Principles of decision-making systems construction based on semantic image analysis

Natallia Iskra
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
Email: niskra@bsuir.by

Vitali Iskra
Omnigon Communications LLC
New York, NY, USA
Email: iskra.vitaly@gmail.com

Marina Lukashevich
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
Email: lukashevich@bsuir.by

***Abstract*—In this paper principles of decision-making systems construction are considered. An approach to image analysis based on semantic model is proposed and studied. The results show an improvement in processing speed and image captioning quality based on Visual Genome dataset.**

***Keywords*—decision-making, video surveillance, neural networks, semantic analysis, image captioning**

I. Introduction

The task of image analysis in decision-making systems using technical vision today is acute. Automatic image interpretation is a non-trivial task. For example, for a video surveillance system, it would be relevant not only to record and save video, but also to analyze what is happening, as well as to signal any suspicious situations - violations, incidents, actions that require an immediate response.

The approach to image analysis considered in this paper proceeds as follows:

Step 1. Individual objects detection. These can be the objects that are significant in the context of the system (for example, traffic participants, road markings and signs in traffic monitoring systems), areas that outline objects (bounding boxes), or more precise object-by-pixel selection.

Step 2. Building a semantic model. At this stage, relations between objects and / or attributes of individual objects are formalized.

Step 3. Model interpretation. According to the constructed model, a textual description of what is happening (an annotation of the image or image caption, for example, for keeping a surveillance log) can be obtained, or specific situations on the image that are of interest (for example, cases of traffic rules violation, traffic accidents, etc.) can be determined. In this case, the interpretation of the model will consist in highlighting only those relationships and attributes that can signal of an abnormal situation.

The most important part in a situational analysis implementing is the construction of an interpretable image model. Modern approaches to models construction have a large number of limitations. In this article, the main focus is on the methodology for constructing this model, the selection of an algorithm for detecting objects in an image, as a preliminary stage of building a model, as well as the principles of quality analysis of the constructed model and decision-making based on it. To represent the obtained model and implement the decision-making process on the basis of the obtained model, it is proposed to use the approaches developed in the framework of OSTIS Technology.

II. Methods overview

A. Object detection in images

The first step during the image analysis is to process the source image and detect the objects automatically. During this step one of the following tasks is performed as a rule [1]:

- Semantic Segmentation – for every pixel in the source image determine its class and category;
- Classification and Localization – determine the class of a single object in the image and its exact location;
- Object Detection – define a class and a rectangular area bounding each of the objects in the image;
- Instance Segmentation – on the image with multiple objects determine the contours (all visible pixels) and the class of each of the objects.

From the standpoint of a semantic model construction last two tasks are of the most interest.

Among the existing modern object detection algorithms, including those based on deep learning methods, the most relevant approaches are:

- Sliding Window [2];
- Region Proposals [3];
- Single Shot Detection [4].

Each of the approaches has its own advantages and disadvantages, in terms of their relevance to application in systems that include image analysis [5].

To construct the model, described in this paper, it seems to be the most promising to use methods, based on the group of neural networks architectures with region proposals, so-called R-CNN, and their development:

- R-CNN [3] – represents a model of sequential image processing pipeline: generation of a set of regional proposals, the use of a pre-trained convolutional neural network with a final layer of support vectors and linear regression for a more accurate regions estimation;
- Fast R-CNN [6] – a model in which, to speed up the performance of the previous processing pipeline, a selection of regions and the union of all neural network models into one are used;
- Faster R-CNN [7] – to accelerate the model even further, a selective search of regions is used;
- Mask R-CNN [8] – unlike previous models, this one uses a binary mask to determine not just a rectangular region - a candidate for objects, but specific pixels belonging to the object, which, in essence, is the solution to the image segmentation problem described above.

B. Semantic model

Within the framework of modern approaches as the basis of the semantic image model the so-called Scene Graph [9] is widely used. A scene graph is a data structure that describes the contents of a scene, which, in turn, can be specified by an image or its textual description. In the scene graph instances of objects their attributes and relationships between objects are encoded.

Formally, a scene graph is defined as follows: let C be a set of object classes, A – a set of attribute types, R – a set of relation types. A scene graph is defined as $G = (O, E)$, where $O = \{o_1, \ldots, o_n\}$ – a set of objects – nodes of a graph, $E \in O \times R \times O$ – a set of graph edges. Every object is represented by $o_i = \{c_i, A_i\}$, where $c_i \in C$ – the class of the object, and $A_i \in A$ – its attributes.

A scene graph can be grounded to an image. Let B be a set of rectangular areas, each of which delineates a certain object in the image (they are generally called Bounding Boxes), then the grounding of the scene graph *G = (O, E)* to the image is the function $\gamma : O \to B$, or γ_o.

To conduct the experiments the dataset Visual Genome [10] is commonly used. It consists of 108 077 labelled images, for which 5.4 millions of the textual descriptions and scene-graphs for the whole images and their sections (regions) were produced using crowd sourcing. All the scene graphs are grounded to either textual descriptions, or images (regions), or both.

The example of scene graph – region grounding in Visual Gnome is shown in Fig. 1.

Grounding scene graphs to a textual descriptions (each object, attribute and relation) in Visual Genome corresponds to WordNet synset [12]. WordNet – network word representation,that is structured according to semantic relations. In WordNet each word is represented as a set of its synonymous meanings, which are called synsets. Each synset comprises of a triplet <word>.<pos>.<number>, where word – is a word itself, pos – its part of speech (n – noun, v – verb, a – adjective, r – adverb), number – index of the meaning in the set. E.g. the term "person" in WordNet is represented by three meanings person.n.01, person.n.02 and person.n.03. Textual grounding of the object "person" in Visual Genome corresponds to the synset person.n.01. In WordNet there are relations of synonymy, antonymy, "part – whole" (meronym – holonym), "general – specific" (hypernym – holonym).

Using a graph representation to describe an image model has a number of significant advantages compared to more traditional approaches to image captioning aimed toward a natural language description (considered in [13] and other works). The graph representation is more unambiguous (invariant) and is much better suited for automatic processing, and, in particular, the interpretation of such a model.

However, despite these advantages, the currently used approach to the scene graph construction has a number of disadvantages that make it difficult to interpret image models presented in this form. The key disadvantage, in our opinion, is the lack of any semantic unification (standardization) in the principles of building scene graphs, in particular, in the principles of distinguishing relations and attributes (and, generally speaking, in this case, there is no clear boundary between the concepts of relation and attribute), in the framework of even one data set, as well as the lack of syntactic unification in the representation of scene graphs in various approaches. In addition, in modern approaches to the construction of scene graphs, as a rule, the problem of internationalization still remains.

In turn, the lack of unification in the representation of scene graphs makes it impossible to build universal processing tools for such graphs, in particular, means for verifying and decision making based on scene graphs.

An ontological approach is currently used as the basis for solving the problem of unification in various fields. In this paper, to implement this approach, it is proposed to use the OSTIS Technology, within the framework of which a unified standard for coding information (SC-code) is proposed [14], the problem of unification of the principles for representing different types of knowledge [15] and the problem of integrating various models for problem solving [16] are solved .

We list some of the advantages of OSTIS Technology that are relevant in the context of solving the problem posed in this paper:

- unification of the representation of semantic image models;
- ensuring the independence of the image model from the external language in which the preliminary description was made;
- the possibility of structuring the model according to various criteria, as well as the possibility of representing meta-information, which, in particular,

Figure 1. An example of an image from Visual Genome with grounding [11].

will allow us to select image regions that are more or less significant for solving the current problem;

- the availability of verification tools and adjustments to the image model, partially considered in [17], which make it possible to verify the model itself for its internal consistency and adequacy of the subject area, and in the future will automatically evaluate the degree of conformity of the automatically obtained model to standard models developed by experts;
- the presence of a large amount of source data that has been accumulated at the moment and which may be useful for further research makes the task of automating the transition from scene graphs, for example, from the Visual Genome dataset, to a unified semantic representation in relevant SC-code.

III. Semantic model construction technique

To build a semantic image model in the form of a scene graph, we must first detect the objects in the image, and then for each pair of objects decide whether they have a relations and which ones [18]. The selection of relations can be greatly simplified by using external knowledge bases (general purpose or specialized for a specific domain) [17]. In both cases, for the image on which n objects are found, it is necessary to consider $(n^2 - n)$ relations. In this paper it is proposed to simplify the solution by identifying the so-called salient (significant or the most relevant) objects [19], and to further consider $(n - 1)$ relationships. This approach corresponds to the scenario of tracking certain objects or situations in surveillance systems.

Frequency analysis of Visual Genome data shows that the most frequent relationships between objects in images are spatial relationships: the "on" relationship occurs 642,150 times, the "in" relation – 218,166, "behind" - 39,833. In addition, due to the hierarchical structure of WordNet grounding, spatial relationships can be described in more detail: for example, "car in a parking lot" or "car is parked in a parking lot". Indeed, when looking at an image, a person first of all notes how the objects are located relative to each other in space. In automatic processing it is also possible to determine semantic spatial relations between objects [20]. In addition, reducing the set of relations of the image model to spatial relations will allow at the current stage to significantly simplify the process of constructing and interpreting the model, while maintaining the ability to assess the anomalies and oddities.

The technique for automatic model construction for spatial relations is presented below.

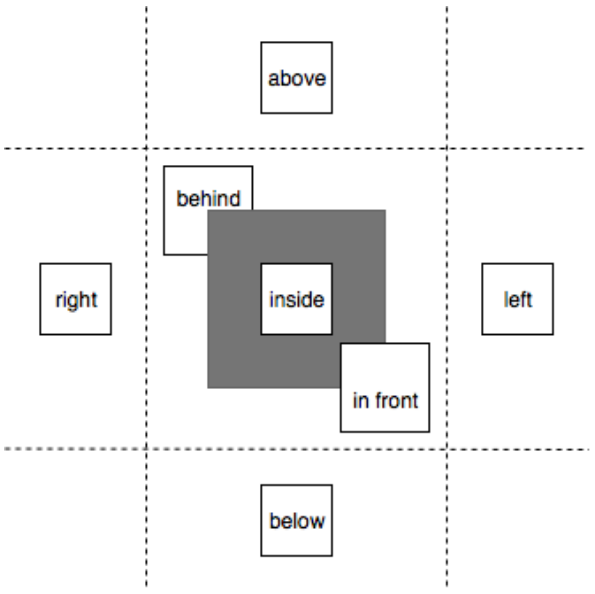

Figure 2. Spatial relations system.

In Fig. 2 the system of all possible spatial relations is visualized: the area of the salient object (subject) is filled, all the other areas are the options for the location of the object of interest (object), for which, using the

decision tree in Fig. 3, the type of spatial relationship in the form “subject-relation-object” will be determined.

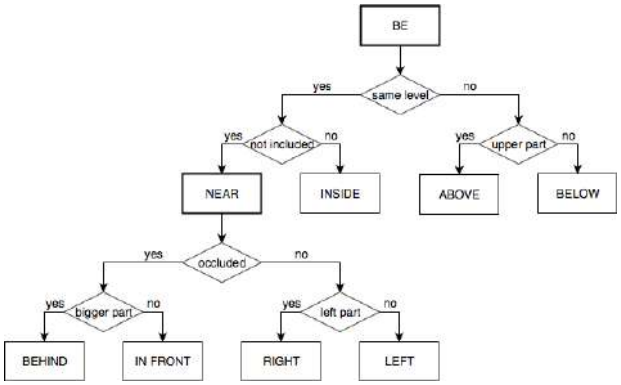

Figure 3. The decision tree for automatic spatial relations model construction.

It should be noted that for the names of types of relations in this model prepositions are used (and prepositions, as it was described above, are not represented in WordNet), i.e. at this stage, grounding to WordNet is not possible, but at the next step (for interpretation), synsets containing these prepositions and their meanings (be.v.01, along.r.01, etc.) will be used.

In the decision tree the rectangles show the blocks corresponding to a certain type of relationship, while more general relationships that need to be specified are highlighted (similar to hypernyms from WordNet). When constructing a tree to speed up the process of final decision-making, the rules (shown in the figure by rhombus’s) were formulated in accordance with the statistical data of the Visual Genome analysis, so that a more frequent case would be to the left of the tree. So, in the initial dataset, the“near” relationship is found more often than other spatial relationships (26,316 times), the “above” is significantly more common than the “below” – 13,767 times and 3,349 times respectively etc.

The implementation of the method used for the experiments described below first detects objects using the Faster R-CNN method, determining the classes of objects and their bounding boxes. The salient object is determined as the object with the largest bounding box.

In natural images the boundaries of the object regions, as a rule, intersect. If the intersection of the regions of the salient object and the object of interest is less than 50% of the area of the object of interest region, the relations corresponding to the decision rule are selected from the set “top”, “bottom”, “left”, “right” (that is, it is considered that there is no intersection). At an intersection of more than 50%, the ratio is selected based on a comparison of the pixel masks of the objects obtained by applying Mask R-CNN to the object regions: if there are more pixels of a significant object in the intersection zone, the “back” relation is selected, and the “in front” relation is the opposite case.

To describe spatial relationships in the framework of OSTIS Technology, a corresponding ontology was developed, the description of which and usage examples are presented in [20].

IV. Experiments

A. *Experimental evaluation of model construction*

For experimental evaluation of the semantic model construction technique from Visual Genome dataset the subset of images was selected. It is a sample of images in which each of the relations under consideration is represented by 50 regions with a grounding to the image (i.e. 50 regions for the relation “above”, 50 regions – “below”, etc. – the total of 350 regions). The examples of the images are given in Fig. 4 - 5.

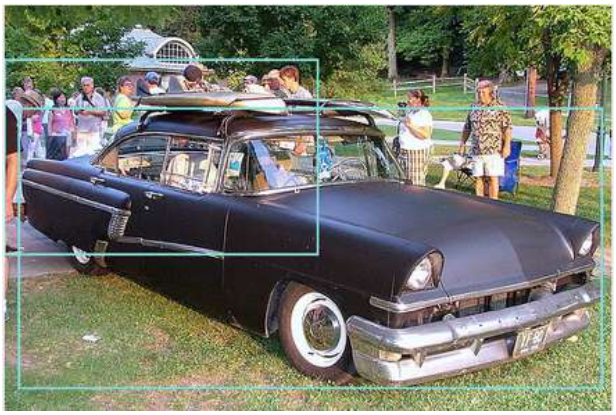

Figure 4. The example of the region grounding for “crowd behind car”.

Figure 5. The example of the region grounding for “car below trees”.

The size of the experimental set is relatively small, since it was planned to manually verify the results of determining the relations in order to evaluate not only the accuracy of the model, but also the “gameability” of the obtained results, i.e. to exclude situations where

a high indicator of the quality assessment metric (the correct result) may correspond to an expression that a human expert considers “unnatural” (for example, “the sky is blue in color” instead of “the sky is blue”) [21].

In the experiment, relationships in the selected regions are automatically determined and the results are compared with the reference (given in the dataset) and evaluated by experts (see Table I).

B. Experimental evaluation of model interpretation

To experimentally evaluate the interpretation of the constructed model for the set of regions, textual descriptions of the regions are generated by replacing the relationships with the most common synonyms from WordNet (for example,“car below tre” turns into “car parked under tree”) and the resulting annotations are compared with the reference using the METEOR metric [13].

The annotation results are also compared with the results obtained using a combined neural network [22] and purely convolutional neural network [23] approaches to annotating image regions without constructing a semantic model (Table III).

As mentioned earlier, the description of the image model in the form of natural language text has a number of significant drawbacks. The rejection of such a description and the transition to graph representation leads to the need for a transition from classical text metrics (METEOR [13], etc.) to metrics that allow us to evaluate the similarity of graph models.

A graph representation makes it possible to simplify the comparison of two models at the syntactic level, however, problems related to the semantic data presented remain urgent, which in the textual presentation faded into the background due to the large number of problems associated with the presentation form.

In general, we can distinguish the following levels of complexity of situations that arise when comparing graph image models:

- the system of terms and the system of concepts (logical ontology) coincide. In this case, the comparison is reduced to the search for isomorphic fragments, however, the problem of assessing the significance of each fragment remains relevant;
- the system of terms does not coincide, but the system of concepts coincides, i.e. the same concepts are used, but they can be named differently (for example, in the context of a street situation, the meaning of the words “car” and “automobile” will coincide). In this case, the identification and gluing of the same concepts, named differently, is additionally required. In the general case, this problem concerns not only concepts, but also specific entities;
- the system of concepts does not match. In this case, the alignment of systems of concepts is additionally required, in this case involving the definition of concepts used in the evaluated model through the concepts used in the example model.

The indicated levels relate to the case when strict coincidence of models is evaluated to the level of specific objects and concepts, however, when interpreting a model it is often enough to use a generalized model (for example, in a situation “a person sits on a chair” and a “person sits in an armchair” it is often important that a person sits and it doesn’t matter where). Thus, the task of generalizing models with subsequent comparison is also relevant. In classical textual approaches, a similar problem is partially solved by identifying synonyms.

Using OSTIS Technology to represent image models and construct relevant metrics has several advantages, in particular, one of them is the availability of means for structuring information and representing meta-information. Fig. 6 shows an example representation of similarities and differences for the two pieces of information presented in SCg-code [24].

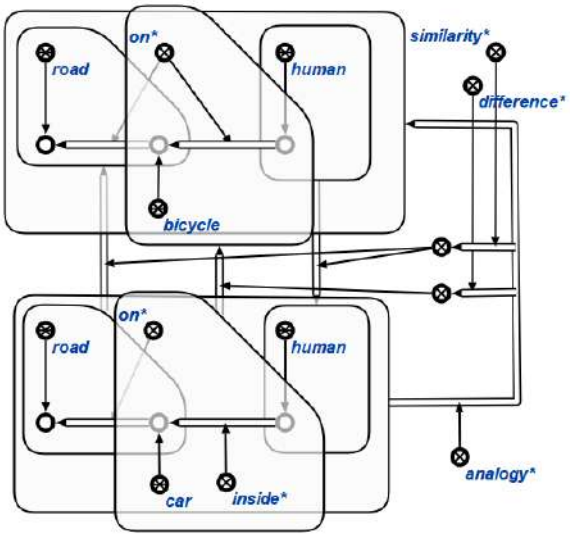

Figure 6. Representation of similarities and differences in SCg.

Note that the development of graph metrics based on OSTIS Technology will allow them to be used for other tasks, for example, to assess the quality and completeness of a student’s response in intelligent training systems, to assess the contribution of each developer to a common knowledge base, etc.

V. Results and discussion

In Table I the results of semantic model construction evaluation are shown. Object detection is considered correct, if the class labels match, the differences in bounding boxes estimation are considered insignificant in given context.

In Table II the results of spatial relations matching are shown.

Table I
Model construction evaluation.

	Number	%
Set size (relations/objects)	350/700	100
Object detection (RCNN-based)	687	98.1
Relations (dataset match)	335	95.7
Relations (visual analysis)	340	97.18

Table II
The analysis of spatial relations estimation.

Spatial relation	Visual analysis (for 50)	Model (for 50)
BEHIND	49	44
IN FRONT	48	45
RIGHT	50	50
LEFT	50	50
INSIDE	50	50
ABOVE	49	48
BELOW	49	48

In Table III the results of image captioning evaluation are shown.

Table III
Region captioning results evaluation

Coder model	METEOR
CNN + RNN [22]	0.305
TCN [23]	0.290
Semantic model	0.515

As shown in the table, the use of a semantic model for encoding information from an image significantly exceeds neural network models when constructing meaningful phrases that describe regions. According to the METEOR metric, which takes into account not only the structure of the annotation, but also its semantic variations, the proposed method shows the results by more than 60 % better than the neural network approaches.

VI. Decision making based on semantic model

To make decisions on the basis of the proposed model at this stage (with a small number of classes of objects and relations between them), a general mechanism can be used, which was examined in detail, in particular, in [25]. The specified technique assumes a reliable logical conclusion based on the logical rules and ontology of contingencies available in the knowledge base, where for each class some recommendations are assumed.

Let us consider in more detail the example of decision-making. The Fig. 7 shows the image from the surveillance camera, on which the regions of objects detected are highlighted. For convenience, some regions are omitted (in the current implementation, the detector on this image detected 25 people, 7 cars, 4 umbrellas, 2 backpacks and 4 bags).

According to the technique described above, a salient object, i.e. the key subject of the relationship, is an instance of the id1 class “pedestrian crossing” (label “crosswalk”, synset crossing.n.05). In the current implementation, this is due to the fact that it has the largest size, but subsequently the application of the ontological approach will also allow contextual information to be taken into account.

The following objects of the corresponding classes were detected in the image:

- id2, id5, id6 – class “car”
- id3, id4, id7, id8, id9 – class “person”

According to the technique for constructing a model based on existing intersections of regions, the following relationships between pairs of objects are estimated:

1) id2 ->id1: “on”
2) id3 ->id1: “on”
3) id4 ->id1: “below”
4) id5 ->id1: “on”
5) id6 ->id1: “above”
6) id7 ->id1: “inside”
7) id8 ->id1: “below”
8) id9 ->id1: “on” (detection error due to camera placement)

In the form of SCg language this model is presented as in Fig. 8.

Based on estimated relations the following captions can be generated:

1) car on crosswalk - car is parked on crosswalk
2) person on crosswalk - person is crossing the road on crosswalk

Based on the of detected objects and relations, decisions in this example are made in the context of “people cross the road, cars let pedestrians pass”. Thus, normal (regular) situations for “person” with respect to “crosswalk” – “on” and “inside”, for “car” with respect to “crosswalk” – the opposite.

The example of formal rule in SCg language is shown in Fig. 9.

Using a rule base, applying the simple inference mechanisms, the following contingencies can be distinguished:

1) traffic rules violation: car on the crosswalk – in pairs 1 and 4
2) traffic rules violation: a person is using a crosswalk – in pairs 4 and 7

Rule Violation in pair 8 will not be determined at the moment, due to the camera placement. To prevent this mistakes, it is possible to detect not regions, but masks, however, in this case, image processing will take much longer.

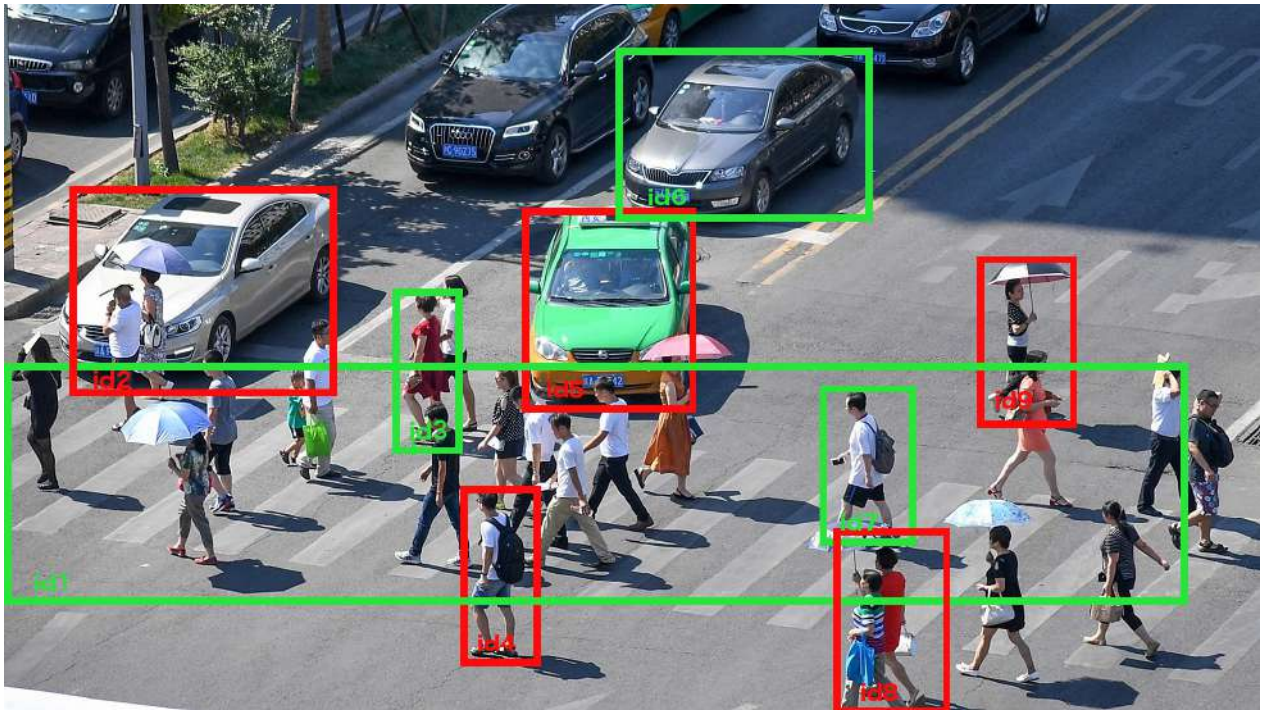

Figure 7. The example of the image for decision-making.

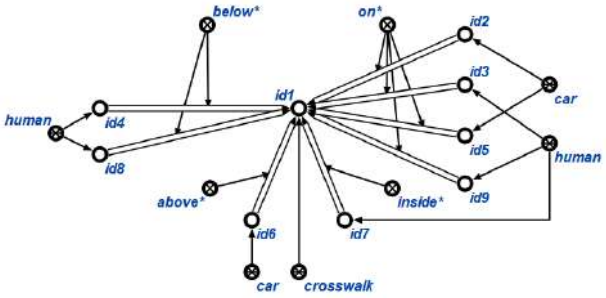

Figure 8. Semantic model of image in SCg.

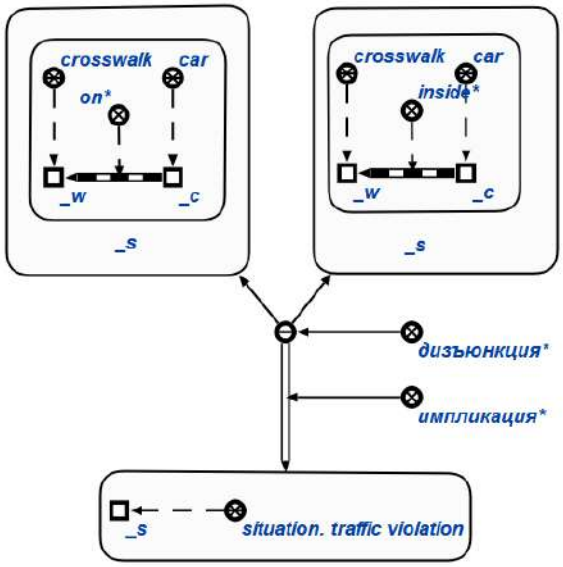

Figure 9. Example rule for decision making.

VII. Conclusion

Thus, the proposed method of constructing a semantic model analyzes less relationships between objects, which can significantly reduce the image processing time on test sets from the Visual Genome dataset and improve the quality of annotation.

It should be noted that this approach contains simplifications - the largest of the objects is considered salient and only relations between two objects are considered (i.e. only fragments of a scene-graph), also attributes of objects are not taken into account.

In further work, it is planned to use more complex approaches to determining a salient object (including based on specific subject area), the complete construction and analysis of graph scenes.

In turn, the use of OSTIS Technology to represent the model and implement the decision-making mechanism makes it possible to ensure the modifiability and scalability of the system, built on the basis of the approaches proposed in this paper, which in the future will allow to eliminate the described limitations.

Acknowledgment

The research presented in this paper was conducted in close collaboration with the Department of Intelligent Information Technologies of Belarusian State University of Informatics and Radioelectronics. Authors would like to thank the research group of the Department of Intelligent Information Technologies for productive cooperation.

References

[1] S. Agarwal, J. O. D. Terrail, and F. Jurie, "Recent advances in object detection in the age of deep convolutional neural networks," *arXiv preprint arXiv: 1809.03193*, 2018.

[2] J. Müller, A. Fregin, and K. Dietmayer, "Disparity sliding window: object proposals from disparity images," in *IEEE/RSJ International conference on intelligent robots and systems*, 2018, pp. 5777–5784.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, Cham, 2016, pp. 21–37.

[5] P. S. Hursov and N. A. Iskra, "Algoritmy detektsii ob'ektov dlya analiza izobrazhenii [Object detection algorithms for image analysis]," in *Informatsionnye tekhnologii i sistemy: materialy mezhdunarodnoi nauchnoi konferentsii [Information Technologies and Systems: materials of the international scientific conference]*, Minsk, 2019, pp. 128–129, (In Russ).

[6] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, pp. 91–99, 2015.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[9] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5410–5419.

[10] "Visual Genome," Available at: https://visualgenome.org.

[11] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. J. Li, D. A. Shamma, and M. S. Bernstein, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, no. 123(1), pp. 32–73, 2017.

[12] G. A. Miller, *WordNet: An electronic lexical database*. MIT press, 1998.

[13] S. Banerjee and A. Lavie, "METEOR: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[14] V. Golenkov, N. Guliakina, I. Davydenko, and A. Eremeev, "Methods and tools for ensuring compatibility of computer systems," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed. BSUIR, Minsk, 2019, pp. 25–52.

[15] I. Davydenko, "Semantic models, method and tools of knowledge bases coordinated development based on reusable components," in *Open semantic technologies for intelligent systems*, V. Golenkov, Ed. BSUIR, Minsk, 2018, pp. 99–118.

[16] D. Shunkevich, "Agent-oriented models, method and tools of compatible problem solvers development for intelligent systems," in *Open semantic technologies for intelligent systems*, V. Golenkov, Ed. BSUIR, Minsk, 2018, pp. 119–132.

[17] N. Iskra, V. Iskra, and M. Lukashevich, "Neural network based image understanding with ontological approach," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system: materialy mezhdunarodnoj nauchno-tekhnicheskoj konferencii [Open semantic technologies for intelligent systems: materials of the international scientific and technical conference]*, Minsk, 2019, pp. 113–122.

[18] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," in *Proceedings of the european conference on computer vision*, 2018, pp. 670–685.

[19] A. Borji, M. M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Computational visual media*, pp. 1–34, 2014.

[20] N. A. Iskra, A. L. Mezhen', and D. V. Shunkevich, "Ontologiya predmetnoj oblasti prostranstvennyh sushchnostej dlya sistemy semanticheskogo analiza izobrazhenij [Ontology of the subject area of spatial entities for the system of semantic image analysis]," in *Informatsionnye tekhnologii i sistemy: materialy mezhdunarodnoi nauchnoi konferentsii [Information Technologies and Systems: of the international scientific conference]*, Minsk, 2019, pp. 112–113.

[21] D. Shunkevich and N. Iskra, "Ontological approach to image captioning evaluation," in *Pattern Recognition and Information Processing: Proceedings of the 14 international conference*. Minsk: Bestprint, 2019, pp. 219–223.

[22] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4565–4574.

[23] N. Iskra and V. Iskra, "Temporal convolutional and recurrent networks for image captioning," in *Pattern Recognition and Information Processing: Proceedings of the 14 international conference*. Minsk: Bestprint, 2019, pp. 346–349.

[24] "IMS metasystem," Available at: http://ims.ostis.net/, (accessed 2020, Jan).

[25] V. Golovko, A. Kroshchanka, V. Ivashenko, M. Kovalev, V. Taberko, and D. Ivaniuk, "Principles of decision-making systems building based on the integration of neural networks and semantic models," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed. BSUIR, Minsk, 2019, pp. 91–102.

Принципы построения систем принятия решений на основе семантического анализа изображений

Искра Н.А., Искра В.В., Лукашевич М.М.

В статье рассматриваются принципы построения систем принятия решений. Предложен и изучен подход к анализу изображений на основе семантической модели. Результаты показывают улучшение скорости обработки и качества аннотирования на основе набора данных Visual Genome.

Received 08.02.2020