Object Recognition in Clutter Color Images using Hierarchical Temporal Memory combined with Salient-Region Detection

Radoslav Škoviera[a,*], Ivan Bajla[b], Júlia Škovierová[a]

[a] *Czech Institute of Informatics, Robotics, and Cybernetics, Czech Technical University in Prague, Czech Republic*
[b] *Institute of Measurement Science, Slovak Academy of Sciences, Bratislava*

Abstract

The essential goal of this paper consists in extending the functionality of the bio-inspired intelligent HTM (Hierarchical Temporal Memory) network towards two capabilities: (i) object recognition in color images, and (ii) detection of objects located in *clutter* color images. The former extension is based on development of a novel scheme for application of three parallel HTM networks which separately processes color, texture, and shape information in color images. For the latter HTM extension we proposed a novel system in which HTM is combined with a modified model of computational visual attention. We adopted the results of [1] and [2], and added new elements [3] for the calculation of image saliency maps. The proposed algorithm enables to locate individual objects in clutter images automatically. For computer experiments a special image database has been created to simulate ideal single object images and cluttered images with multiple objects on inhomogeneous background. The recognition performance of the HTM alone and in combination with the salient-region detection method has been evaluated. We showed that the attention subsystem is able to satisfactorily locate objects in clutter color images with inhomogeneous background. We have also carried out benchmark calculations for two selected computer vision methods used for object detection in color clutter images. Namely, the methods of cascade detectors and template matching have been used. Our study confirmed that the proposed attention system can improve the HTM's capabilities for object classification in cluttered images. The compound system of visual attention and HTM outperformed the compared methods in both criteria (recall and correct detection rate). However, as expected, the system cannot match the HTM's recognition accuracy achieved on single object images and the further research is needed.

Keywords: machine learning, pattern recognition, Hierarchical Temporal Memory, image saliency

1. Introduction

The tremendous rise of digital techniques for image acquisition and ubiquity of web connections in recent decade evoked the creation of an enormous number of image databases accessible to web users. A searching/retrieving of images from such databases for various aims becomes an everyday task in many application areas. Especially requested are the so-called content-based image retrieval (CBIR) techniques and systems [4, 5]. Although a number of image retrieving methods have been proposed and explored up to now, no satisfying general solutions still exist. The core problems are: i) choice of suitable image features for image content representation, and ii) efficient image object detection/recognition in cluttered images which are frequently occurring in CBIR tasks. In papers [6], [7] several aspects of solving these two problems by application of an bio-inspired Hierarchical Temporal Memory (HTM) system have been addressed. This system is catagorized by [survey od Rada] in to the class of deep learning algorithms. Pal et al studied using a color histogram technique together with the HTM model for retrieving the final images post classification of the query image. [7] focused on exploration of possibilities of the HTM network to be applied to CBIR, when color features are used. Application of HTM to CBIR tasks revealed, however, similar problems, as outlined above, because the training images for HTM are characteristic by single centered objects, placed usually at homogeneous background.

It appeared, that to overcome the limitations of bio-inspired feed-forward models of visual cortex, their application to multiple object recognition tasks, requires a support of a computational model of visual attention. Several papers have been devoted to this research topic. In [8] a question was explored, to what extent saliency-based bottom-up attention can extract useful information about the location, size, and shape of objects in cluttered scenes. The rigorous quantitative analysis of the authors showed the usefulness of the synergy between recognition and attention. Chikkerur et al [9] proposed a two-stage approach to recognizing objects in clutter. In the first stage, a Bayesian (feature-based) model of visual attention isolates the target object while suppressing the clutter around it. The results are then fed to a hierarchical feed-forward model of object recognition in the ventral stream. They showed that attentive processing improves recognition in comparison to purely feed-forward processing. An improvement of feed-forward object recognition by biologically plausible saliency mechanism has

*Corresponding author
Email address: radoslav.skoviera@cvut.cz (Radoslav Škoviera)
DOI of the published paper: 10.1016/j.neucom.2018.04.030

been explored and demonstrated in the paper [10]. Recently two different models have been proposed in which saliency function is directly combined with a network for object class recognition. In [11] a salient hierarchical model for object recognition was proposed that is characterized by two contributions: (1) a traditional saliency model is modified to achieve more robust saliency estimation, and (2) this part is combined with the Hierarchical Maximization architecture (HMAX) of immediate object recognition in primate visual cortex.

To our knowledge, the first attempt to combine the HTM network model with an image saliency approach represents the recent work of [12]. The authors proposed a supervised learning method for recognition of objects in different orientations. Instead of conventional color image input vectors, the combined model exploits a preceding saliency detection step that isolates the region of interest, releasing the HTM learning procedure from redundant information. The proposed approach has several limitations that hinder from its application to colored clutter images containing many different objects located anywhere in the image. Therefore, the basic motivation of our research was to extend the possibilities of both parts of the combination of the HTM system and some visual attention system including saliency. We propose the novel version of such a combination with the following contributions: (1) the saliency of color images is calculated as a combination of a suppression map with a contrast features, (2) then, a combined feature based on discriminative local regions is calculated, (3) finally, instead of the standard HTM implementation with gray-level images as inputs, a system of three parallel HTM networks is proposed - each HTM network processes a separate image feature map and a coupled k-nn classifier is proposed for weighted supervised classification of belief vectors having been inferred by individual HTM networks. Computer experiments incorporate generation of multiobject images that simulate clutter images in a similar way as in [9] (placement of individual objects at random position and scale in a test multiobject image). In comparison to the above mentioned combined models of feed-forward processing and image saliency maps [11, 12], **which are applied exclusively to images with one object dominated**, in our approach we attempt to solve a more sophisticated multiple object recognition task in clutter images (simulated models). The important goal of the paper is to prove a feasibility of the one-level HTM network application to the task of object recognition in multiobject clutter images.

This paper is organized as follows. In Section2 the basic description of a memory-prediction HTM network is provided. Also, a novel system of parallel coupling of three HTM networks is proposed that processes three separate image features of edge, texture, and color. Section 3 is devoted to the description of image data set construction needed for implementation of computer experiments with the proposed combined system. Two classes of images are used, namely, partially cluttered and fully cluttered images. In Section 4 the details of the proposal of a novel color image saliency model are explained. This is a key section in which the main contributions of our approach to combination of the HTM network with color image saliency mapping (ISM) model are presented. Section 5 is oriented towards implementation of computer experiments. Two benchmark methods are selected and shortly described. The proposed combined system of (ISM + HTM) is compared to cascade detectors, and to template matching applied to the identical testing images. In Section 6 the obtained results are presented and discussed. We used five different characteristics of object detection/recognition success: accuracy, overlap, recall, clutter, and hits. Lastly, Section 7 presents the conclusions of our findings and discusses the contributions and limitations of the proposed approach, presented in this paper. Open issues of a further research in this domain are also formulated.

2. Hierarchical Temporal Memory

2.1. Basic description

The HTM is a memory-prediction network proposed by [13, 14] and distributed initially by Numenta, Inc., as a free software package NuPIC [15, 16]. Promising results in applications of HTM have been achieved especially in the field of visual object recognition/classification, e.g. [17], [18], [19], [20], [21], [22]. It represents a hierarchical Bayesian network and can be assigned to the class of Deep Belief Networks in artificial intelligence [23, 24], [25]. In more details, it can be described as a hierarchy of several layers (levels) consisting of basic operational units called nodes.

The effective area from which a node receives its input is called *field of view* or *receptive field* of the node. The individual levels are ordered in a hierarchical tree-like structure (see Fig.1 as a prototypical HTM network). There is a zero sensory level of the HTM which serves as an input to the first level of nodes. In our case, zero level represents a visual field of image pixels or feature maps derived from it. At the top level, there is only one node that serves for classification. In this role various classifiers can be used. Each HTM node works in two modes – learning and inference. In the learning mode the node performs two operations, spatial pooling and temporal pooling. Once these two steps are completed, the node is switched to the inference mode. In practice, all nodes within one hierarchy level are considered to be equivalent. Since the use of smooth temporal dependencies of input spatial patterns is essential characteristic of the HTM, its learning process utilizes either native sequence of images (e.g., video captured by a camera), or (in case of static images) an artificially generated sequence of images using various exploring schemes.

In the first step of the learning process, the node memorizes the representative spatial patterns (coincidences) from its receptive field that results in creating a codebook of image patterns. After reaching the requested number of quantization centres, the memorization process is stopped. The ultimate goal of the HTM learning is to detect correct invariant representations of the input world based on the temporal relations contained in the learning sequence. To achieve this, one needs a frequency of transition events, i.e., co-occurrences of the memorized coincidences in adjacent time instances. A sequence of the input patterns generates a sequence of the n coincidences within the node. In the HTM theory [14, 15], the temporal relations are

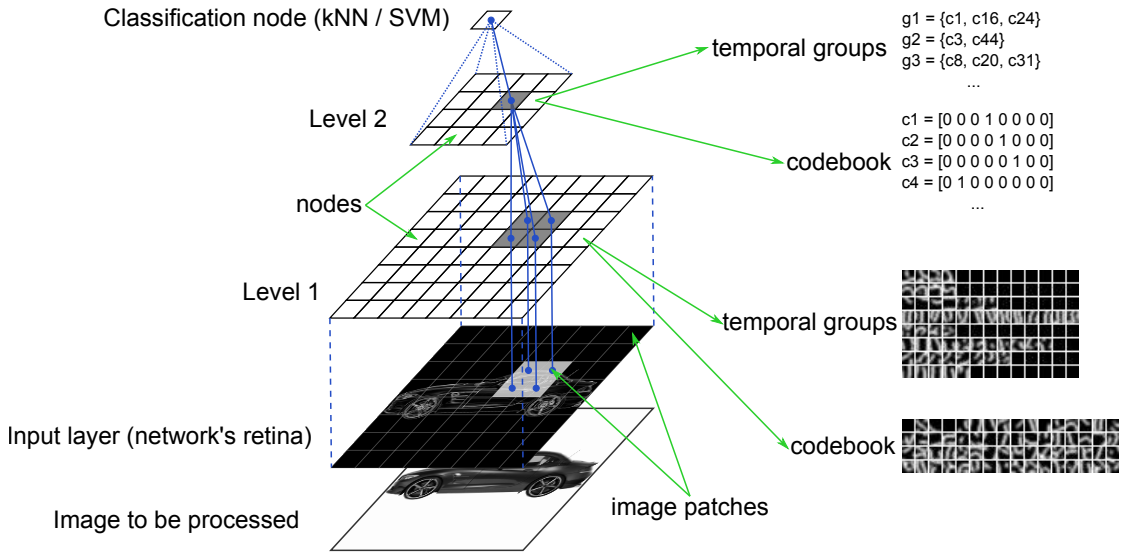

Figure 1: Structure of a 3-layer HTM network with examples of codebook patterns and temporal groups.

described in a form of the first-order Markov graph where vertices represent the memorized coincidences and links stand for the transitions between coincidences in time. The last step of the learning process in each HTM node is to analyze the normalized Markov graph with the aim to partition it into a set of temporal groups. The goal of this partitioning is to group together coincidences (i.e., vertices of the Markov graph) which highly likely follow one another. A node that has completed its learning phase can be switched into the inference mode. In this mode, the node produces an output vector for every input pattern provided. This vector indicates the degree of membership of the input pattern into each of the temporal groups. There are two phases of the inference process, inference in the "spatial pooler" followed by inference in the "temporal pooler".

Typically, most of the input patterns do not perfectly match any of the patterns stored in the node's memory. Let d_i be the Euclidean distance of the i-th stored pattern from the input pattern. The larger is the distance, the smaller should be the match between the input pattern and the stored coincidence. It can be assumed that the match of the patterns can be expressed as a Gaussian function of their Euclidean distance, with the zero mean: $y_i = e^{-d_i^2/\sigma^2}$, where σ is a parameter of the node. By calculating this quantity for all n memorized coincidences, one can produce an overall belief vector $\mathbf{y} = [y_1, y_2, \ldots, y_n]$ that represents closeness of the input pattern to all memorized coincidences [14]. In the second phase of the inference, the temporal pooler makes use of the learned temporal groups and calculates the output vector for the nodes that are above in the HTM hierarchy. Each individual component of the output vector represents belief that $\mathbf{y}$ comes from a particular temporal group.

2.2. *Image features as an input to HTM*

Initial implementations of the HTM networks worked with gray-level input images exclusively. In some cases an input transformation block with Gabor filters was included in the HTM network that produced input in the form of image local feature vectors. On the other side, color images are obvious for CBIR applications. Therefore to apply the HTM network to a CBIR task required to explore possibilities of HTM to deal with color images or features. Consequently, in [7] we explored various combinations of features as input vectors for the HTM network. For color features, we selected simple images converted to one of the selected color spaces and then fed to HTM a vector composing of individual color components for each pixel. As an alternative to the two gray-scale texture features, we used Color Co-occurrence Matrix (CCM), in particular, the Reduced Space variant of the CCM (RSCCM) defined in [26]. We selected six possible color component combinations. Then, for each image patch to be fed as an input to the HTM network, we calculated six RSCCMs for the color combinations. We concentrated ourselves on four color spaces: the standard RGB space, the independent component space I1I2I3, and two perceptually uniform UVW and Lab color spaces. The results of our computer experiments and mutual comparison of the Recognition Accuracy values showed ([22]) that the highest values are reached for two gray-level image features, produced: (i) by application of the Canny edge detector, and (ii) by calculation of the Grey-Level Differences (GLD), which are completed by the color features derived for the Lab color space.

2.3. *Object recognition in color images based on parallel application of three HTM networks*

To improve the recognition accuracy of object detection system based on the HTM network we propose a system that uses information from three HTMs, each focused on different aspect of the images (i.e. group of image features). The selected aspects are edges, texture and color information. For each aspect, we choose the feature with the best recognition accuracy (mentioned in the previous section).

All three HTMs are trained individually, using a different sensory layer, but on the same set of images. After the training, each HTM is used to infer belief vectors from all of the train-

ing images, to create its own set of training samples for k-NN classifier – $\boldsymbol{V}_{\text{edge}}$, $\boldsymbol{V}_{\text{texture}}$, $\boldsymbol{V}_{\text{color}}$.

Then, for the object classification in an input image **I,** the following scheme is proposed:

- the image **I** is inferred separately for each of the feature applied, using the corresponding learned HTM network; as a result three belief vectors $\mathbf{v}\boldsymbol{I}_{\text{edge}}$, $\mathbf{v}\boldsymbol{I}_{\text{texture}}$, $\mathbf{v}\boldsymbol{I}_{\text{color}}$ are obtained,
- for each of these vectors, k nearest neighbors are searched for in the corresponding set of feature training samples (we experimentally determined that the best performance of the k-NN is achieved for $k = 2$); for each of the selected neighboring vectors, only information about its class and distance to the corresponding vector $\mathbf{v}\boldsymbol{I}$ is retained,
- all the nearest neighbors are then analyzed together in order to find out which of them have maximum incidences of identical class membership; the class having this property is declared to be the class to which the image **I** belongs; if two or more classes have the same count, then the output class is chosen out of these classes from which the neighboring vector is the closest to the vector $\mathbf{v}\boldsymbol{I}$.

3. Image data set

For testing the combined system of the saliency map generation with the HTM network, similarly to [9], we created three different image data sets – single object images, partially cluttered images, and fully cluttered images (see Fig. 2). The single object images have resolution of 128×128 pixels, while the partially and fully cluttered images 512×512 pixels.

The first data set consists of images with a single object on a homogenous (black) background. These images were obtained from free image sources from the internet that satisfied the size requirements and had transparent background. The objects are located approximately in the center of the image and they are resized to preserve the aspect ratio of the objects with the aim to occupy maximum possible image area (not protruding outside of the image boundaries). We acquired 300 unique images from 10 object categories, i.e. 30 images per category. For extending the data set by additional images, we applied rotation around the image center and mirroring about the vertical axis, to the original images. The rotation range was ±40º with 10º step. Altogether 3000 images have been obtained (1 original image, 8 rotated image versions, and 1 mirrored exemplar). These individual object images served for the basic cross-validation scheme necessary for computer experiments. The set was 10 times randomly split into two subsets - 60 % constituted a set of training images for the HTM network, while the remaining - 40 % was used as testing images and for the generation of simulated clutter images. The procedure of composing partially and fully cluttered testing images can be described as follows.

We decided to generate four basic types of multiobject clutter images, namely, containing 1, 2, 4, or 8 objects (denoted as the order of the image). For each of these cases, an initial object image is randomly selected of the given set of testing images

Figure 2: Examples of the used images: The top row shows examples of single object images of four different object categories. Both middle and bottom rows represent examples of multiobject images of the 1st (left) and 8th order (right). Partially cluttered images are depicted in the middle row and fully cluttered images (nonoverlapping objects on inhomogeneous background) in the bottom one.

generated in the previous step. The object image is then randomly embedded in a new clutter image having homogeneous or inhomogeneous background. The location of the object image is random but with a restriction not to overlap with other already placed, object images. For the 2nd, 4th and 8th orders, this procedure is repeated corresponding number of times. Similarly to Chikkerur et al [9], we enforced up to one instance of any object class per every generated image. We also generated ground truth images (GTI) containing pixel-by-pixel information on locations of the bounding boxes of the images and their classes of all objects present in cluttered testing images.

The difference between partially and fully cluttered images consists in the background used. In the partially cluttered images, the individual objects are separated by homogenous areas and as a clutter information, the presence of other object images is considered. The fully cluttered images are generated with noisy background that creates additional clutter and makes the segmentation of individual objects more difficult. The heterogeneous background is generated using a combination of colored perlin noise textures of various grain and image blurring.

4. Color image saliency model

As outlined in Introduction, the goal of our research into the object recognition in clutter color images is to combine the HTM network with some model of image saliency explored

within the field of computational methods of visual attention. Several models of image saliency calculation have been mentioned in Introduction, however, due to specificity of clutter images (a number of different objects occurring on inhomogeneous backgrounds), we would need a saliency model that is able to ignore regions with similar visual characteristics and to provide the recognition part of the combined system preferably with the information on true attention regions. We found out that the model of Hu and coauthors, published in [2], can serve a suitable candidate satisfying such a requirement. On the other hand, according to the work of Fukun Bi et al [1], it is possible to improve detection of salient regions in images using an alternative integrated bottom-up model. Namely, the authors suppose that discriminative local regions (DLR) are closely related with spatial entropy. They speculate on biological plausibility of the DLR-based mechanism in relation to the finding that human does not perceive local details of an image at pre-attention stage, however, they are focused on the concentrative distribution of such regions. Exactly DLR-based mechanism of attention can simulate the visual property of spatial entropy. In our approach we attempt to merge the advantages of the both methods in a way capable to generate necessary position of a salient image window to which the HTM network is to be applied. We first outline our modifications to basic steps of both methods of saliency map calculation, and then, based on (Fig. 3), we will describe our approach of merging and extending them.

Hu et al [2] proposed a method of adaptive local context suppression of multiple cues for visual saliency calculation. Similarly, we use the color features defined in the *Lab* color space. In particular, the intensity feature is computed from the L channel, the individual color components are computed from the color channels of this model, and, eventually, the texture feature is computed using all the channels (Fig. 3).

Let us consider an image divided into blocks, called *attention patches*, each containing $p \times q$ pixels. The contrast of a particular feature at the patch centered at (i, j) is calculated using this formula:

$$FV_k(i,j) = \frac{1}{N} \sum_{[u,v] \in (U \times V) \setminus \{[0,0]\}} |MF_k(i,j) - MF_k(i+u, j+v)|,$$

where $U = \{-1, 0, 1\}$, $V = \{-1, 0, 1\}$, are the sets of relative indices of the patch centered at (i, j), and $MF_k(.,.)$ are the means of the $k-th$ feature in the central or the corresponding neighboring patch, and N is the number of patches in its neighborhood. The contrasts at the patch (i, j) for $k = 1, 2, \ldots n$ features/(attention cues) are normalized to the interval $< 0, 1 >$. Then each patch can be represented by the n dimensional feature contrast vector over its neighborhood. To get the *combined feature map*, the individual components of the contrast vector of the given image patch are summed up. The contrast measure for the given set of features is suppressed, if the patch and its neighbors are *similar*. The *similarity* is estimated by the variance of data along eigenvectors of an $n \times n$ covariance matrix. This matrix is formed from the feature contrast vectors at the patch (i, j) and its neighborhood. The eigenvalues $\bar{\lambda}$ of this matrix represent the extent of similarity or dissimilarity among the attention cues. For example, a large eigenvalue indicates a large variance along the direction of its corresponding eigenvector, which implies higher discriminating power [2]. Thus, the suppression factor SF for the patch (i, j) is defined through the product of the eigenvalues: $\tau(i,j) = \prod_{s=1}^{t} \bar{\lambda}_s$, where the eigenvalues $\bar{\lambda}_s$ are sorted in ascending order, and the parameter t (number of accepting eigenvalues) controls the degree of suppression. For obtaining the ultimate saliency value $S(i, j)$ for the patch (i, j), the multiple attention cues represented by the combined map, are modulated by SF according to this formula

$$S(i,j) = \tau(i,j) \cdot \sum_{k=1}^{n} FV_k(i,j).$$

Such a product, calculated in the block, depicted in Fig. 3 as the "saliency map", should contain true Attention Regions.

The authors of [1] argued that the basic property of human visual attention consists in two complementary mechanisms: 1-suppressing the response to frequently occurring features, and 2-simultaneous enhancing unexpected ones. Their algorithm of detecting the saliency from discriminative local regions (DLR map in Fig. 3) simulates this property. We outline only key steps

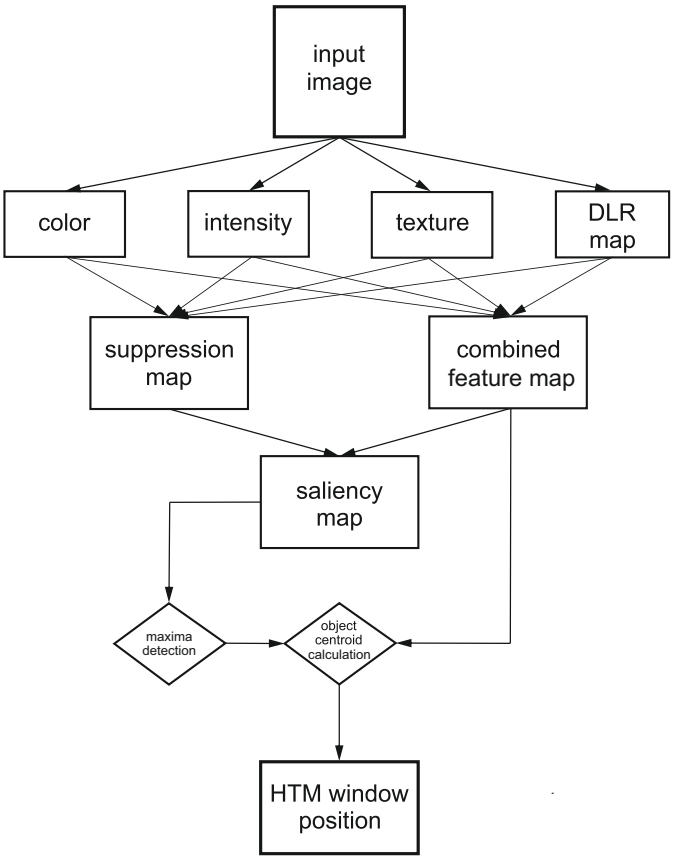

Figure 3: Scheme of the proposed compound algorithm in which two salient region detection methods ([2] and [1]) are combined with the goal to generate an appropriate window position for the application of the HTM to color clutter images in the task of object detection and recognition.

of the DLR map calculation, for details the reader is referred to the original paper [1].

- First, using the standard scale-space approach, DoG images are generated from the intensity map M_I of the given color input image M by repeatedly convolving M_I with the Gaussians,
- the SIFT keypoints are found by detecting local extrema in adjacent scales of DoG images, the $i-th$ keypoint is denoted as $P_i(x,\, y,\, \sigma_{point}\,(i))$, where$(x,\, y)$are coordinates of the keypoint location, and $\sigma_{point}\,(i)$ indicates the keypoint scale,
- as robust saliency detection requires extraction of merely stable keypoints, an iterative process is proposed for repeated image resampling and keypoints re-extraction,
- the final keypoints set $Z^k = \{P_1, P_2, \ldots, P_{N_k}\}$ is generated by k iterations, and the downsampled image M_I^k is eventually reached,
- the parameter σ_T characterizing influence degree within the keypoint neighborhood is defined as $\sigma_T = \sigma_{point} \times \left(size\left(M_I^k\right)/size\left(M_I\right)\right)$, where $size\,(\cdot)$means the image size,
- the range of the applied discrete Gaussian function is denoted by L_T, and given as $L_T = 2 \cdot (2\sigma_T + 1) + 1$,
- then, for each extracted keypoint, a neighborhood template T_{KP} is defined by the formula:

$$T_{KP}\,(x, y, \sigma_T, L_T) = \frac{1}{2\pi\sigma_T^2} exp\left(-\left(x^2+y^2\right)/\left(2\sigma_T^2\right)\right),$$

- the saliency map SM_{DLR} for representation of the saliency of DLR is calculated according to

$$SM_{DLR} = \sum_{i=1}^{N_k} T_{KP}^i\,(x,\, y,\, \sigma_T,\, L_T)\,.$$

Instead of the calculation of Early Visual Features (EVF), used in the paper [1] to combine them with the DLR map, we propose to merge the partial result – maps of three features, generated by the procedure of Hu et al. – with the DLR saliency map SM_{DLR}. Thereby we get a novel combined feature map (SM_{CF}). As the inspiration for such an integration of the DLR map into our attention system, the human face recognition methodology, proposed in [3] and [27], has served: –face and skin detection was added to color, intensity, and texture features – to find true salient regions in images with human faces. However, our system should work with general images, thus, instead of the face and skin detection, we extend the set of initial features by the DLR map. Eventually, two branches of the diagram of our method, depicted in Fig. 3 as A, and B, represent the following operations.

First, regional maxima are detected (branch A) in the saliency map S and their coordinates are used to extract windows $\{W_i\}$ from the combined feature map SM_{CF} ($SM-CF$ in Fig. 4).

The cut-out windows are centered in the local maxima and their size is identical with the size of the square-shaped field of view of HTM.

Second, for each generated window $W_i \subset SM_{CF}$ we calculate its centroid (branch B). Since some objects can generate more than one regional maximum in the saliency map SM_{CF}, we calculate the distance between all centroid pairs and merge those pairs that are closer to each other than a certain threshold. The result of such a merging procedure is a new "centroid" with the coordinates which represent the final position of the HTM window to be extracted from the original input image at this location. The HTM window position is also adjusted do not protrude outside the original image.

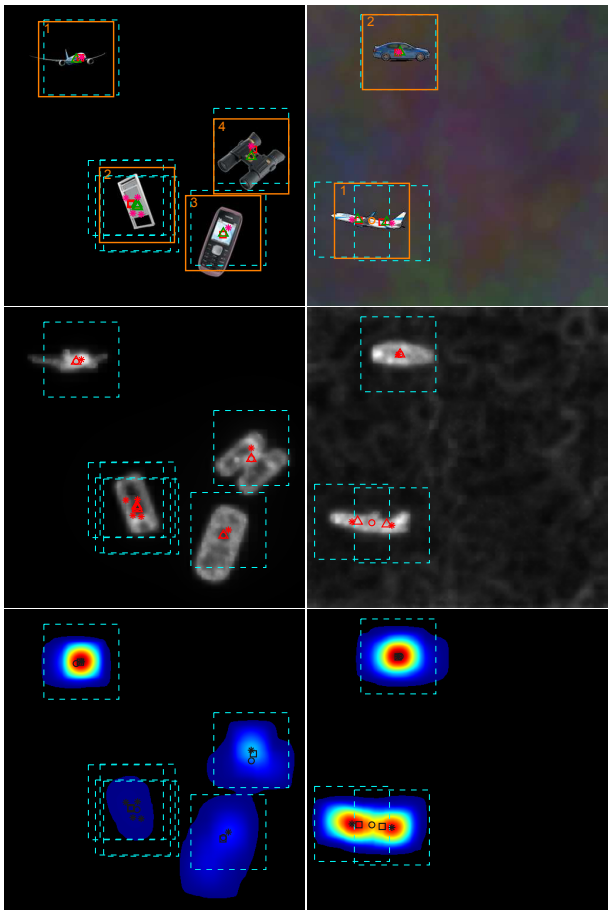

Figure 4: Illustration of the saliency and feature map and processes of the HTM window position calculation. The top row shows examples of two multiobject images from the input data set (left - partially cluttered, right - fully cluttered) with the final HTM cut-out window marked as solid-line squares (each window is numbered in its top left corner). In the middle row the corresponding combined feature maps are visualized, and at the bottom, the final saliency maps are displayed. The asterisks mark the local maxima found in the saliency maps. The dashed rectangles illustrate the cut-outs from the combined feature map, for which the individual centroids (triangles) are calculated (see Section 4). For illustration, the image centroids calculated from the values of the saliency map, are marked as small squares. In the window N.2, occuring in the top left image, the final windows accepted (after centroid merging procedure) as the HTM window, are enhanced. The window N.4 illustrates the situation when the centroid, generated from the feature map, improves the accuracy of the window placement.

5. Computer experiments

5.1. Image object detection and recognition tasks

The visual object detection task is often considered independently of an image object recognition (classification) task. However, there exist computer vision and processing applications in which these two approaches emerge in a concerted and complex task. The proposed compound system, including the image saliency mapping (ISM), as well as the HTM part, is one of such examples of "classifier-based object detection". The evaluation of its performance required a design of specific tools that is addressed in the following subsection 5.2 On the other side, for a comparison of our method to other methods solving this type of the tasks, we needed to choose suitable techniques. In subsection 5.3 we describe two such benchmark techniques. The results of the evaluation of the experiments of both types are discussed in details in Section 6 – Results.

5.2. Design of image object detection and recognition experiments

The compound system, we proposed for object recognition in simulated clutter images, consists of two key subsystems, - the intelligent HTM network, and - the image saliency mapping (ISM) system that is responsible for detecting geometrical locations of individual objects occurring in a model of the complex clutter image. Our primary goal was not to optimize the HTM network with respect to its object recognition performance. The goal was to concentrate ourselves on exploring a principal possibility to link the HTM network with ISM system which would give the HTM multiobject detection capabilities. Therefore we decided to reduce the computational complexity of the experiments by applying the HTM network in its basic mode (one level) that enabled, at the same time, to prepare a consistent platform for its benchmarking with other methods. For evaluation of the performance of the whole system, we proposed the following experimental conditions:

1. first, the evaluation of the performance of the HTM network in its basic mode applied to the task of **single object recognition**,
2. second, the success of the HTM application to clutter images depends not only on the HTM system performance alone, but primarily on the type of image saliency model selected, and on the method of extracting windows of interest in the generated saliency map. Therefore, before evaluating the performance of the whole (ISM+HTM) system, the performance of the ISM system alone had to be evaluated,
3. finally, the overall recognition accuracy was evaluated for the whole system, applied to the task of object recognition in clutter images.

Based on the overall evaluation methodology, we elaborated the following more detailed steps of calculations:

A) to test the performance of the HTM subsystem in ideal conditions, images with the single centered object and no clutter (the input images) are randomly split into training (60%) and testing (40%) sets. After the HTM network is trained on the training set, the testing images are consecutively inferred and classified by the learned HTM. We also generate a set of testing images with random translation of the object by 0 to 10 pixels in either direction and test the recognition performance of the HTM on this set separately. This is done to have the HTMs performance evaluated not only in ideal conditions but also in conditions similar to the image windows generated by the saliency system which are expected to be slightly offset even in the best case scenario. This whole process is repeated 10 times to achieve more consistent results.

B) the objective of the image saliency calculation is to provide a response function, the local maxima of which can indicate areas of object occurrence in the given multiobject image; exactly to such image areas should the HTM subsystem be applied (see Fig.4). For characterization of the accuracy with which the saliency map satisfies this objective, we proposed to use two measures – *localization accuracy* and *overlap*. The *localization accuracy* (*LA*) indicates how well a saliency map pinpoints objects of interest in the images. *The overlap* indicates the portion of the area of the object's square bounding box (defined in GTI) by which it overlaps the cut-out window.

We shortly describe the calculation of these two measures for arbitrary cut-out window generated by ISM system. All these windows have square shape and identical size with the length of the diagonal denoted by $\overline{d}$. First, we identify all such bounding boxes of the individual objects from the GTI, which have nonempty intersection with the given cut-out window. The object whose bounding box has the greatest overlap with the cut-out window represents the dominant object. We consider then the centroid C_o of the dominant object bounding box and the centroid C_w of the given cut-out window. The maximum allowed distance (deviation) of these two centroids can be limited by the threshold $\overline{d}/2$. Since distances $d(C_w, C_o)$ can vary over the set of all simulated objects from GTI, we should normalize them by the maximum allowed distance, i.e. by $\overline{d}/2$. We exclude from the succeeding process the cases for which inequality $d(C_w, C_o) > \overline{d}/2$ holds, and assign the $LA = 0\%$. For other cases, the *localization accuracy* is calculated using the formula:

$$LA = (1 - \frac{d(C_w, C_o)}{\overline{d}/2}) * 100\% \,.$$

We see that the maximum possible *LA*=1 is reached when the centroids C_o and C_w coincide.

The *overlap* is a ratio of the area of the bounding box of the dominant object intersecting the cut-out window to the total area of the bounding box in the original GTI. We also evaluated the *recall* of the attention system that represents the percentage of correctly located objects of all objects present in the image. The *clutter* measure shows the ratio of bounding boxes of all other, non-dominant objects to the area of the cut-out. Finally, we also counted the average number of cut-outs per image, i.e. the number of object locations proposed by the ISM. In ideal case, the number of cut-outs windows per image would be the same as the order of that image.

The overall recognition accuracy for the whole (ISM+HTM) system is calculated as follows. First, the saliency maps are calculated for each of the testing images. Then, the saliency maps and their local maxima are used to generate object candidate square windows. These are afterwards used for calculatation of the final position of the given cut-out window (Fig.4). The boundix boxes of the objects of the GTI are identified with the individual windows. The final cut-out windows are fed to the trained HTM. So, for each cut-out window, the inference is performed using the HTM subsystem. The generated belief vectors are afterwards classified using the algorithm described in Section 2.3. For each image the number of correctly detected and classified objects is calculated. The object detection is valid iff the window is classified as one of the class of the GTI regions that have a non-zero overlap with the window.

5.3. The benchmark experiments for comparison of the (ISM+HTM) system to other techniques

Based on practical reasons (an application of standard and verified software tools), for the benchmark experiments we have selected two algorithms available as toolboxes in the MATLAB software, namely: i) cascade object detectors [28], and ii) the algorithms of object template matching in images [29].

5.3.1. Cascade detectors

The cascade detectors consist of several stages formed by ensembles of weak learners. Each stage is trained using the boosting technique and it labels the currently located region as either positive or negative. Positive indicates that an object was found and negative indicates no objects were found. The stages are designed to reject negative samples as fast as possible. If the label is negative, the classification of this region is complete, and the detector slides the window to the next location. If the label is positive, the classifier passes the region to the next stage. The detector reports an object found at the current window location when the final stage classifies the region as positive. The individual cascades, which tend to be more and more complex with higher stages, can incorporate various image features. In our experiments we used: i) histograms of oriented gradients (*HOG*) [30][31], ii) local binary patterns (*LBP*) [32][33], and iii) "slanted" Haar features [34][35]. A true positive occurs when a positive sample is correctly classified. A false positive occurs when a negative sample is mistakenly classified as positive. A false negative occurs when a positive sample is mistakenly classified as negative. Even if the detector incorrectly labels a non-object as positive, the mistakes can be corrected in subsequent stages.

Using the MATLAB implementation we trained individual detectors for three mentioned features. We used the same set of training images of ten categories as in the case of (ISM+HTM) system, i.e., single dominant objects on homogeneous background for each detector. To avoid the situation when detectors are focused just on the background, we extended the set of negative examples of objects by images consisted of merely black background. Altogether each detector has been trained on 180 positive instances of objects from the given class and 1800 negative instances of other nine classes plus black background. The number of the detector cascades was 4. After the extended testing we have found optimized parameters of the applied algorithms. The results of object detection via cascade detectors applied to the set of multiobject images, introduced in Section 3, obtained for the optimized parameters, are described in Section 6.

5.3.2. Template matching

In the domain of computer vision and image object detection, template matching represents another well-known and established method that is suitable for our benchmark experiments. Provided a set of templates – prototypes of the image objects being detected in an image – is available, this method is based on calculation of the 2D cross-correlation function between the given template and the input image. There are several modifications of the definition of the cross-correlation function. In the MATLAB implementation, we utilized, two definitions are used, i) *the sum of the squared differences* (*SSD*), and ii) *the normalized cross-correlation* (*NCC*). The identical training images to those used for the HTM network and the cascade detectors have been used as templates in our application of the template matching. The organization of the experiments was as follows. For each class of ten classes of image objects used (see Section 3) in experiments with HTM, we calculated the cross-correlation maps $M_i(x, y)$ for all *n* object instances in the class. Then a conjunctive cross-correlation map $MAP(x, y)$ for the selected object (template) class was calculated. The conjunctive map can be computed either as a maximum, that is $MAP(x, y) = max_{i=1}^{n} M_i(x, y)$, or as an average of the individual maps, i.e. $MAP(x, y) = \left[\sum_{i=1}^{n} M_i(x, y)\right] / n$.

6. Results

6.1. Evaluation of the performance of the (ISM+HTM) system

As described in the previous chapter, we first evaluated the HTM's performance on the single object images where HTM achieved 91.13% recognition accuracy (with standard deviation (STD) of 0.76%) and 73.23% (STD 2.99%) on the randomly translated images. We also tested the HTM's performance on the multi-object images of the 1st order without using the attention system, where it achieved recognition accuracy of 10.04% (STD 0.4%), which is roughly equivalent to the random chance (10% in our case – classification into 10 object classes). To apply the HTM system to classification of objects present in clutter images is not adequate, as the system is learned on a set of single object images and it was not intended to be applied directly to the images of such a type.

implementation details can be found at the Matlab web page: http://www.mathworks.com/help/vision/ref/vision.cascadeobjectdetector-class.html

available at the web portal Matlab Central File Exchange: http://www.mathworks.com/matlabcentral/leexchange/24925-fast-robust-template-matching

Next, we evaluated the performance of the attention system, see Table 1. As we expected, the system performs better in images with partial clutter than in fully cluttered images. However, the accuracy was above 80% in all cases, with the maximum absolute average deviation from the true center of the object of 15.25 pixels (considering 128×128 pixel windows) and the minimum absolute average deviation 9.75 pixels. More important is that even though there are some decreasing trends, the accuracy and overlap values are relatively similar regardless of the number of objects present in the image, considering the partially and fully cluttered images alone. The recall values were above 90% and in most cases equal or very close to 100%. This means the system is almost always successful in finding all of the objects in the images. Interestingly, we can see that the recall for images of the 2nd and 4th order is slightly lower for partially cluttered images. The trend is even more noticeable, if one analyzes STD values. We hypothesize that inhomogenous background can sometimes help the objects to stand out more than at homogenous background. This can be seen in Fig. 4. The saliency map of the partially cluttered image (bottom left) contains multiple maxima but only one of them is significant, the remaining regions have very low saliency. However, both regions in the fully cluttered saliency map (bottom right) contains maxima with roughly the same significance. The clutter expectably increases with the image order, but it is at most 10.33%. In all experiments, the number of windows found by the system was approximately the same as the number of the objects per image in that particular data set, even though the system had no a priori knowledge of the image order.

Image Type	Measure	Order of Images			
		1	2	4	8
Partially Cluttered	localization accuracy	88.63 ±7.37	89.23 ±5.55	89.04 ±4.32	87.04 ±3.92
	overlap	90.26 ±6.22	93.09 ±4.82	94.23 ±3.79	93.87 ±3.36
	recall	100.00 ±0.00	98.24 ±9.18	97.62 ±7.55	95.91 ±6.80
	clutter	0.00 ±0.00	0.78 ±3.41	2.91 ±4.62	10.25 ±5.51
	hits	1.11 ±0.32	2.17 ±0.63	4.31 ±1.19	8.53 ±2.35
Fully Cluttered	localization accuracy	83.14 ±11.55	84.09 ±8.54	84.42 ±6.27	83.63 ±4.72
	overlap	85.63 ±9.56	89.22 ±7.38	91.00 ±5.61	91.36 ±4.28
	recall	100.00 ±0.00	99.98 ±0.52	99.48 ±3.59	93.51 ±8.45
	clutter	0.00 ±0.00	0.82 ±3.35	3.07 ±4.59	10.33 ±5.47
	hits	1.22 ±0.46	2.35 ±0.72	4.59 ±1.30	8.55 ±2.44

Table 1: Evaluation of the attention system. Each column represents the values for different order of images.The values of accuracy, overlap, recall, and clutter are shown as percentage; hits represents the absolute number of windows. The values of the standard deviation (STD) are situated below the corresponding characteristic value.

Further, we evaluated the HTM's performance when it was applied to the cut out windows generated by the attention system. The classification accuracy (*CA*) values are shown in Table 2. The best *CA* of 64.65% is achieved in the simplest case – partially cluttered images of the first order – which is 29% decrease compared to centered ideal case images and 11.7% decrease when comparing to the translated single object images. Moreover, the *CA* drops rapidly with increasing image order. As it could be expected, the performance for the fully cluttered images is much lower than for the partially cluttered images. However, even in the worst case – fully cluttered image of the 8th order – the *CA* is almost 3 times higher (29.44%) than the random chance.

Image Type	Order of Images			
	1	2	4	8
Partially cluttered	64.65 ±1.05	51.01 ±1.93	43.60 ±1.43	37.18 ±1.18
Fully cluttered	39.51 ±3.71	34.57 ±3.21	32.32 ±2.49	29.44 ±1.29

Table 2: Classification accuracy-*CA* of the HTM. The individual columns correspond to the specific image order. The values represent the percentual *CA* of the classification with its STD below.

6.2. *Comparison of the image object detection accomplished by the (ISM+HTM) system and by two benchmark techniques*

In the comparison experiments with object detection in multi-images we used two important performance characteristics: "*correct detection rate*" (*CDR*) and "*recall*" (*R*). The correct detection rate is defined as a ratio of the number of correctly detected image objects (true positives) to the number of all detected objects (true positives + false positives). The correct detection rate decreases with increased number of false positives. On the other hand, the recall is defined as a ratio of the number of correctly detected objects in a multi-image to the number of all objects present in the image. Naturally, the increasing number of improperly rejected positions in the image (false negatives) results in decreased recall value.

The figure 5 shows the information flow in the system. Analysis of the flow suggests the following alternative representa-

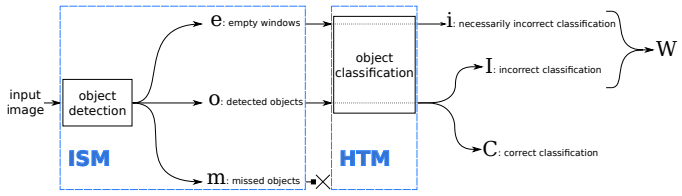

Figure 5: Flow of information in the proposed system composed of ISM and HTM and quantities used for the definition of the evaluation measures: *correct detection rate* CDR and *recall* R.

	cascade detector parameters			
feature	scaleFact	minSize	maxSize	mergeThr
HOG	1.05	108	130	350
LBP	1.05	110	130	360
Haar	1.01	120	130	500

Table 3: The optimized parameters for the cascade detectors applied to multi-image object detection.

tions of the used measures:

$$CDR = \frac{correctly\,classified\,objects}{all\,classified\,objects} = \frac{C}{e+o} = \frac{C}{C+W},$$

$$R = \frac{correctly\,classified\,objects}{all\,objects\,present\,in\,the\,image} = \frac{C}{o+m},$$

therefore, it can be seen that the CDR of the whole system is the same as the classification accuracy of the HTM.

The simulation experiments were organized as a multiple (10-times) random generation of the individual multi-images to which the proposed compound (ISM+HTM) system and both benchmark methods were applied. The generated multi-images were used as ground true images with known location of objects needed for calculations of comparison characteristics. The condition of correct detection of an object, described in Section 5.2, was applied to each cut out window.

6.2.1. *Implementation and evaluation of the cascade detectors performance*

The MATLAB implementation (see Section 5.3.1) of the cascade detectors required setting of four basic controlling parameters. We carried out a number of tests for finding optimum values of these parameters. Although the spatial resolution of testing images was 128×128 pixels, based on the tests we set the parameter $minSize$ (initial expected size of objects) to the value 100, i.e. we used the images of 100×100 pixels. The parameter $scaleFactor$ defines the step of gradual object enlargement, beginning at $minSize$ up to the maximum size $maxSize$. The object detection was repeated at each scale. The parameter $mergeThreshold$ determines the number of scales at which the object has to be detected at the given location to be accepted as object position. The maximum admissible number of erroneous positive detections at each cascade ($FalseAlarmRate$) was set to 10 %. The optimized values of the controlling parameters of the cascade detectors are listed in Table 3.

image type	order	features			ISM+HTM
		HOG	LBP	Haar	
partially cluttered	1	*19.35*	18.28	18.73	**64.65**
	2	*18.57*	17.16	17.90	**51.01**
	4	*19.54*	18.84	17.83	**43.60**
	8	20.78	*23.26*	20.00	**37.18**
fully cluttered	1	4.90	*12.04*	8.13	**39.51**
	2	6.18	*13.54*	10.28	**34.57**
	4	7.60	*14.31*	12.62	**32.32**
	8	9.63	15.13	*15.14*	**29.44**

Table 4: The values of the **CDR** (in %) obtained by three versions of the cascade detectors and by the proposed (ISM+HTM) system for partially and fully cluttered testing multi-images.

image type	order	features			ISM+HTM
		HOG	LBP	Haar	
partially cluttered	1	47.03	42.88	*59.34*	**71.76**
	2	41.05	36.92	*52.48*	**55.35**
	4	35.63	31.99	**47.36**	46.98
	8	27.99	25.15	*36.80*	**39.64**
fully cluttered	1	44.24	39.58	**56.12**	48.20
	2	37.11	35.45	**48.15**	40.62
	4	30.45	31.63	**41.50**	37.09
	8	25.15	25.51	**33.60**	31.46

Table 5: The values of the **recall** (in %) obtained by three versions of the cascade detectors and by the proposed (ISM+HTM) system for partially and fully cluttered testing multi-images.

In Table 4 the values of the correct detection rate achieved by three versions of the cascade detectors for partially cluttered and fully cluttered multi-images are mentioned. The values of the same characteristic reached by the compound system (IMS+HTM) are located in the last column of the table. In all test cases, our system achieved the highest CDR. It should be emphasized that the results achieved by our compound system applied to fully cluttered images are significantly better than for any of the cascade detectors.

In Table 5 the values of the recall are shown which were achieved by three versions of the cascade detectors for partially cluttered and fully cluttered multi-images. The values of the recall reached by the compound system (IMS+HTM) are again located in the last column of the table. In the partially cluttered images, our system achieved the best recall in 3 out of four cases. For the fully cluttered multi-images, the best results were reached by the cascade detector using the Haar feature, however the results of the application of (IMS+HTM) were comparable.

6.2.2. *Implementation and evaluation of the template matching performance*

In the case of template matching we used MATLAB implementation (see Section 5.3.2). This implementation comprises the code of both above mentioned methods NCC and SSD (subsection 5.3.2). For merging the individual maps we applied the average method (denoted as ϕ), as well as maximum method (denoted as M). Altogether we tested four versions of the tem-

plate matching procedures. In the final step of detection, it is necessary to set a suitable threshold for accepting some local maxima of the conjunctive cross-correlation maps for the templates of the given class. If the value of the local maximum of the map is lesser than the threshold, the presence of the object in the detected position is refused. Otherwise, the object is accepted and its center is identified as the position of the corresponding map maximum. For the verification of the correctness of the object detection, the method identical with that used in the case of the (ISM+HTM) system was used. The values of the *CDR* obtained by the template matching methods in two versions for partially and fully cluttered testing multi-images are listed in Table 6. To enable direct comparison we included in the table also the *CDR* values obtained by the (IMS+HTM) system.

image type	order	methods				ISM+HTM
		NCC		SSD		
		ϕ	M	ϕ	M	
partially cluttered	1	10.06	*15.52*	1.94	3.25	**64.65**
	2	13.80	*20.48*	6.46	6.62	**51.01**
	4	17.96	*30.57*	12.39	12.51	**43.60**
	8	21.72	45.52	19.13	22.14	**37.18**
fully cluttered	1	2.51	*6.26*	0.08	1.90	**39.51**
	2	4.15	*9.33*	0.10	3.02	**34.57**
	4	7.04	*14.85*	0.29	5.08	**32.32**
	8	11.08	*21.99*	0.86	9.18	**29.44**

Table 6: The values of the **CDR** (in %) obtained by four versions of the template matching methods and by the proposed (ISM+HTM) system for partially and fully cluttered testing multi-images.

Similarly to the cascade detectors, our system outperformed all template matching methods in all testing conditions. Comparing solely the template matching methods, the the best overall *CDR* was achieved by the (NCC -M methods, both in partially and fully cluttered multi-images.

The values of the recall measure achieved in testing all four versions of the template matching algorithms are included in Table 7. In partially cluttered images, the best performing method was the (NCC-M) template matching. However, as soon as clutter was introduced, its recall dropped below the values achieved by our method.

7. Conclusions

The basic goal of our research was to extend the domain of present applications of the HTM network (exclusively to one object images) to classification of objects located in "clutter" (multi-object) images. We have proposed to enhance the classification functionality of the HTM network by means of a visual attention-based system. Similarly to the HTM inteligent

image type	order	methods				ISM+HTM
		NCC		SSD		
		ϕ	M	ϕ	M	
partially cluttered	1	**100.00**	97.58	31.58	81.67	71.76
	2	69.08	**84.63**	33.54	59.75	55.35
	4	44.92	**69.81**	30.00	43.44	46.98
	8	27.16	**54.10**	23.55	32.50	39.64
fully cluttered	1	32.83	*44.00*	8.25	29.33	**48.20**
	2	27.25	*35.83*	9.34	25.21	**40.62**
	4	23.31	*31.56*	9.77	21.77	**37.09**
	8	17.13	*24.68*	10.31	18.93	**31.46**

Table 7: The values of the **recall** (in %) obtained by four versions of the template matching methods and by the proposed (ISM+HTM) system for partially and fully cluttered testing multi-images.

network, also the interconnection of HTM to ISM was biologically inspired. The basis of this system was adopted from two papers [2] and [1]. We have modified it and appended to it a novel algorithm for the image saliency map calculation. We have also developed a particular scheme for the combination of three parallel HTM networks which can separately process color, texture, and shape information from color images. This scheme is based on the results of our preceding research into optimum features used as input data of the individual HTM networks. We have evaluated the attention system and the HTM's performance separately. The evaluation of the attention system shows promising results in the sense that the system can satisfactorily locate objects in images with several objects on inhomogenous background. Albeit the HTM's performance is significantly lower than in the ideal case (centered single object images), it is still relatively high in most cases, considering the classification into 10 object classes.

Two benchmark methods of object detection in multiobject images have been deployed for comparison of their performance with the performance of the proposed combined system (IMS + HTM). We selected the cascade detectors, as well as the template matching algorithms from the MATLAB set of Toolboxes. For the performance evaluation of all three approaches we used two basic characteristics, the correction detection rate (*CDR*) and Recall (*R*). The obtained results of computer experiments with cluttered multiobject images are summarized in previous Section 6. Based on the detailed discussion of these results we can draw the following conclusions:

- the visual attention system alone achieves satisfactory high values of the recall, for the highest simulated image complexity (eight objects) in the fully cluttered (inhomogeneous background) images, the value $R = 93.51$ is reached; for partially cluttered multi-images (type 8) the highest recall value 54.10 was achieved for template mnatching method (NCC-M), however the recall value 39.64 achieved

by the (IMS+HTM) system was the second one,

- on the other side, for fully cluttered multi-images (type 8) the maximum recall 33.6 was obtained for the cascade detector with Haar basis; the recall value 31.46 obtained for (ISM+HTM) system in this case is quite comparable,
- the values of the CDR of the system (IMS+HTM) were significantly higher than the values reached by both benchmark methods,
- in this paper we did not address the issues related to the optimization of the HTM network itself, even though it turned out that the combination of the HTM with a sophisticated system of visual attention opens possibilities of the HTM application to more complex (clutter) images in tasks of object detection,
- on the other side, there are a number of open issues which can be tackled in the future research, e.g., making the learning phase of the HTM network more robust to the training images shifted in various directions, resulting in increasing the value of *CDR* for the whole (IMS+HTM) system, also more levels of the HTM network could certainly contribute to better generalization of the learning process,
- especially challenging directions of the future research appeared recently, namely, to explore possibilities of cooperating the HTM with a suitable convolution neural networks (CNN), and/or research into relations among IMS, HTM, and CNN in various combinations.

Acknowledgement

This work has been supported by the Slovak Grant Agency for Science (research projects: VEGA 2/0138/16 and VEGA 2/0011/16) and by the European Regional Development Fund under the project Robotics 4 Industry 4.0
(reg. no. CZ.02.1.01/0.0/0.0/15_003/0000470).

References

[1] F. Bi, M. Bian, L. Gao, T. Long, Improvement of salient-region detection using an integrated bottom-up model, in: Proceedings of the 10th IEEE International Conference on Signal Processing (ICSP)), 2010, pp. 836–840.

[2] Z. Hu, D. Rajan, L. Chia, Adaptive local context suppression of multiple cues for salient visual attention detection, in: Proceedings of the IEEE International Conference on Multimedia and Expo, 2005, pp. 1–4.

[3] J. Kučerová, Saliency map augmentation with facial detection, in: The 15th Central European Seminar on Computer Graphics, 2011, pp. 61–66.

[4] T. Deselaers, D. Keysers, H. Ney, Features for image retrieval: An experimental comparison, Journal of Information Retrieval 11 (2008) 77–107, dOI 10.1007/s10791-007-9039-3.

[5] Y. Mistry, D. Ingole, Survey on content based image retrieval systems, International Journal of Innovative Research in Computer and Communication Engineering 1 (8) (2013) 1828–1836.

[6] D. K. Pal, S. S.Tripathy, V. Ranjan, A. Das, Robust content based image retrieval using Hierarchical Temporal Memory with query by example, in: Proceedings of the International Conference on Electronics and Computer Technology (ICECT), Kanyakumari, India, 2013, pp. 1–6.

[7] R. Škoviera, I. Bajla, Image classification based on Hierarchical Temporal Memory and color features, in: J. Manka, M. Tysler, V. Witkovsky, Frollo (Eds.), Proceedings of the 9th International Conference on Measurement, Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Smolenice castle, Slovakia, 2013, pp. 63–66.

[8] U. Rutishauser, D. Walther, C. Koch, P. Perona, Is bottom-up attention useful for object recognition?, in: Proceedings of the IEEE International Conference on CVPR, Vol. 2, 2004, pp. 37–44.

[9] S. Chikkerur, T. Serre, T. Poggio, Attentive processing improves object recognition, Technical report MIT-CSAIL-TR-2009-046, CBCL-279, Computer Science and Artificial Intelligence Laboratory, MIT (October 2009).

[10] S. Han, N. Vasconcelos, Biologically plausible saliency mechanism improve feedforward object recognition, Vision Research 50 (2010) 2295–2307.

[11] W.-B. Yang, B. Fang, Z.-W. Shang, B. Lin, A salient hierarchical model for object recognition, in: Proceedings of the 2012 International Conference on Wavelet Analysis and Pattern Recognition, Xian, 2012, pp. 244–249.

[12] I. Kostavelis, L. Nalpantidis, A. Gasteratos, Object recognition using saliency maps and htm learning, in: Proceedings of the IEEE International Conference on Imaging Systems and Techniques, IST 2012, Manchester, UK, 2012, pp. 528–532.

[13] J. Hawkins, S. Blakeslee, On intelligence, Henry Holt and Company, New York, 2004.

[14] D. George, J. Hawkins, Towards a mathematical theory of cortical microcircuits, PLoS Computational Biology 5 (10) (2009) e1000532. doi:10.1371/journal.pcbi.1000532.

[15] Numenta, Hierarchical Temporal Memory, concepts, theory, and terminology, Document version 1.8.0 (June 2008).

[16] Numenta, Numenta node algorithms guide, NuPIC 1.7 (October 2009).

[17] J. Hartung, J. McCormack, F. Jacobus, Support for the Use of Hierarchical Temporal Memory Systems in Automated Design Evaluation: A First Experiment, ASME Conference Proceedings 2009 (49057) (2009) 853–862.

[18] I. Kostavelis, A. Gasteratos, On the optimization of Hierarchical Temporal Memory, Pattern Recognition Letters 33 (5) (2012) 670–676. doi:10.1016/j.patrec.2011.11.017.

[19] X. Chen, W. Lei, W. Wang, An overview of Hierarchical Temporal Memory: A new neocortex algorithm, in: Proceedings of thye 2012 International Conference on Modelling, Identification and Control (ICMIC), 2012, p. P1004.

[20] F. Sassi, L. Ascari, S. Cagnoni, Classifying human body acceleration patterns using Hierarchical Temporal Memory, in: R. Serra, R. Cucchiara (Eds.), Proceedings of the International Conference AI*IA 2009: Emergent Perspectives in Artificial Intelligence, Berlin, Heidelberg, 2009, pp. 496–505.

[21] S. Štolc, I. Bajla, On the optimum architecture of the biologically inspired Hierarchical Temporal Memory model applied to the hand-written digit recognition, Measurement Science Revue 10 (2) (2010) 28–49.

[22] S. Štolc, I. Bajla, K. Valentín, R. Škoviera, Pair-wise temporal pooling method for rapid training of the htm networks used in computer vision applications, Computing and Informatics 31 (4) (2012) 901–919.

[23] G. E. Hinton, Learning to represent visual input, Philosophical Transactions of the Royal Society B: Biological Sciences 365 (1537) (2010) 177–184.

[24] G. E. Hinton, Deep Belief Networks, Vol. 4 of 5947, Scholarpedia, 2009.

[25] J. Schmidhuber, Deep learning in neural networks: An overview., Neural Networks 61 (2015) 85–117.

[26] A. Porebski, N. Vandenbroucke, L. Macaire, Supervised texture classification: color space or texture feature selection?, Pattern Analysis and Applications 16 (2013) 1–18, dOI 10.1007/s10044-012-0291-9.

[27] R. Vargic, J. Kučerová, J. Polec, Wavelet-based image coding using saliency map, Journal of Electronic Imaging 6 (2016) 061610–1 – 061610–10.

[28] P. A. Viola, M. J. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the International Conference CVPR,

Vol. 1, IEEE Computer Society, 2001, pp. 511–518.

[29] R. O. Duda, P. E. Hart, D.G.Stork, Pattern classification, Vol. second edition, John Wiley and Sons, INC., 2001.

[30] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the International Conference on Computer Vision and Pattern Recognition, CVPR 2005, Vol. 1, IEEE Computer Society, 2005, pp. 886–893.

[31] Q. Zhu, M.-C. Yeh, K.-T. Cheng, S. Avidan, Fast human detection using a cascade of histograms of oriented gradients., in: Proceedings of the Computer Vision and Pattern Recognition Conference, Vol. 2, IEEE Computer Society, 2006, pp. 1491–1498.

[32] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns., IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI 7 (24) (2002) 971–987.

[33] T. Ojala, Nonparametric texture analysis using spatial operators, with applications in visual inspection., Phd thesis, Department of Electrical Engineering, University of Oulu, Finland (1997).

[34] R. Lienhart, J. Maydt, An extended set of haar-like features for rapid object detection, in: Proceedings of the International Conference ICIP, 2002, pp. 900–903.

[35] R. Lienhart, A. Kuranov, V. Pisarevsky, Empirical analysis of detection cascades of boosted classifiers for rapid object detection, in: M. Bernd, K. Gerald (Eds.), DAGM-Symposium, Lecture Notes in Computer Science, Vol. 2781, Springer, 2003, pp. 297–304.