U.P.B. Sci. Bull., Series A, Vol. 72, Iss. 1, 2010 ISSN 1223-7027

CELLULAR AUTOMATA BAG OF VISUAL WORDS FOR OBJECT RECOGNITION

Ionuţ Mironică[1], Bogdan Ionescu[2], Radu Dogaru[3]

În această lucrare vom propune și analiza o nouă metodă pentru rezolvarea problemei de recunoaștere de obiecte, algoritmul fiind inspirat din teoria automatelor celulare. Metoda propusă este eficientă și are un cost computațional scăzut, fiind încorporată în cadrul modelului standard Bag-of-Visual-Words. Testele experimentale efectuate pe o serie de baze de imagini standard au arătat că metoda propusă obține o performanță superioară altor algoritmi existenți din literatură.

In this paper we propose and analyze a novel method for object recognition, inspired by the cellular automata theory. The proposed method has a low computational complexity, and can be incorporated in the standard Bag-of-Visual-Words framework. Experimental tests conducted on several standard image databases show that the proposed method provides a significant improvement in the classification performance, outperforming some other classic approaches.

Keywords: object recognition, image classification, local descriptors, Bag-of-Visual-Word.

1. Introduction

During the last 20 years Content Based Image Retrieval (CBIR) has been steadily gaining importance in the computer vision community. This issue became more critical due to the extent development of technology, e.g., portable multimedia terminals, wireless transmission protocols, imaging devices which basically unlimited the access to information everywhere. Information retrieval has become now a part of our daily social interaction.

The main idea behind CBIR is to compactly describe an image by using a digital signature which best represent the underlaying visual contents. These descriptors are to be stored by the system and then used to match a

[1]Postdoc Researcher, LAPI, University Politehnica of Bucharest, Romania, e-mail: imironica@imag.pub.ro

[2]Associate Professor, LAPI, University Politehnica of Bucharest, Romania, e-mail: bionescu@imag.pub.ro

[3]Professor, ETTI, University Politehnica of Bucharest, Romania, e-mail: radu_d@ieee.org

user query image to the most resembling image within the data set (e.g., Internet, databases, etc). This is carried out by employing some similarity criteria [1]. Due to the subjective nature of the process, the system typically provides the user with not only one response but a ranked list of possible choices to select from. A major part of the CBIR work focused on object recognition in images. Generally, the object recognition problem can be regarded as a labeling problem based on models of known objects. Formally, given an image containing one or more objects of interest (and background) and a set of labels corresponding to a set of models known to the system, the system should assign correct labels to regions, or a set of regions, in the image. The object recognition involves a various number of difficult problems: starting from the fact that objects may vary somewhat in different view points, in many different sizes / scale or even when they are translated or rotated to illumination challenges. Also, the objects should be recognized even when they are partially obstructed from view.

Today, most of common object detection techniques are based on local features, which have been widely used and demonstrated high performance rates. Actually, the most representative approach of this family, the Bag-of-Visual-Words (BoVW) algorithm [2] became a facto standard for image retrieval and recognition. BoVW models are very popular in object recognition due to their robustness to noise and occlusions.

The remainder of the paper is organized as follows. Section 2 presents a state-of-the art of the literature and situates our work accordingly. Section 3 depicts the algorithm of the proposed approach. Experimental validation is presented in Section 4 while Section 5 presents the conclusions and discusses future work.

2. **Previous Work**

Image classification remains one of the most challenging problems, mainly, because it implies to predict complex semantic categories, like scenes or objects, from raw pixels. The are several major approaches that have emerged in the last decade towards this goal. The first is the design of global descriptors, e.g., image histograms. However, this class of approaches obtain low results for complex object categories, mainly, because these are not invariant to object variations, such us illumination, rotation, shape, color or translation.

The second class of algorithms is represented by the notion of mid-level representations inspired from the text retrieval community, based on the Bag-of-Visual-Words model [2]. The BoVW model can be applied to image classification, by treating image features as words. The typical BoVW model works as follow. Firstly, the algorithm identifies a list of local patches from the image, either by densely sampling [3] or by a interest point detector [4]. These local patches, represented by vectors in a high dimensional space are often called keypoints. Next step for the BoVW model is to generate a codebook or dictionary (analogy to a word dictionary). A codebook can be considered

as a sum of the most representative keypoints. One simple method to create the dictionary is to perform a simple k-means clustering over all the training keypoints and the codewords are then defined as the centers of the learned clusters. To reduce the high computational cost of k-means, several methods were proposed, namely hierarchical k-means in [11] and random forests in [3]. In order to efficiently handle these key points, the key idea is to quantize each extracted keypoint into one of visual words (from a already created dictionary), and then to represent each image by a histogram of the visual words. This vector quantization procedure allows to represent each image by a histogram of the visual words, which is often referred as the bag-of-words representation, and consequently converts the object categorization problem into a text categorization problem.

Many different techniques for describing local feature keypoints have been developed. The simplest descriptor is a vector of image pixels [12]. However, the high dimensionality of such a description results in a high computational complexity for recognition and low recognition rates.

Lowe [4] proposed a scale invariant feature transform (SIFT). Today, the SIFT descriptor represents a standard for the local image description. The compute of SIFT descriptor consists in several steps. First, a set of orientation histograms is created on 4x4 pixel neighborhoods with 8 bins each. These histograms are computed from magnitude and orientation values of samples in a 16 x 16 region around the keypoint such that each histogram contains samples from a 4 x 4 subregion of the original neighborhood region. The magnitudes are further weighted by a Gaussian function with standard deviation σ equal to one half the width of the descriptor window. Finally, the descriptor becomes a vector of all the values of these histograms.

In [13], Ke et al. proposed the use of PCA-SIFT. In other words, PCA-SIFT uses Principal Component Analysis (PCA) to normalize and decrease the SIFT feature. The feature vector is significantly smaller than the standard SIFT feature vector, and it can be used with the same matching algorithms.

Another robust local feature representation was proposed by Herbert Bay et al. in 2008 [14]. Speeded Up Robust Features (SURF) uses the sum of the Haar wavelet responses around the point of interest, that can be calculated very fast with the aid of the integral image. Mikolajczyk and Schmid [15] use a multi-scale version of the Harris interest point detector to localize interest points in space and then employ the Harris [16] algorithm for scale selection and affine adaptation.

In this paper we propose a strategy for building local feature descriptors that capture local information by using the Cellular Automata (CA) theory [7]. The CA theory have been successfully applied for many image processing fields from edge detection [8] to skin detection [9], but, to our knowledge, the CA have never been used in object recognition using a similar framework. This idea was exploited in [10] but in the context of texture categorization.

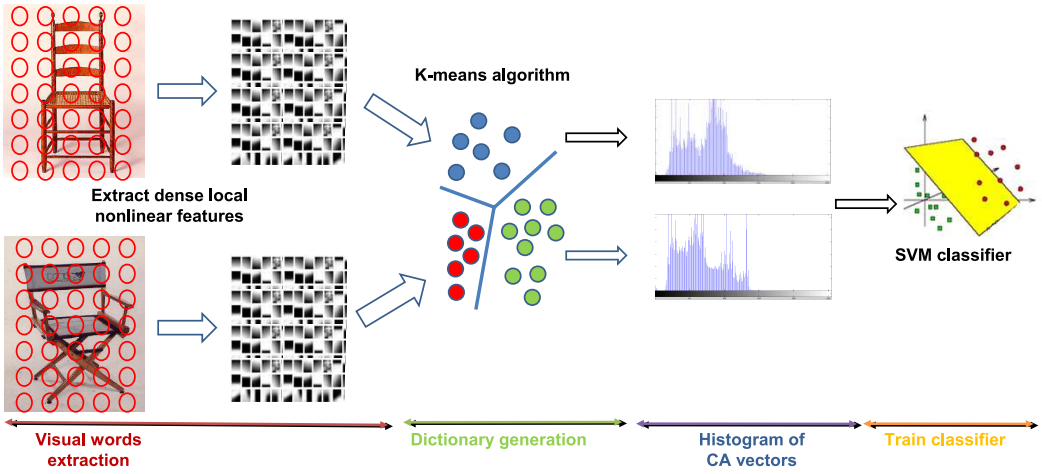

FIGURE 1. The proposed object recognition algorithm.

However, the texture processing is a fundamentally different and more easier problem because the texture categories contains a lower intra variability class. The main difference between these approaches is that the proposed algorithm is applied as a local descriptor in a Bag-of-Visual-Words framework, while in [10] a global image feature representation is created.

Our main contributions are summarized as follows: (1) we propose a new robust local feature for object categorization; (2) we incorporate the new features in a general and flexible learning framework, namely the Bag-of-Visual-Words representation and apply a popular nonlinear SVM classifier on our descriptors to classify the objects; (3) we achieve promising results comparable to the state-of-the-art results on Caltech101 [5] and Caltech256 [6], and significant improvements over state-of-the-art SIFT features.

3. Proposed method

3.1. The Bag-of-Visual Words architecture

The architecture of the proposed system is presented in Figure 1, and it consists of four different layers. First, we extract a list of image keypoints using a dense sampling strategy, as proposed in [5]. For each keypoint, we compute a local descriptor. A classical approach to compute local keypoints is to calculate SIFT [4] or SURF [14] descriptors. SIFT had proven high stability in many situations, but with slow computational speed and high sensitivity at illumination changes. SURF provides faster speed but also it has many drawbacks, e.g., it is not stable to rotation and illumination changes. In order to suppress these disadvantages we propose a new local non-linear feature, inspired from the cellular automata theory (presented in Section 3.**2**).

Then, we create a dictionary of visual words. By using the k-means algorithm, the sampled features are clustered in order to quantize the space

into a discrete number of visual words. We use a visual vocabulary of 4,096 words (which represent a common value for video related tasks and gives good results on both the TRECVID and Pascal VOC datasets [17, 18]) and final descriptors are represented at two different spatial scales of a spatial pyramidal image representation (entire image and quadrants) [27]. Afterwards, for each image a global histogram descriptor is computed, by assigning each visual word to the nearest cluster center from the previous step.

The final layer is represented by the classification algorithm. Support Vector Machines (SVMs) are a very popular classifier due to its robustness against large feature vectors and sparse data. The choice of SVM-kernel has a large impact on performance. In our experiments, we test three types of SVM kernels: a fast linear kernel and two non-linear kernels, namely the RBF and Chi-Square kernels. While linear SVMs are very fast in both training and testing, SVMs with an non-linear kernel are more accurate in many classification tasks.

3.2. Cellular automata approach for local feature description

A cellular automata (CA) represents a discrete model and consists of a regular grid of cells, that contains a finite number of states, such as on and off (or "0" and "1"). Also, in more complex simulations the cells may have more different states and the grid can have any finite number of dimensions.

Our algorithm is inspired by the cellular automata theory [7]. The proposed cellular structure consists on a Moore Neighborhood Model (see Figure 2 (a)) with two distinct states: 0 and 1. The red cell is the center cell, the blue cells are the neighborhood cells. The states of these cells are used to calculate the next state of the (red) center cell according to the defined rule.

Therefore, the first task is to transform the local patch in a binary lattice. To create binary images, we use a thresholding process with a various number of limits. During the thresholding process, individual pixels in an image are are given a value of "1" if their value is greater than some threshold value and as "0" otherwise. We have used in our experiments a fixed number of equally spaced thresholds T (from one to 10 thresholds). Afterwards, by using these binary images, we compute a global feature, by using the following formula:

$$C = \frac{1}{N \times M} \sum_{i=1}^{M} \left[\sum_{j=1}^{N} |F(i,j)| \right] \tag{1}$$

where M and N are the image width and height (see more details about the image size in Section 4) and $F(i,j)$ is a kernel function, computed on current pixel neighborhood. The kernel function is defines as:

$$F(i,j) = \sum_{k \in N_{i,j}} |I_{i,j}(k) A(k)| \tag{2}$$

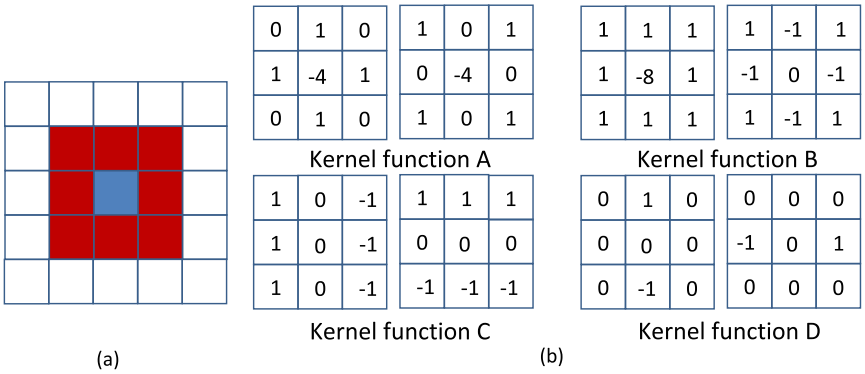

FIGURE 2. (a) The Moore neighborhood for a cellular automata (b) Four function kernels used for the computing of the non-linear parameters.

where $N_{i,j}$ represents the 3×3 neighborhood centered around the (i, j) location, $I_{i,j}(k)$ is a pixel value at location k ($k = 1..9$) in the neighborhood centered on (i, j) and $A(k)$ is one of the 3×3 template function presented in Figure 2 (b). Finally, the features size becomes $K \times T$, where T represents the number of the thresholds and K is the number of kernel functions (2 for our scenario).

The use of presented architecture contains several motivations to implement it in a general object recognition framework. It was demonstrated in [28], that a value of C close to 1 indicates a homogeneous state while a value of $C = 0.5$ is a measure of a perfect (high frequency) chaotic pattern. At the other extreme $C = 0$ indicates the presence of perfectly regular chess-board pattern. Consequently, such synthetic indicators as C are strongly correlated with the human perception. Using various templates ensures that various directions of interests in the image are better characterized.

Our key idea is to define a local feature that, instead of being composed of a single SIFT descriptor, is a multi-resolution set of descriptors. By using several thresholds and non-linear functions, it allows us to capture the appearance of a local patch at multiple levels of detail and to maintain the distinctiveness, all while preserving invariance at each level of resolution.

4. **Experimental results**

The validation of the proposed object recognition approach was carried out on two standard image datasets: *Caltech-101* [5] and *Caltech-256* [6].

The Caltech-101 dataset (collected by Fei-Fei et al. [5]) consists of 9,144 images from obtained using Google Image Search. The dataset includes 101 object categories (such as animals, vehicles, flowers and objects), and contains from 31 to 800 images per category. Most images have medium resolution, about 300x300 pixels. The significance of this database is its large inter-class variability. As suggested by the original dataset and also by many other

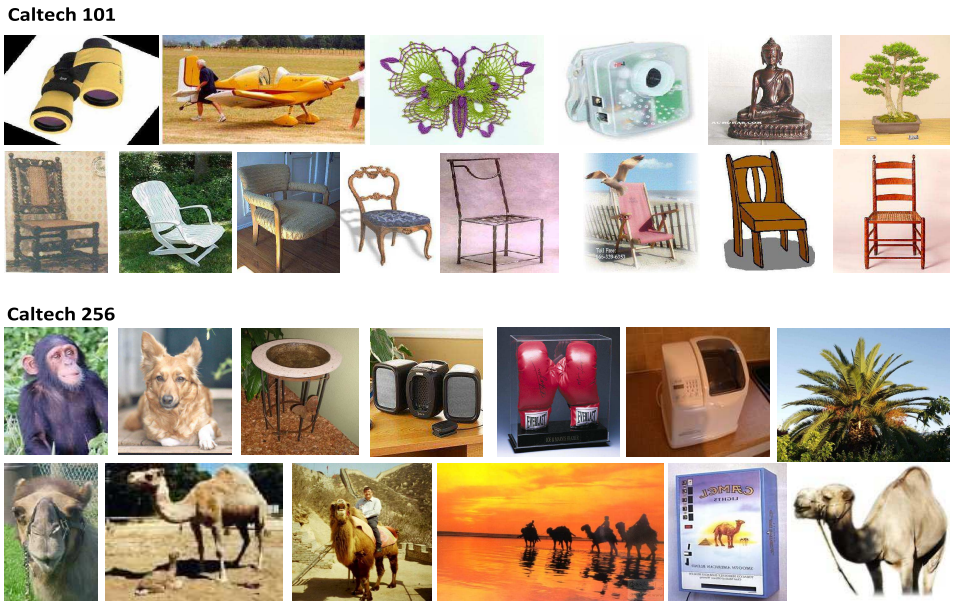

FIGURE 3. Sample images from the *Caltech101* [5] and *Caltech256* [6] datasets.

researchers [20, 21, 22], we partitioned the whole dataset into 10, 20 and 30 training images per class and no more than 50 testing images per class, and measured the performance using average accuracy over 102 classes (i.e., 101 classes and a "background" class).

The Caltech-256 dataset (collected by Griffin et al. [6]) consists of images from 256 object categories and is an extension of Caltech-101. It contains from 80 to 827 images per category. The total number of images is 30,608 with a average resolution of 300x300pixels. The significance of this database is its large inter-class variability, as well as a larger intra-class variability than in Caltech-101. Moreover there is no alignment amongst the object categories. Our experimental results are generated by using 10, 20 or 30 images per category for training, while for testing, we used 50 images per category. These number of train and test images are typically used for this dataset [6, 23, 24].

These datasets have a significant variation in the position of the object instances within images of the same category, and also different background clutter between images. Figure 3 illustrates some image examples in this respect. In order to asses the system performance, we use the accuracy measure, which represents the official metric for the Caltech-101 and Caltech-256 datasets.

4.1. Choose of the parameters

In this experiment we analyze the influence of the parameters on the system performance. We perform the optimization of the parameters on a quarter of the Caltech-101 dataset.

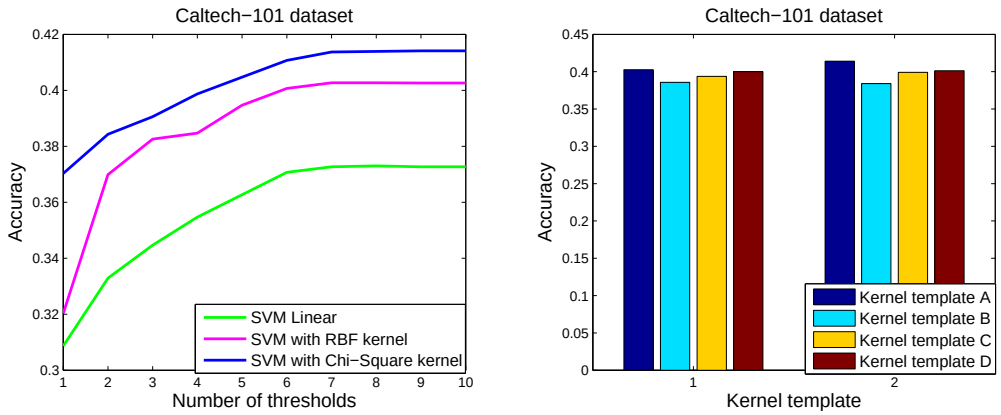

FIGURE 4. The influence of parameters on the system performance: (1) the variation of the thresholds number and the SVM kernel and (2) the performance of various template functions.

In the first test we study the influence of the number of thresholds (see Section 3.**2**). The results are exposed in Figure 4.1. It can be observed that the best results are obtained using 7 thresholds. After this value, the performance remains constant.

The second parameter is represented by the choice of kernel functions. We tested four different kernel functions (see Figure 4.2.). The first kernel function obtain the best results. Also, it can be observed that a dense sampling strategy leads to better performance with more than 8 percents.

The last parameter that has to be taken into consideration is the SVM kernel. Initial experiments showed that we obtained better results with non-linear Chi-Square kernel (see Figure 4.1.). In the next experiments, we will use the non-linear Chi-Square kernel.

4.**2**. **Comparison with SIFT features**

In this section, we compare the performance of the proposed approach with the classical SIFT local features. We perform this comparison on the entire Caltech-101 dataset. The values are presented in Table 1. The proposed approach tends to provide better retrieval performance in all cases (see bold values). Also, the proposed keypoints descriptor is more compact. The feature size of our approach is equal to 14 values (7 thresholds $\times$ 2 kernel functions), while the SIFT representation is 9 times bigger (128 values).

Overall, for the Caltech-101 dataset, the BoVW with SIFT features and SVM non-linear kernels provide good results, but the proposed approach is better. The presented algorithm obtains an accuracy of 73.76%, while the BoVW-SIFT approaches is lower with more than 9 percents. At the other end, the smallest performance is obtained using BoVW with Linear SVM kernel.

TABLE 1. Comparison of accuracy between SIFT features the proposed approach (the highest values are depicted in bold).

Method	Accuracy
Caltech-101 dataset	
BoVW - SIFT [2] - SVM with Linear kernel	58.12%
BoVW - SIFT [2] - SVM with RBF kernel	62.88%
BoVW - SIFT [2] - SVM with Chi-Square kernel	64.17%
BoVW with CA features (this paper)	**73.76%**

To conclude, using the CA features to model the keypoints improves the BoVW model, yielding much better results than the state-of-the-art SIFT features on the Caltech-101 dataset.

4.3. Comparison with state-of-the-art

In order to compare our algorithm with other object recognition approaches, we have selected the settings that provides the greatest improvement in performance: seven thresholds and SVM classifier with Chi-Square kernel. All the experiments from this section are performed on the entire Caltech-101 and Caltech-256 datasets.

In the following, we compare our approach against other validated algorithms from the literature. In [20], the authors proposes a new convolutional factor analysis model that uses a deep feature learning. Aflalo et al. [21] present a novel algorithm based on Multiple Kernel Learning (MKL) strategies. Also, Ma et al. [22] suggested a computational system of object categorization based on decomposition and adaptive fusion of visual information. A coupled Conditional Random Field is developed to model the interaction between low level cues of contour and texture, and to decompose contour and texture in natural images. In order to improve the BoVW model, Germert et al. [23] proposes a new kernel codebook strategy, and Wang et al. [25] use a new Locality-constrained Linear Coding algorithm.

The results can be visualized in Table 2. As can be seen, our approach yields the highest accuracy for both dataset, namely 73.67% for Caltech-101 dataset and 42.12% for the Caltech-256 dataset. This shows that the proposed local representation is effective for a object recognition approach. We conclude that our framework yields better performance than other state-of-the-art algorithms. Figure 5 presents several system responses, when we use the proposed system configuration: first four lines represent the true positives (TP) examples in which the object found by the system are correctly identified according to the ground truth (note the scenario difficulty, different fields of view, object dimension, different object color, illumination, camera noise and other objects around the object of interest). Anyway there are also false negatives situations

TABLE 2. Accuracy for various state-of-the-art algorithms for Caltech-101 [5] and Caltech-256 [6] datasets (the highest values are depicted in bold).

Method	Accuracy
Caltech-101 dataset	
Chen et al. (ICML) [20]	65.80%
Aflalo et al. (ML) [21]	67.07%
Ma et al. (CVPR) [22]	70.38%
This paper	**73.67%**
Caltech-256 dataset	
van Gemert et al. (ECCV) [23]	27.17%
Yang et al. (CVPR) [24]	34.02%
Griffin et al. [6]	34.10%
Wang et al. (CVPR) [25]	41.19%
This paper	**42.12%**

(line five and six) in which the system is unable to classify correctly (according to ground truth) the object detected due to the similarity between classes.

5. Conclusions

We demonstrate how our local descriptors improve image classification results for a standard Bag-of-Visual-Words approach. Experimental results on a wide spectrum of benchmark problems suggest that given its simplicity, our approach may be a good alternative for local keypoints descriptors. In all the experiments our approach achieves the best results on recognition scenarios. Also, the proposed features demonstrate their robustness, compact representation and significance for human perception.

Future work will mainly consist of fine tuning and adapt the method to address to object recognition and event detection in video documents. Also, we will try to integrate the algorithm in other object recognition framework, such as Fisher kernel.

Acknowledgment

This work has been supported by the ESF POSDRU/159/1.5/S/132395 InnoRESEARCH programme.

References

[1] *A. W. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain*, Content-based Image Retrieval at the End of the Early years, IEEE Transansactions on Pattern Analysis Machine Intelligence (PAMI), vol. **22(12)**, pp. 1349-1380, 2000.

[2] *G. Csurka, C. R. Dance, L. Fan, J. Willamowski, C. Bray*, Visual Categorization with Bags of Keypoints, European Conference on Computer Vision (ECCV), pp. 1-2, 2004.

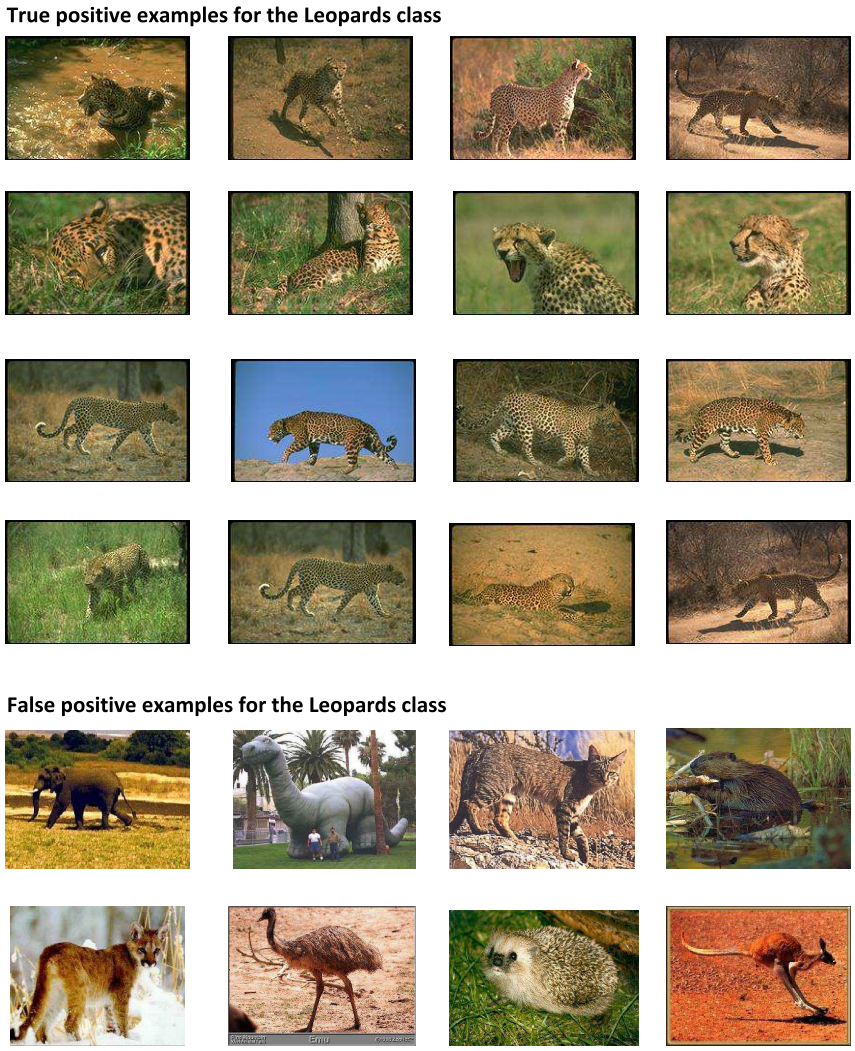

FIGURE 5. Examples of system classification responses: first four lines represent examples in which the object found by the system are correctly identified according to the ground truth, and the last two lines provide false negative examples (images from the Caltech-101 dataset [5]).

[3] *J.R.R. Uijlings, A.W.M. Smeulders, R.J.H. Scha*, Real-Time Visual Concept Classification, In IEEE Transactions on Multimedia, vol. **12(7)**, pp. 665-681. 2010.

[4] *D. Lowe*, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, vol. **2(60)**, pp. 91-110, 2004.

[5] *L. Fei-Fei, R. Fergus, P. Perona*, Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories, CVPR 2004, Workshop on Generative-Model Based Vision. 2004.

[6] *G. Griffin, A. Holub, P. Perona*, Caltech-256 object category dataset, 2007.

[7] *S. Wolfram*, Statistical mechanics of cellular automata, Reviews of Modern Physics, **55(3)**, pp. 601-644, 1983.
[8] *C. L. Chang, Y. J. Zhang, Y. Y. Gdong*, Cellular automata for edge detection of images, In Machine Learning and Cybernetics, vol. **6**, pp. 3830-3834, 2004.
[9] *A. A. Ahmad, F. Mehran, S. Kasaei*, A new dynamic cellular learning automata-based skin detector, in Multimedia systems, vol. **15(5)**, pp. 309-323, 2009.
[10] *Ionuţ Mironică, Radu Dogaru*, A novel feature-extraction algorithm for efficient classification of texture images, in Scientific Bulletin of UPB, Seria C - Electrical Engineering, vol. **75(2)**, pp. 101-114, ISSN 2286 - 3540, 2013.
[11] *D. Nister, H. Stewenius*, Scalable recognition with a vocabulary tree. In Computer Vision and Pattern Recognition (CVPR), pp. 2161-2168, 2006.
[12] *K. Mikolajczyk, C. Schmid*, A performance evaluation of local descriptors, Pattern Analysis and Machine Intelligence, IEEE Transactions on **27(10)**, pp. 1615-1630, 2005.
[13] *Y. Ke, R. Sukthankar*, PCA-SIFT: A More Distinctive Representation for Local Image Descriptors, In Computer Vision and Pattern Recognition, pp. 511-517, 2004.
[14] *H. Bay, A. Ess, T. Tuytelaars, L. Van Gool*, Speeded-up robust features (SURF). Computer vision and image understanding, vol. **110(3)**, pp. 346-359, 2009.
[15] *K. Mikolajczyk, C. Schmid*, An affine invariant interest point detector, In Proceedings of the 7th European Conference on Computer Vision, pp. 128-142, 2002.
[16] *C. Harris, M. Stephens*, A combined corner and edge detector, In Alvey Vision Conference, vol. **15**, 1988.
[17] *P. Over, G. Awad, J. Fiscus, B. Antonishek, M. Martial, F. Smeaton, W. Kraaij, G. Qunot*, TRECVID 2011-an overview of the goals, tasks, data, evaluation mechanisms and metrics, In TRECVID 2011-TREC Video Retrieval Evaluation Online. 2011.
[18] *M. Marszalek, C. Schmid, H. Harzallah, and J. van de Weijer*, Learning representations for visual object class recognition, ICCV Pascal VOC 2007 Challenge Workshop, 2007.
[19] *S. Lazebnik, C. Schmid, J. Ponce*, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, CPVR, 2006.
[20] *B. Chen, G. Polatkan, G. Sapiro, L. Carin, D. B. Dunson*, The hierarchical beta process for convolutional factor analysis and deep learning, In Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 361-368, 2011.
[21] *J. Aflalo, A. Ben-Tal, C. Bhattacharyya, J. S. Nath, S. Raman*, Variable sparsity kernel learning. The Journal of Machine Learning Research, vol. **12**, pp. 565-592, 2011.
[22] *X. Ma, W. E. L. Grimson*, Learning coupled conditional random field for image decomposition with application on object categorization, In Computer Vision and Pattern Recognition (CVPR), 2008.
[23] *J. C. van Gemert, J. M. Geusebroek, C. J. Veenman, A. W. Smeulders*, Kernel codebooks for scene categorization, In Computer Vision (ECCV), pp. 696-709, 2008.
[24] *J. Yang, K. Yu, Y. Gong, T. Huang*, Linear spatial pyramid matching using sparse coding for image classification, In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1794-1801, 2009.
[25] *J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong*, Locality-constrained linear coding for image classification, In CVPR, pp. 3360-3367, 2010.
[26] http://trec.nist.gov
[27] *S. Lazebnik, C. Schmid, J. Ponce*, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, In CVPR, pp. 2169-2178, 2006.
[28] *R. Dogaru, R. Tetzlaff, M. Glesner*, Semi-Totalistic CNN Genes for Compact Image Compression , In Cellular Neural Networks and Their Applications (CNNA), 2006.