BULETINUL INSTITUTULUI POLITEHNIC DIN IAŞI
Publicat de
Universitatea Tehnică „Gheorghe Asachi” din Iaşi
Volumul 67 (71), Numărul 3, 2021
Secţia
ELECTROTEHNICĂ. ENERGETICĂ. ELECTRONICĂ
DOI:10.2478/bipie-2021-0017

SHORT LITERATURE REVIEW FOR VISUAL SCENE UNDERSTANDING

BY

ȘTEFAN-DANIEL ACHIREI*

“Gheorghe Asachi” Technical University of Iaşi,
Faculty of Automatic Control and Computer Engineering, Iași, Romania

Received: October 11, 2021
Accepted for publication: December 5, 2021

Abstract. Individuals are highly accurate for visually understanding natural scenes. By extracting and extrapolating data we reach the highest stage of scene understanding. In the past few years it proved to be an essential part in computer vision applications. It goes further than object detection by bringing machine perceiving closer to the human one: integrates meaningful information and extracts semantic relationships and patterns. Researchers in computer vision focused on scene understanding algorithms, the aim being to obtain semantic knowledge from the environment and determine the properties of objects and the relations between them. For applications in robotics, gaming, assisted living, augmented reality, etc a fundamental task is to be aware of spatial position and capture depth information. First part of this paper focuses on deep learning solutions for scene recognition following the main leads: low-level features and object detection. In the second part we present extensively the most relevant datasets for visual scene understanding. We take into consideration both directions having in mind future applications.

Keywords: object detection; relationship detection; classification.

*Corresponding author; *e-mail*: stefan-daniel.achirei@academic.tuiasi.ro

1. Introduction

The universal functionality of computer vision in order to understand a scene starts with detection, localizing, recognizing and in the end understanding it (Aarthi *et al*., 2017). Besides detecting some of the visual features like edges or corners, the system requires innovative computer vision functionalities in order to learn, compare results and adapt the feedback loop. This process is meant to improve the analysis and result interpretation. Ideally, a system designed for scene understanding should easily adapt to new and varied environments, even anticipate. Communication with other systems and interaction with humans is also an important task. A few features which can be interpreted as visual information are: color, luminance, contour, shape, texture and semantic context (Aarthi *et al*., 2017). The path of scene understanding involves a few main directions: perception and awareness, maintaining the consistency, recognizing events, continuous evaluation and learning, extracting knowledge from collected data (Aarthi *et al*., 2017).

A prediction example of scene understanding for the outdoor environment image shown in Fig. 1 is: *beach (0.301), coast (0.214), beach house (0.109)* and *lagoon (0.108)*. The outputs of the Convolutional Neural Network (CNN) are in decreasing order of confidence. Besides the scene class, there are given some attributes that define this specific scene, such as: *natural light, open area, far away horizon, sunny, natural, warm, boating, dirt, clouds* and the environment type – *outdoor*.

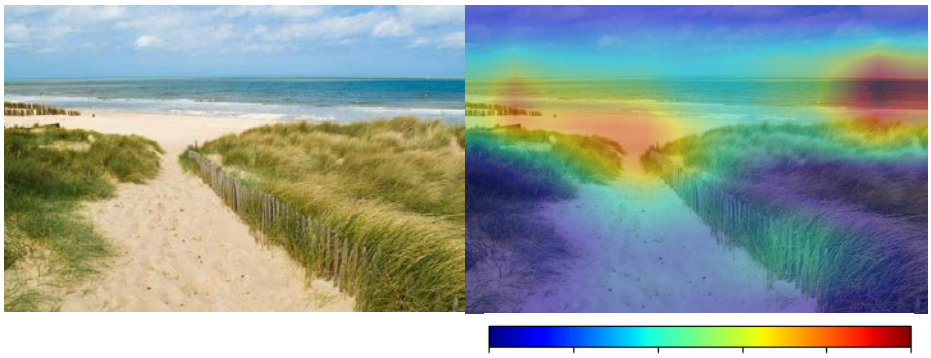

Fig. 1 – Input image (left); informative region for prediction (right).

Fig. 1 – left represents the information of regions for the category prediction with highest confidence - "beach", also an output of the CNN (Zhou B. *et al*., 2017). The colors vary from red (hot) – strong confidence for the area to be “beach” to blue (could) – low confidence for the area to be a “beach”.

Another example of scene understanding is given by the researchers of Facebook (Aarthi *et al*., 2017). Their solution, based on Visual Relationship

Dataset (VRD), is a CNN that learns and predicts relationships between the detected objects in an image. An output example of the network is shown in Fig. 2. The model is able to recognize relationships between 53.000 object categories and summarizes 29.000 relation categories.

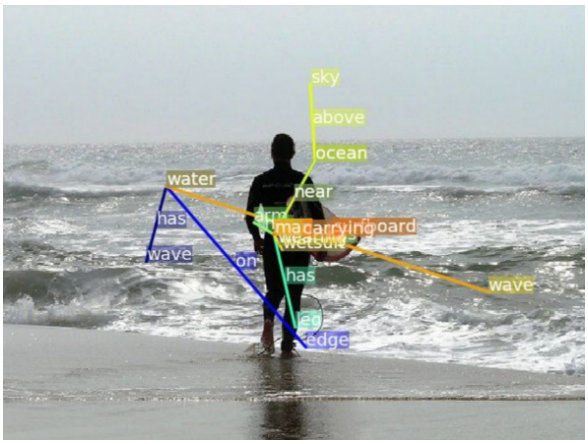

Fig. 2 – Example of VRD prediction.

Even with the research accomplishments in computer vision in the past few years, scene understanding is still a problem to be solved. This turns out to be even more difficult because the acquired data is noisy, sometimes incomplete and the range of object characteristics and elements within the scene varies a lot when the environment is changed. A number of works have shown that this problem can be overcome by using 3Dreconstruction and object classification. Both tasks use algorithms that require great computational power making them unusable on real-time systems with limited resources (low-powered and memory-limited). Nevertheless, real-time perception of scenes is required to work on mobile platforms such as smart phones and embedded platforms in order to implement applications like augmented reality, autonomous driving or intelligent robots (Wald *et al*., 2018).

The main challenge of scene understanding is to detect useful features and characteristics for some objects of importance. Taking into consideration their actions and behaviors, all the relationships should be explored methodically. A golden rule to evaluate performance is to compare the results of the model with the ground truth. Some important factors that affect scene understanding to be taken into consideration are:

- motion
- changes in illuminance
- occlusions of objects
- camera noise

This paper focuses on two directions of scene understanding: based on low-level features and based on object detection followed by visual relationship detection between them. The state-of-the-art for both directions will be presented along with their accomplishments and limitations.

2. Scene Classification, Recognition and Understanding

First, we have to define the concept of "scene" in a technical context. One author describes a scene as a place in which a person can act and navigate (Xiao *et al*., 2010; Singh *et al*., 2017). Accordingly, the concepts of scene recognition, classification and understanding are related to the semantic understanding of the scene.

Aude Oliva defines a scene as a real-world environment view containing multiple objects and surfaces that are organized in a certain way (Oliva, 2008). A distinction between scenes and objects must be made from the very beginning: "objects are compact and act upon, while scenes are extended in space and act within" (Oliva, 2008). Fig. 3 exemplifies a wide range of images with scenes and objects.

The problem of scene understanding has been studied and many researchers documented various methods. In general, the proposed solutions start from the concept of neural networks, making the system able to learn similar to humans. A number of papers have shown that recognizing a scene involves understanding the visuals (Ali *et al*., 2017), object detection (Zhou X. *et al*., 2017) and estimating geometric features. A first scene classification divides them into *object-centered* and *scene-centered* (Pawar and Devendran, 2019). An important result in feature extraction reached 70% accuracy on *Sun397*dataset, it uses Places-CNNs and ImageNet-CNNs for feature extraction (Herranz *et al*., 2016).

A proposed bottom-up architecture estimates the room layout by using semantic segmentation and optimization of hypotheses. It implements RoomNet focusing on low level features, then produces a hypothesis by semantic segmentation (Lee *et al*., 2017).

Dahua Lin and Jianxiog Xio documented another model (Lin and Xio, 2013) which uses the geometrical pixel arrangement for semantic interpretation and segmentation (Pawar and Devendran, 2019). The model creates structural layers for outdoor scenes with notable experimental results for semantic segmentation and scene classification.

One way to overcome the problem of scene understanding is situation recognition. Using a structural prediction model one research (Yatskar *et al*., 2016) achieved activity and object recognition, resulting in an overview of situation related to subjects, objects, activities and location. Once again we look up to the human understanding when analyzing an image and try to implement the same behavior and knowledge extractor on computer vision based systems.

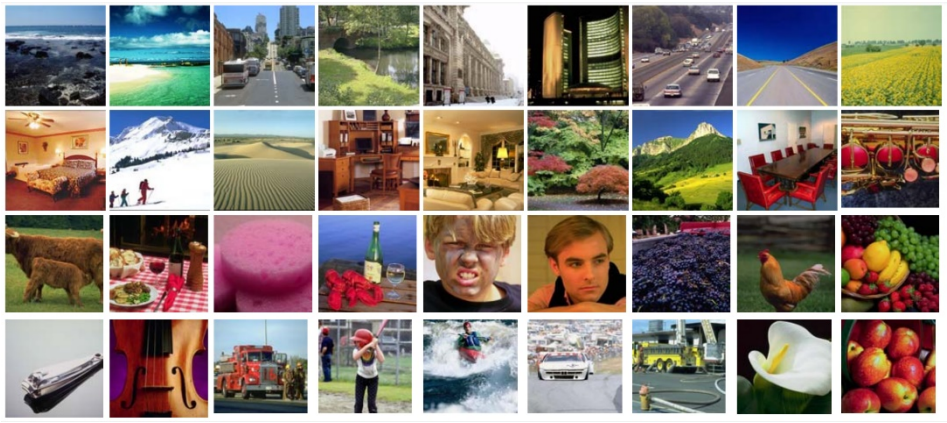

Fig. 3 – Natural images containing objects and scenes (Oliva, 2008).

One approach to solve this problem involves the use of a fully connected layer(FCN) that predicts the geometric context using mid-level features like informative edges (Mallya and Lazebnik, 2015). This can be a hint to identify some spatial attributes specific to a scene. The method uses very large datasets, like LSUN which contains millions of images, objects, subjects and scenes (Pawar and Devendran, 2019). 25 years back, in 1995, image understanding was defined as a verbal description of the image (Ralescu *et al*., 1995).

There are different definitions of image understanding (IU) depending on the objective. An opinion broadly agreed describes IU as content, objects, subjects, relations between them and events (Singh *et al*., 2017).With very little or no effort humans can quickly analyze a scene. Furthermore, are able to categorize scenes from different natural environments (Li *et al*., 2003; Thorpe *et al*., 1996).

In front of a picture, a human needs about 100ms of exposure time in order to process and identify the scene. This performance is still subject of intense research, an understanding of the human vision system was documented 40 years ago by Marr (Marr, 1982; Singh *et al*., 2017).

In general, this problem can be tackled in two different ways. Depending on what is the starting point, the scene understanding algorithms can be split into two categories (Singh *et al*., 2017):

- based on object detection,
- using low-level features.

3. Scene Understanding Based on Object Recognition

To overcome the problem of scene understanding, some approaches use object detection and recognition as a starting point. This method is able to analyze

complex scene which can be difficult using low-level image features implementation.

Some authors consider that for high-level visual tasks the low-level image features algorithms will not work well enough (Li *et al*., 2010). A possible solution to the problem at hand is the Object Bank, an image representation which integrates the result of numerous object detectors. The researchers declare that on one hand using object detectors which are impartial to the testing dataset or the visual task and on the other hand implementing regularized logistic regression, the model will achieve better performance. For pre-training the object detectors, this method makes use of different results, like those of Felzenszwalb *et al.* (2010) and Hoeim *et al.* (2005).

In Fig. 4, the first row represents the weights of the Object Bank dimensions; in the middle row there is the heat map of the highest weight. The last ones are the scene images with the mask of the most relevant object dimensions.

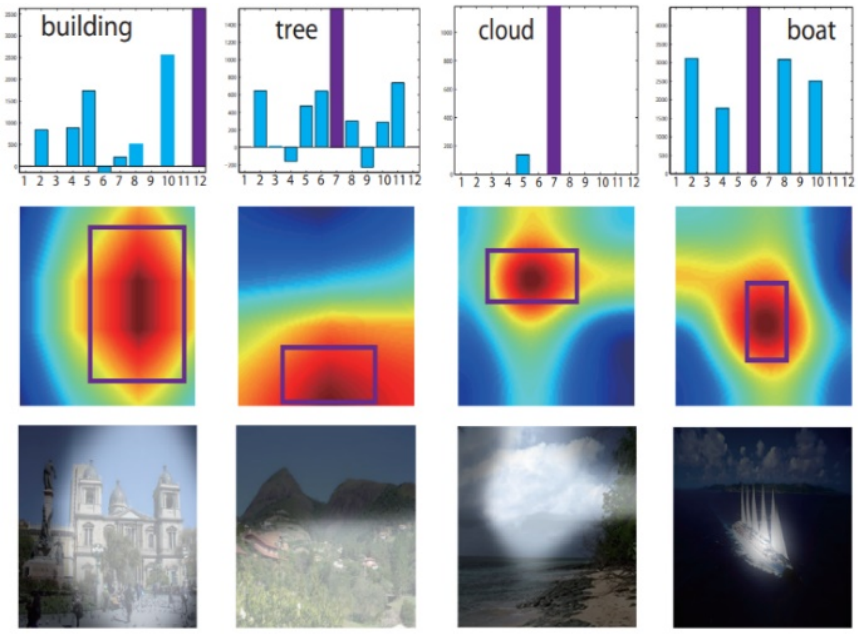

Fig. 4 – Illustration of the learned coefficient within an object group (Li *et al*., 2010).

Fig. 5 represents the object-wise coefficient having the scene class, for scene "beach" the objects with a very high weight are "sky", "water", "sand", "grass", "cloud" and "ocean".

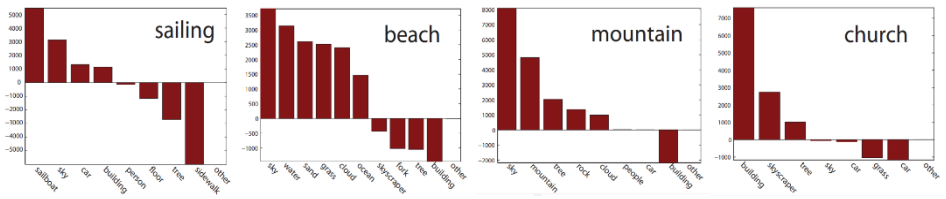

Fig. 5 – Object-wise coefficients (Li *et al.*, 2010).

4. Scene Understanding Using Low-Level Features

Some methods focus on overcoming the problems of scene understanding by proposing different plans using low-level image features. This comes as an alternative to recognize a scene without the need to detect and recognize the objects within it first. One of many reasons for avoiding object recognition as a first step in scene understanding is that the errors from this part propagate further to the scene recognition algorithm. The object detection segment is treated as a black box and still has problems because of illumination conditions or occlusions (Singh *et al.*, 2017). The researchers are looking for methods to find low-level features in images that are relevant to the semantic context, such as: texture, orientation, edges and color.

Based on texture analysis, Renninger and Malik describe a method which rapidly provides clues for identifying a scene after limited exposure (Renninger and Malik, 2004). Julesz defines a core concept of text on (Julesz, 1986; Julesz, 1981), as the image element that lead to our perception of texture. The model based on text determine local features of textures that correspond with a certain probability to a scene class. The histogram of features within an image is then compared to the database resulted from training examples. This is an early-stage algorithm for scene classification and identification, based on a texture recognition model.

Other approaches have had the objective of understanding human perception several years ago; in 1994 Gorkani and Picard published a paper documenting how to quantify the "dominant perceived orientation" (Gorkani and Picard, 1994). Researchers conclude that orientation is a very important feature for classifying textures. The dataset they used contains vacation photos in which it was possible to distinguish between a city or a suburb. The results show that approximately 92.93% of scenes were classified the same humans did.

The research done by Guerin-Dugueetal (Guerin-Dugue and Oliva, 2000) uses a similar approach. The classification task uses a scaling factor and is based on the local dominant orientation feature to decide between four categories: outdoor urban scenes, indoor, closed landscapes, open landscapes.

Another solution to scene understanding uses key points for visual categorization is proposed in (Csurka *et al*., 2004). The algorithm is implemented as follows: first the image patches are detected and described, after which SIFT descriptors are used to create a vocabulary of image descriptors, then a bag of key points is formed. Lastly, a support vector machine (SVM) classifier decides what the image category is.

After analyzing some methods, we can observe that a counting is done in a form or another. This pushed some authors to implement probabilistic based models to understand the scene using feature extraction. One direction is given by Fei-Fei and Perona (Li and Pietro, 2005), they use low-level texture features as descriptors. In his algorithm initially the regions are accumulated into intermediate themes, after which the classification is done. In contrast to other models for scene understanding, this approach generates a collection of themes that could be correlated with the image.

Fig. 6 – Test images for “office” and “living room” categories (Li and Pietro, 2005).

Fig. 6 exemplifies a few testing images for classes office and living room. The first three columns on the left are correctly recognized images, in the last column fewer significant code words are detected and the image is not classified correctly.

An alternative probabilistic approach to detect contents within a scene is documented by Singhal *et al*. (2003). The authors document a method based on material detection and a probabilistic model. Material detection is implemented to solve the problem of key semantic objects for the scene, for example: water, grass, snow, sky, etc. The algorithm analyses low-level features and feeds the output to a classifier in order to get the material.

Another method focuses on understanding scenes based on a spatial pyramid (Lazebnik *et al*., 2006). The authors state that this algorithm improves the bag of features model by using a geometric correspondence. The implementation frequently divides the image and calculates its histogram of local features. It uses the idea of combining multiple resolutions in such a way that the

best results can be obtained. The list of resulted features is one of the following: oriented edge points and SIFT-descriptors, after which the classification is achieved using a SVM.

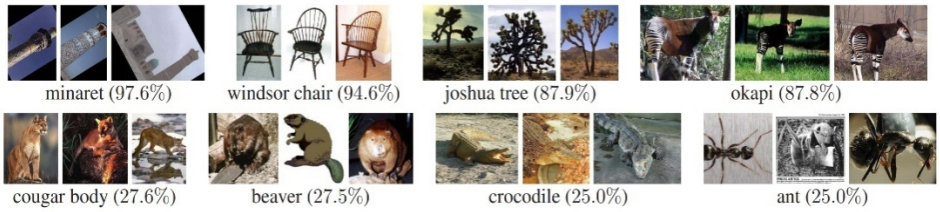

Fig. 7 – Results of (Lazebnik *et al*., 2006) on Caltech-101 dataset.

Fig. 7 illustrates the results of Svetlana Lazebnik *et al.* (2006) on Caltech-101 dataset. On the top row are represented some classes on which their method achieves high performance and on the bottom row are some classes on which the proposed algorithm performs poorly.

One scenario very little considered by researchers is using color based features for distinguishing between scenes. By using color descriptors the illumination invariance increases; Van De Sande *et al*. (2010) used this method for object and scene recognition. The authors state that classification and category recognition are affected by variance of light intensity changes and light color changes and thus an implementation taking into account color-based features is a necessary update. The comparison between histogram based descriptors, moment based descriptors and color SIFT descriptors revealed that the last ones are the best choice.

For distinguishing between indoor and outdoor scenes a solution is presented by Szummer and Picard (1998), they use low-level image features for high-level scene properties and their classification. The features they extract and use are: coefficient of shift discrete cosine transform (DCT), ohta space histogram and autoregressive model parameters (Singh *et al*., 2017).

One classification for indoor and outdoor scenes is documented by Serrano *et al.* (2004). The scientist use low-level image features and a Bayesian network for the final result. The paper stands out by using the wavelet texture features as a replacement for MSAR texture features in order to reduce the computational time.

5. Scene Understanding Using Other Approaches

Vogel and Schiele (2004) introduce for the first time the concept of *semantic typicality* used to categorize natural scenes. This measurement is used to classify an image. The researchers define the *typicality* for the uncertainty of annotation judgment. In cognitive research notions like typicality and prototype

made an impact after the research began by Eleanor Rosch (Rosch, 1973; Rosch *et al*., 1975; Rosch *et al*., 1976). One problem arises when the category annotation is influenced by the perspective of a person, and so it is absolutely necessary to model the typicality of a scene after manually annotating it. In this example six scenes are considered: coasts, forests, mountains, rivers/lakes, plains and sky/clouds; similarities are measured with respect to these categories.

Fig. 8 – Examples of correctly (top) and incorrectly categorized images (Vogel and Schiele, 2004).

In Fig. 8 it is presented the output of Julia Vogel's method: best and worst categorized images (Vogel and Schiele, 2004).

Another approach is proposed by Lipson *et al*. (1997); they use the configural recognition for encoding scene class structure as a model of important image regions and the relations between them. For instance, if there are defined three regions: B – blue region, W - white region and G - gray region; a mountain covered by snow always has region B above region W that is above region G. One class model is defined by seven types of relationships, each taking the value of: less than, equal to or greater than. The encoded relations are related to color, illuminance, spatial relationships and size of the patch. After that, the region is classified into *above* or *below*. This model acts as an adjustable template so that it can be set to best match the image regarding the photometric attributes and luminance.

For scene recognition another concept is introduced by Pandey and Lazebnik (2011). *Deformable part-based models* (DPM) catch persistent visual elements and pertinent objects. The image is defined as the change in histogram of oriented gradients features, which are then used to classify scenes applying SVM.

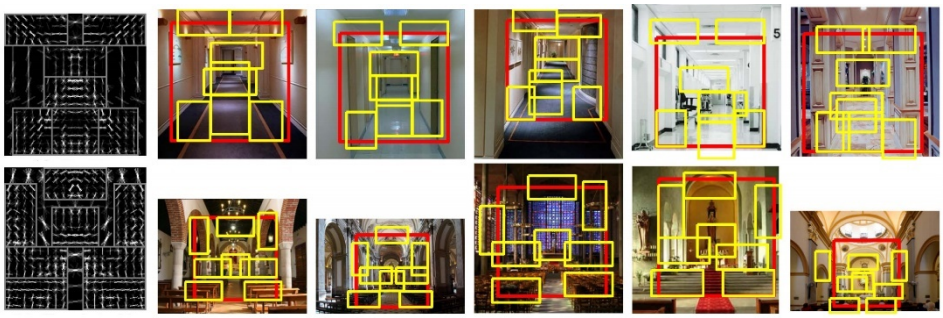

Fig. 9 – Scene models for “corridor” (top left) and “church inside” (bottom left) and test images with the root filter (red) and part filters (yellow) (Pandey and Lazebnik, 2011).

An image classification at scene level is also proposed by Yang *et al*. (2016). The researchers document an algorithm based on covariance descriptors as a matrix of certain features: spatial location, color and the gradient. The algorithm is structured in three parts: extract covariance descriptors, perform collaborative coding using dictionary coefficients and classify, and generate the vector label. The results are satisfying on high resolution images.

A different approach for the scene recognition problem is presented by Olivia and Torralba (2001); they propose a model that analyzes a scene as a single object with a specific shape not a composition of objects. The algorithm determines the spectral signature of the scene categories from labeled data, after which a regression is used to find links between global and spectral features.

Based on the fact that humans first perceive the bigger picture and then use that information to extract details, some algorithms detect the scene first and based on the result keep searching for more objects or different structures in the image (Singh *et al*., 2017). Such an example is described by Murphy *et al*. (2003); based on a natural approach, the algorithm detects the scene and then the presence of an object. The big difference here is that the researchers use all the image as an overall feature to avoid the uncertainties that might appear at a smaller level.

The solution described by Yao *et al*. defines comprehension of the whole image as a holistic scene understanding (Yao *et al*., 2012). The output is an aggregated conclusion which connects different aspects: class label and bounding box, regions, location and scene type. Authors chose to implement two levels of segmentation: *segments* and *super segments*. As we would have expected, compared with the first layer of segmentation super segments are computationally more efficient; they are used for dependencies in a longer range.

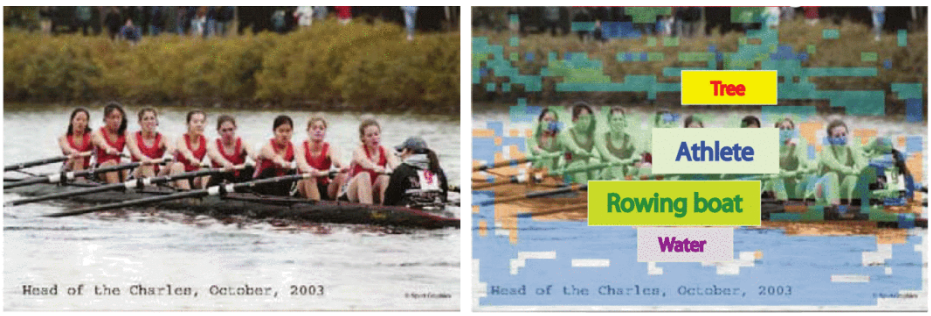

Fig. 10 – Example of WHAT, WHERE and WHO story: *rowing* on a *lake*, objects - *threes*, *athletes*, *rowing boat* and *water* (Li and Fei-Fei, 2007).

A step further is taken by Li *et al*., they describe an event recognizing algorithm (Li and Fei-Fei, 2007).The researchers define the event as a human activity taking place in a specific environment. The desired results are: object detection, scene recognition and event classification.

Given the example image from Fig. 10 left, the system responds to the *what?* question: the event of rowing; *where?* does it happen: on a lake and *who* or *what* objects are in the image: threes, athletes, rowing boat and water.

6. Conclusions

When choosing a scene understanding algorithm there are two main criterions to take into consideration: the level of scene understanding needed for the specific computer vision application and the available hardware resources. For instance a simple “indoor”/”outdoor” classification task can easily be done using a feature extraction algorithm, while for a more complex classification of scenes into specific places (living room/ bedroom/ terrace/ backyard/ garage/ hallway/ etc) a machine learning approach such as a neural network for image classification will provide better results. For high level scene understanding such as interactions between objects and people and context/activity recognition a more complex solution is needed. An algorithm for that type of task would be object detection and recognition followed by activity and human-object interaction recognition.

Scene recognition and classification offers an overview and a general description, producing some of the following outputs: scene class (*beach, coast, beach house, lagoon, etc*), environment type (*indoor/outdoor*), scene attributes (*natural light, open area, far-away horizon, sunny, natural, etc*) and the interest regions for *Top-1* predicted category.

We propose to examine and test alternative neural networks to handle an initial scene recognition, classification and a general description, but also take a

step further and integrate methods for detecting objects and the relationships between them in order to give a complete visual scene understanding. We'll be looking to build a neural network that can be trained to improve both scene recognition and <*subject, predicate, object*> detection.

REFERENCES

Aarthi S., Chitrakala S., *Scene Understanding - A Survey*, in IEEE International Conference on Computer, Communication, and Signal Processing (ICCCSP-2017), 2017.

Ali A.R., Shahid U., Ali M., *High-Level Concepts for Affective Understanding of Images*, IEEE Winter Conference on Applications of Computer Vision, 2017, 1-9.

Csurka G., Dance C.R., Fan L., Willamowski J., Bray C., *Visual Categorization with Bags of Key Points*, Workshop on Statistical Learning in Computer Vision, ECCV, Vol. **1**, 2004.

Felzenszwalb P.F., Girshick R.B., McAllester D., Ramanan D., *Object Detection with Discriminatively Trained Part-Based Models*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 1627-1645.

Gorkani M.M., Picard R.W., *Texture Orientation for Sorting Photos "at a Glance"*, Proceedings of 12th International Conference on Pattern Recognition Vol. **1**, 1994, 459-464.

Guerin-Dugue A., Oliva A., *Classification of Scene Photographs from Local Orientations Features*, Pattern Recognition Letters, 2000, 1135-1140.

Herranz L., Jiang S., Li X., *Scene Recognition with CNNs: Objects, Scales and Dataset Bias*, International Conference on Computer Vision and Pattern Recognition (CVPR16), 2016.

Hoiem D., Efros A.A., Hebert M., *Automatic Photo Pop-Up*, ACM Transactions on Graphics (TOG), 2005, 577-584.

Julesz B., *Texton Gradients: The Texton Theory Revisited*, Biological Cybernetics, 1986, 245-251.

Julesz B., *Textons, the Elements of Texture Perception and Their Interactions*, Nature, 1981, 91-97.

Lazebnik S., Schmid C., Ponce J., *Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories*, in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2169-2178, 2006.

Lee C.-Y., Badrinarayan V., Malisiewicz T., Rabinovich A., *Roomnet: End-to-End Room Layout Estimation*, IEEE International Conference on Computer Vision, 2017, 4875-4884.

Li F.F., VanRullen R., Koch C., Perona P., *Natural Scene Categorization in the Near Absence of Attention: Further Explorations*, Journal of Vision, 2003, 331-331.

Li L.-J., Su H., Xing E. P., Fei-Fei L., *Object bank: A High-Level Image Representation for Scene Classification and Semantic Feature Sparsification*, Advances in Neural Information Processing Systems, 2010, 1378-1386.

Li L., Fei-Fei Li, *What, where and who? Classifying Events by Scene and Object Recognition*, in IEEE 11th International Conference on Computer Vision, pp. 1-8, 2007.

Li F.-F., Pietro P., *A Bayesian Hierarchical Model for Learning Natural Scene Categories*, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Vol. **2**, pp. 524-531, 2005.

Lin D., Xio J., *Characterizing Layout of Outdoor Scenes Using Spatial Topic Processes*, IEEE International Conference on Computer Vision, 2013, 1-8.

Lipson P., Grimson E., Sinha P., *Configuration Based Scene Classification and Image Indexing*, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1007-1013, 1997.

Mallya A., Lazebnik S., *Learning Informative Edge Maps for Indoor Scene Layout Prediction*, IEEE International Conference on Computer Vision, 936-944, 2015.

Marr D., *Vision: A Computational Approach*, Henry Holt and Co., 1982.

Murphy K., Torralba A., Freeman W.T., *Using the Forest to See the Trees: A Graphical Model Relating Features*, Objects And Scenes, in Advances in Neural Information Processing Systems, pp. 1499-1506, 2003.

Oliva A., Torralba A., *Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope*, in International Journal of Computer Vision, pp. 145-175, 2001.

Oliva A., *Object Recognition and Scene Understanding Course*, http://people.csail.mit.edu/torralba/courses/6.870/slides/lecture5.pdf, 2008.

Pandey M., Lazebnik S., *Scene Recognition and Weakly Supervised Object Localization with Deformable Part Based Models*, in IEEE International Conference on Computer Vision, pp. 1307-1314, 2011.

Pawar P.G., Devendran D.V., *Scene Understanding: A Survey to See the World at a Single Glance*, 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT), 2019, 182-186.

Ralescu *et al*., *Image Understanding = Verbal Description of the Image Contents, in SOFT*, Journal of the Japanese Society for Fuzzy Theory, pp. 739-746, 1995.

Renninger L.W., Malik J., *When is Scene Identification Just Texture Recognition?*, Vision Research, 2004, 2301-2311.

Rosch E.H., *Natural Categories*, in Cognitive Psychology, pp. 328-350, 1973.

Rosch E., Mervis C.B., *Family Resemblances: Studies in the Internal Structure of Categories*, in Cognitive Psychology, pp. 573-605, 1975.

Rosch E., Mervis C.B., Gray W.D., Johnson D.M., Boyes-Braem P., *Basic Objects in Natural Categories*, in Cognitive Psychology, pp. 382-439, 1976.

Serrano N., Savakis A., Luo J., *Improved Scene Classification Using Efficient Low-Level Features and Semantic Cues*, in Pattern Recognition, pp. 1773-1784, 2004.

Singh V., Girish D., Ralescu A., *Image Understanding - A Brief Review of Scene Classification and Recognition*, MAICS 2017: The 28th Modern Artificial Intelligence and Cognitive Science Conference, 2017, 85-91.

Singhal A., Jiebo Luo, Weiyu Zhu, *Probabilistic Spatial Context Models for Scene Content Understanding*, in 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, 2003.

Szummer M., Picard R.W., *Indoor-Outdoor Image Classification*, in 1998 IEEE International Workshop on Content-Based Access of Image and Video Database, pp. 45-51, 1998.

Thorpe S.J., Fize D., Marlot C., *Speed of Processing in the Human Visual System*, Nature 381, 1996, 520.

van de Sande K., Gevers T., Snoek C., *Evaluating Color Descriptors for Object and Scene Recognition*, in IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1582-1596, 2010.

Vogel J., Schiele B., *A Semantic Typicality Measure for Natural Scene Categorization*, in Joint Pattern Recognition Symposium, pp. 195-203, 2004.

Wald J., Tateno K., Sturm J., Navab N., Tombari F., *Real-Time Fully Incremental Scene Understanding on Mobile Platforms*, IEEE Robotics Andautomation Letters, Vol. **3**, 2018, 3402-3409.

Xiao J., Hays J., Ehinger K.A., Oliva A., Torralba A., *Sun Database: Large Scale Scene Recognition from Abbey to Zoo*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, 3485-3492.

Yang C., Liu H., Wang S., Liao S., *Scene-Level Geographic Image Classification Based on a Covariance Descriptor Using Supervised Collaborative Kernel Coding*, in Sensors, pp. 392, 2016.

Yao J., Fidler S., Urtasun R., *Describing the Scene as a Whole: Joint Object Detection*, Scene Classification and Semantic Segmentation, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 702-709, 2012.

Yatskar M., Zettlemoyer L., Farhadi A., *Situation Recognition: Visual Semantic Role Labeling for Image Understanding*, IEEE Conference on Computer Vision and Pattern Recognition, 2016, 1-9.

Zhou X., Gong W., Fu W., Du F., *Application of Deep Learning in Object Detection*, IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), 2017, 1-4.

Zhou B., Lapedriza A., Khosla A., Oliva A., Torralba A., *Places: A 10 Million Image Database for Scene Recognition*, IEEE, 2017.

SCURTĂ RECENZIE A LITERATURII DE SPECIALITATE PENTRU ÎNȚELEGEREA VIZUALĂ A SCENEI

(Rezumat)

Persoanele sunt extrem de precise în ceea ce privește înțelegerea vizuală a scenelor naturale. Prin extragerea și extrapolarea datelor ajungem la cel mai înalt stadiu de înțelegere a scenei. În ultimii ani perceperea de nivel înalt a scenei s-a dovedit a fi o parte esențială în aplicațiile de computer vision. Soluția de înțelegere vizuală a scenei merge mai departe, adăugând un nivel de abstractizare suplimentar peste detecția și recunoașterea obiectelor. În acest fel percepția sistemelor se apropie de cea umană: integrează informații semnificative și extrage relații și modele semantice. În domeniul *Computer Vision* cercetarea s-a concentrat pe algoritmi de înțelegere a scenei, scopul fiind obținerea de cunoștințe semantice din mediu și determinarea proprietăților obiectelor și a relațiilor dintre acestea. Câteva domenii care aplică soluții de înțelegere vizuală a scenei sunt: robotica, industria jocurilor, assisted living, realitatea augmentată, etc. Un task fundamental pentru astfel de aplicații este conștientizarea poziției spațiale și capturarea informațiilor de adâncime.

Prima partea acestei lucrări abordează soluții bazate pe detecția și recunoașterea obiectelor, iar în a doua parte sunt prezentate propuneri care pornesc de la caracteristicile de nivel scăzut ale obiectelor din imagine.