)	Dataset Caltech– Classification Accuracy (%)
MFHLIB	80	70.6	18.75
SFHLIB + Circular Gabor	85	80.4	22.4
SFHLIB + S3/C3 Layers (60% features)	86	76.6	19
SFHLIB + S3/C3 Layers (100% features)	81	80.4	21.4

Table 1. Average percentage classification accuracies over 3 independent runs for the three datasets. Note that descending order algorithms in the left column include the enhancements of the previous algorithms. All results typically vary at ±1.5% (see our discussion for more detail).

From table 1, the results portray for all enhancements a gradual improvement over both the accuracy itself and time. CUUD being uncluttered, presents minimal difficulty for an algorithm and classification accuracies were overall the highest. Under this dataset, a 6% percentage improvement was observed between MFHLIB and SFHLIB variants (excluding SFHLIB with 100% features).

A higher difference between the MFHLIB and our enhancements was noticed in CUCD. In this dataset even though the number of classes remains the same, the added

background information and more complicated poses, affect the performance of all algorithms, particularly in MFHLIB. As a first step by increasing the resolution and tampering the number and size of patches has increased performance by 6% and a total of 10% better performance was achieved by using SFHLIB with circular Gabor filters. A drop of nearly 10% for MFHLIB between CUUD and CUCD signifies its inefficiency as a dataset becomes more realistic. A decrease in performance (4.5%) can be also observed for SFHLIB though it is almost half compared with MFHLIB.

Experiments with the benchmark Caltech 101 dataset have revealed a decrease in performance with respect to the other two datasets which was primarily caused by the large number of classes and different setup. However, within this set of experiments an incremental difference between FHLIB and SFHLIB is apparent.

Classification accuracies for S3/C3 layers show that although for the cluttered datasets an improvement can be claimed the trend is not followed in CUUD. A major difference between previous variants of the code is that the number of features required to achieve this performance was lower and thus computationally cheaper. Having selected a fixed number of features (10000) for the library, by running the S3/C3 on the CUUD, reductions of an average of 15% were observed for a 100% of the features used. Similarly for the CUCD, the average percentage of identical feature discards reached 22% and for the Caltech dataset 10%. The difference of this percentage between the three datasets can be explained by the larger number of images used in the Caltech data. The same total number of features corresponds to fewer features per image thus reducing the probability of identical patches extracted randomly across salient regions. Discarding identical features improves time (by approximately the same percentage) and computational requirements.

6 Conclusions

Following the basic cortex-like machine vision models of FHLIB and GBVS, the contribution here has been to enhance object recognition performances in these early models by incorporating the visual saliency into the ventral stream process. The SFHLIB version being a fusion of salience and recognition, has achieved ~6% classification accuracy for the CUUD, 10% for the CUCD and 4.5% for the Caltech dataset better than that of the MFHLIB model. The present work has also highlighted the need of an efficient feature extraction method from the dataset and further alterations on the mechanism of the algorithm revealed the significance of refining the extracted features to improve the integrity of the feature library itself. It has been also found that the computational time of the proposed SFHLIB is faster by a significant percentage than the MFHLIB.

It is planned to use more extensive datasets to verify the newly developed SFHLIB algorithm against its portability and adaptability. Moreover, it is planned to employ a pulsed (spiking) neural network to replace the SVM classifier for object classification in the near future.

Acknowledgements

The authors thank Drs C Lewis, R Bower & R Botton of the CPNI. AT thanks EPSRC for the provision of DTA grant and TC & KH thank the DCMT internal funding of their studentships.

References

1. Treisman, A., Gelade, G.: A feature-integration theory of attention. Cognitive Psychology 12(1), 97-136 (1980)
2. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on pattern analysis and machine intelligence 20(11) (1998)
3. Itti, L.: Visual Attention. In : The Handbook of Brain Theory and Neural Networks. MIT Press (2003) pp. 1196-1201
4. Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nature Neuroscience 2(11), 1019-1025 (1999)
5. Riesenhuber, M., Poggio, T.: Models of object recognition. Nature Review (2000)
6. Serre, T., Wolf, L., Poggio, T.: Object recognition with features inspired by visual cortex. CVPR (2005)
7. Fukushima, K., Miyake, S., Ito, T.: Neocognitron: a neural network model for a mechanism of visual pattern recognition. In : IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-13(Nb. 3, p.826—834 (September/October 1983)
8. Wysoski, S., Benuskova, L., Kasabov, N.: Fast and adaptive network of spiking neurons for multi-view and pattern recognition., 2563-2575 (2008)
9. Zhang, W., Deng, H., Diettrich, G., Mortensen, N.: A Hierarchical Object Recognition System Based on Multi-scale Principal Curvature Regions. In : 18th International Conference on Pattern Recognition (ICPR'06) (2006)
10. Elazary, L., Itti, I.: A Bayesian model for efficient visual search and recognition., 1338–1352 (2010)
11. Borji, A., Itti, L.: Scene Classification with a Sparse Set of Salient Regions. In : IEEE International Conference on Robotics and Automation (ICRA), Shanghai, China (February 2011)
12. Harel, J., Koch, C., Perona, P.: Graph-Based Visual Saliency., Eds. Cambridge, MA: MIT Press (2007)
13. Mutch, J., Lowe, D.: Object class recognition and localisation using sparse features with limited receptive fields. International Journal of Computer Vision 80(1), 45-57 (2008)
14. Ungerleider, L., Mishkin, M.: Two cortical visual systems. MIT Press, Cambridge, USA (1982)
15. Hubel, D., Wiesel, T.: Receptive fields and functional architecture of monkey striate cortex. Journal of Physiology 195 (1967)
16. Tsitiridis, A., Yuen, P., Hong, K., Chen, T., Ibrahim, I., Jackman, J., James, D., Richardson, M.: An improved cortex-like neuromorphic system for target recognitions. In : Remote Sensing SPIE Europe, Toulouse (2010)
17. Zhang, J., Tan, T., Ma, L.: Invariant Texture Segmentation Via Circular Gabor Filters. In : 16th International Conference on Pattern Recognition (ICPR'02) (2002)
18. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative models from few training examples: an incremental bayesian approach tested on 101 object cagories. In : CVPR Workshop on Generative-Model Based Vision (2004)
19. Serre, T., Wolf, L., Bilecshi, S., Riesenhuber, M., Poggio, T.: Robust Object Recognition with Cortex-like Mechanisms. In : IEEE transactions on pattern analysis and machine intelligence, vol. 29 (3), pp.411-425 (March 2007)