Robotic System Operation Specification on the Example of Object Manipulation

Leonid Mylnikov, Pavel Slivnitsin and Anna Mylnikova
Perm National Research Polytechnic University, 29 Komsomolsky avenue, Perm, Russia
leonid.mylnikov@pstu.ru, slivnitsin.pavel@gmail.com, novikova@yandex.ru

Keywords: Function Modeling, Diagram, Metalanguage, Recognition, Identification, Positioning, SLAM, Computer Vision, Robotic System.

Abstract: Currently, robotic system tasks are formalized with help of procedural programming languages that do not take into account the specificity of robots and are not generic in their application. The goal of the paper is to develop a method of semantic description of the sequence of operations performed by a robotic system on the example of object manipulation around them. To achieve the goal, a method of a graphical representation of a robotic system operation specification and its semantic description (metalanguage) are proposed. The paper considers the approaches to the objects' representation, determines the way object characteristics are stored, and provides the list of possible operations with objects. The obtained methods of graphical and semantic robotic system operation specification allow to assign the task without being bound to a specific technical solution. In addition, the paper provides the examples of operation assignments for the robotic arm.

1 INTRODUCTION

The current development trends of automation technologies and robotic systems are their autonomation and more enhanced complexity of the operations performed. In industry, this is caused by the need to increase production performance and quality, the need to provide high-precision machining in the context of enhanced complexity, reduced size, and shorter life cycles of products. The automation of routine operations such as welding, painting, assembly, and sorting [1] is becoming insufficient.

These operations cannot be performed in the context of a dynamic environment without solving the task of object recognition and considering the variability of motion paths. More complex operations such as manipulation of various objects, assembly and disassembly, repair, adjustment are not possible at all without taking into account the peculiarities of the environment.

Object manipulation tasks are becoming more and more widespread. They are often used for robotic systems positioning, such as warehouse maintenance robots [2], urban infrastructure facilities' maintenance [3], etc.

In order to improve the efficiency of such systems, there are various competitions. For example, the competition on the automatic object picking and sorting [4], [5].

2 THE DESCRIPTION OF OBJECT MANIPULATION OPERATIONS

The task of objects' manipulation performed by a robotic system is implemented as a sequence of operations with the manipulator and objects, which can be represented by the Figure 1, where Subj is a subject that performs the operation (manipulator), Obj is an object of the operation, V is the operation to be performed, NS are features that distinguish the subject, NV is the goal (where to move/shift/place/etc. an object), NO are the features that distinguish the object.

To specify operations and their describing structures, we will use the following symbols [6]: «=» – clarifying the concept; «{}» – merge; «<>» – mandatory part; «[]» – optional part; «|» – or; «&» – and; «\» – clarifying a new variable, «""» – rigidly given element.

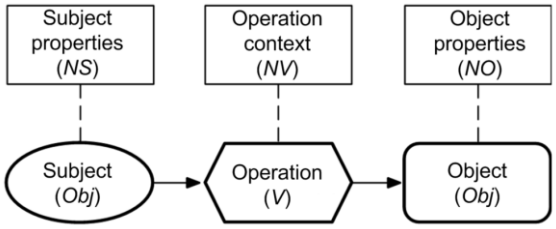


Figure 1: The functional diagram of the robotic system task assignment.

Then, syntactically a task can be described as the following structure:

<Subj>[NS]<V>[NV]<Obj>[NO].

For example, the task "Use the manipulator A to move the object B from X to Y" can be described as follows:

\Subj="A"
\V="move"
\Obj="B"
\NV=Y
\NO=X
Subj V NV Obj NO.

Let's consider the task of outdoor luminaire replacement described in [3], which is represented in natural language by the following expression: **«Use the manipulator A, replace the luminaire B with the luminaire C on the lighting column D».**

In this example, we are dealing with a complex opeartion that can be decomposed into a number of operations (e.g., remove and install) and a complex object (consisting of the objects B and C). The Figure 2 shows the structure of this task.

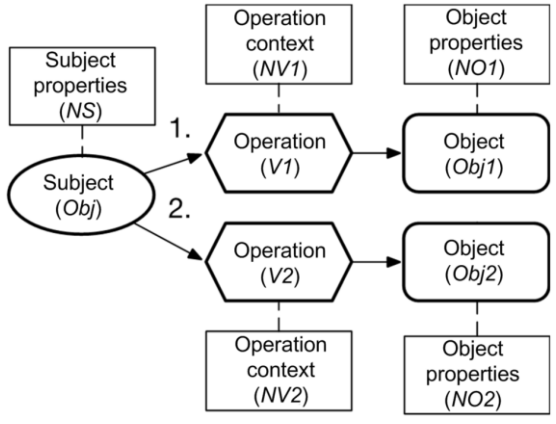


Figure 2: The functional diagram of the robotic system task assignment with a complex operation and a complex object.

Semantically, this operation can be represented as follows:

\Subj="A"
\V1="remove"
\Obj1="B"
\NV1=D
Subj V1 NV1 Obj1
\V1="install"
\Obj2="C"
\NV2=D
Subj V2 NV2 Obj2.

From the above examples we can make a conclusion that the robotic system operation can be defined by a set of operations **V={produce, repair, assemble, disassemble, replace, take, lift, put, place, insert, throw, separate the gripper, bring the gripper together, move the gripper to some location, etc.}** (some of which can be complex **V={V1,V2, etc.}**), as well as by the ability to identify objects and their specified features **NO={location, shape, color, material, etc.}** based on which a number of certain operations can be performed.

Complex operations can be decomposed in different ways into elementary ones, provided that the principle of equivalence is observed. This provides an opportunity to consider the optimization of the robotic system operation.

The optimization process of the robotic system operation is defined as follows: 1) the algorithms that implement elementary operations, 2) object recognition algorithms, 3) the knowledge about the environment (the underlying circumstances of the operation).

For example, the movement of the manipulator from the position X to the position Y can be implemented by the application of the shortest path algorithm or the ant colony optimization algorithm. In the first case, if obstacles occur in the environment, the operation cannot be performed. Moreover, different implementations of this operation will have different power consumption, different execution time, as well as they can have different execution risk assessments.

The efficiency of operation execution algorithms is not relevant if it is not possible to identify the object or there is a lack of information about the operation context.

3 APPROACHES AND METHODS FOR OBJECT IDENTIFICATION AND LOCATION DESCRIPTION IN MANIPULATION TASKS

The recognition of objects in the environment is a complex task, that includes the localization of objects, their identification, search for interaction tools with objects, the mapping of the environment and building the knowledge bases that describe the environment.

Existing object recognition methods are based on the selection of certain features, a set of elements or templates that are used to identify objects. Along with the feature detection, it is also important to consider the relationships between features that have an impact on recognition (for example, see the Thatcher effect or Thatcher illusion [7] shown in the Figure3).

Figure 3: The Thatcher Effect presented by Peter Thompson.

The recognition process and its application can be represented by the scheme shown in the Figure 4. In the paper we will consider several steps, which indicated in the figure by the numbers 2-4. Depending on the implementation algorithm, some blocks can be combined or can be executed simultaneously.

3.1 Approaches to Object Representation

The ideas of object representation (block 3 in Figure 4) used in computer vision systems are based on the theories of human object recognition. At the moment, there is a number of theories that describe the approaches to the recognition and classification of objects by humans.

The template matching theory (exemplar theory) [8] assumes that for each object there is a template in the memory. By the processing of new information, object identification requires an exact match between the object and the template from the memory. The disadvantage of the template matching theory is the need to store a large number of templates.

Another theory is the **prototype theory** [9], [10]. It involves the comparison of new information not with templates, but with some abstract object prototypes (see the Figure 5). A prototype is based on a set of examples of the object and describes their common features. There are two models for prototype formation: the central tendency model and the attribute-frequency model. [10].

In the context of computer vision, the algorithms SIFT [11] and SURF [12] are well-known, which have elements of both theories. They are based on the attribute extraction of template objects, which is used to detect objects in the image.

Later theories have developed the ideas of attribute extraction to form a prototype. For example, according to the **feature analysis theory** (see [5] and [6]), the human visual system includes feature detectors and object recognition is based on the extraction of the simplest features of objects (see Figure 6). The theory assumes a layered recognition. The simplest feature detectors detect simple features. The next feature detectors are capable of detecting more complex features. In the case of the occurrence of same features or the same combination of features for certain objects, it becomes possible to determine that these objects belong to the same class.

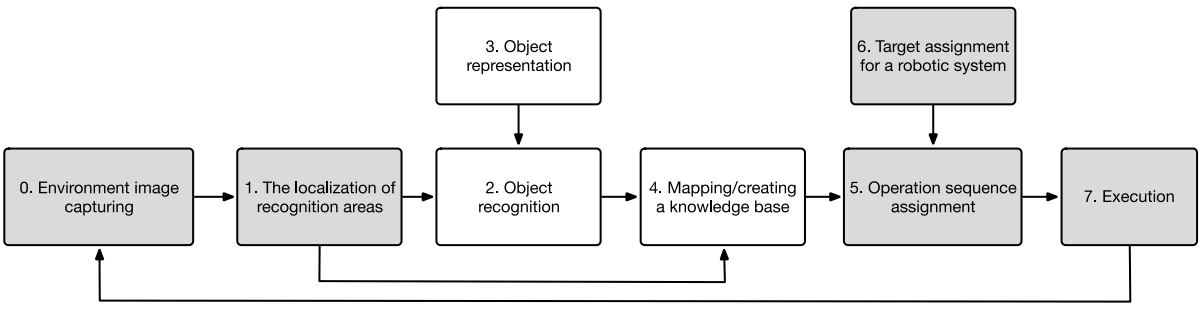
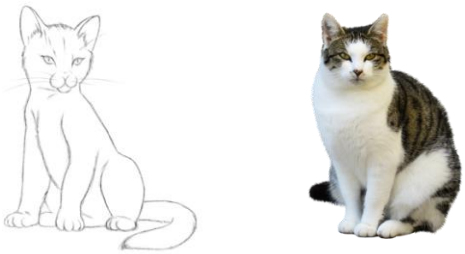


Figure 4: The sequence of steps performed by a robotic system to collect and use information about the environment.



a) *b)*

Figure 5: Prototype theory: *a)* abstract prototype, *b)* class instance.

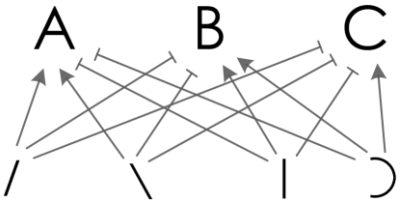
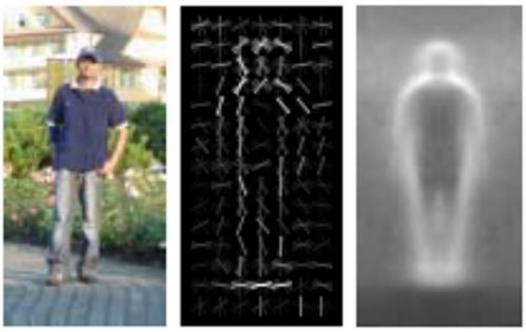


Figure 6: An example of identifying English alphabet characters using features.

The elements of this approach can be traced in the Dalal-Triggs method [15] and the Viola-Jones method [16].

The Dalala-Triggs method is based on the calculation of histograms of oriented gradients (HOG) (see the Figure 7).

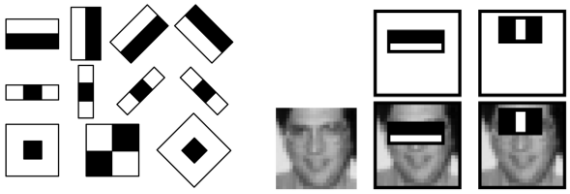


a) *b)* *c)*

Figure 7: The Dalal-Triggs Method: *a)* an image of a person, *b)* oriented gradients (features), *c)* trained structure (prototype) [15].

The Viola-Jones method is based on an integral image representation and describes objects using a combination of typical features from a limited set (Figure 8).

The ideas of feature analysis are used in convolutional neural networks (CNN) (Figure 9) [17].



a) *b)*

Figure 8: The Viola-Jones Method: *a)* a set of features, *b)* matching the features to the object (face) [16].

In this approach, a convolution operation is used, which is based on image processing with help of a convolution filter, which allows to extract the object features. The result of the convolution operation in the image is a new image. The intermediate (hidden) convolutional layers represent a matrix (a feature map, which can also be represented as an image) or a set of matrices, depending on the number of filters applied to the previous layer. Each next hidden layer of the neural network detects more complex object features compared to the previous one.

The approach described by A. Kononyuk in [19] can also be associated with the feature analysis theory. It is based on the **feature extraction of benchmark objects** from the training set **using predicate logic**.

In this approach features can capture color, spatial arrangement, etc. Each object is represented by a specific set of features. Each feature is described by a predicate. Predicate arguments are elements of the benchmark images that indicate the occurrence of the feature in it. The representation of the benchmark image (and object in it as well) is the conjunction of all the object features. An object class is represented by a training set as a disjunction of all benchmark images that contain an object of the class $\bigcup_{\omega_k \in \Omega} \bigcap_{i=1}^{h} P_i(c_{1i}, \ldots, c_{ni})$, where ω_k is the k-th benchmark image of the class Ω, h is the number of features, c_i are parts of the image that describe the occurrence of the i-th feature in the benchmark image.

Based on the classes described in the way as specified above, the object recognition task is performed as a logical inference task, which allows to identify object classes on the new image. Thus, the recognition task is to prove equivalence of objects.

Within this approach, the same set of primary features can occur for different objects, which corresponds to the feature analysis theory. For example, the primary features are lines (vertical, horizontal, parallel), and the secondary features represent more complex combinations of lines (rectangles, etc.).

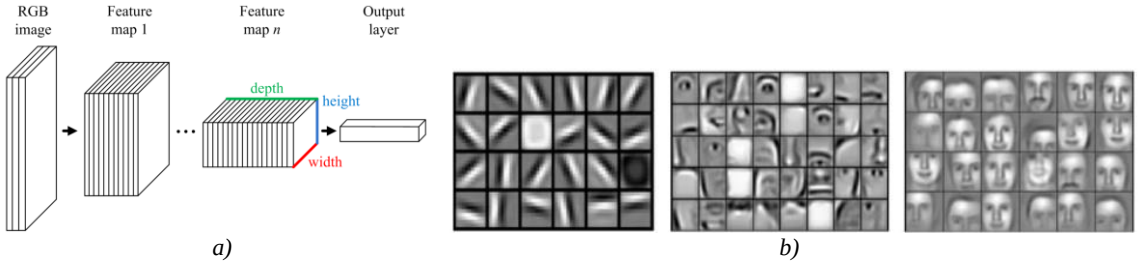


Figure 9: A convolutional neural network: a) structure, b) feature maps [18].

The implementation of feature analysis methods generated a group of approaches to object recognition that are based on the filter and mathematical functions' application for image processing and assigning objects to certain classes [20], [21].

A further development of recognition theories is D. Marr's **computational theory of human stereo vision** [22]–[24], which assumes that recognition is multistage and involves more enhanced degree of object details.

At the first stage the information about contours, edges and spots is processed. At the second stage information about the depth and object surfaces position is processed. After that, at the third stage a three-dimensional model of the detected object is formed. According to this theory, a three-dimensional representation is based on the canonical forms (e.g., cylinders). Thus, objects can be represented as a set of cylinders of different sizes with an axis, depending on the degree of detail (see [15] and [16], the Figure 10).

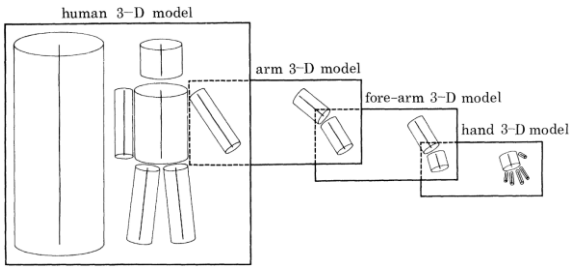


Figure 10: A visual representation of a three-dimensional model of a human using the computational theory [23].

I. Biederman developed the **recognition by component theory** [25]. According to this approach, a human being perceives objects of the real world through a certain set of geometric primitives called geons and relations among them. Thus, every object and every scene in the image can be represented by a set of primitives (Figure 11).

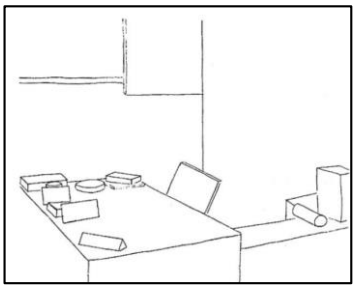


Figure 11: Scene presentation using geons [26].

Each primitive can also be described through nonaccidental properties of shapes, i.e. properties that do not change when the angle of view changes (e.g. collinearity, curvilinearity, symmetry, etc.). Thus, each component of an object can be represented by the relationship of a number of primitives, which are described by a number of nonaccidental properties. Biederman distinguishes the following relations between geons: verticality, relative size, centering, surface size join [26].

In position recognition systems, such as human pose detection [27] (see the Figure 12) or hand gesture detection [28], there are methods based on the extraction of key points or key structures of objects (e.g., skeleton structures).

Figure 12: Human pose detection using a skeleton structure [27].

The main differences of the analyzed approaches are shown in the Table 1.

Table 1: The comparison of object representation approaches.

	Features	Relations
1. The template matching theory (exemplar theory)	A set of templates	—
1.1. SIFT	The extraction of the features of the template objects to detect them in the image	—
1.2. SURF		
2. The prototype theory	A set of prototypes	—
3. The feature analysis theory	The sets of geometric features	—
3.1. Dalala-Triggs Method	A unique set of features forming the prototype	—
3.2. Viola-Jones Method	A set of features is formed for each required class	A class is defined by a combination of features
3.3. Neural networks	Each layer of the neural network represents a map of features	—
4. The computational theory	Features are relations between a number of primitive objects of the same type	Relations between the primitives form the class
5. Recognition by components theory	Features are a set of geometric primitives	The relation type is considered as a feature
6. Extraction of key points or key structures	Features are key points or key structures that describe the object's position	Relations between key points
6.1. Neural networks	The feature is a set of key points that form the object prototype structure (object's configuration)	

3.2 Object Recognition Using Features

Object recognition on an image includes the object detection [29] and object classification, and in some cases, image segmentation (semantic or instance [29]) or detection of object parts (key points) [27].

In the object recognition task, it is necessary to define classes of objects, each of which is assigned a specific set of features (based on information that is extracted from the image, including the object location on the image). For example, in the case of using object parts as features, the difference between classes can be demonstrated as shown in the Figure 13.

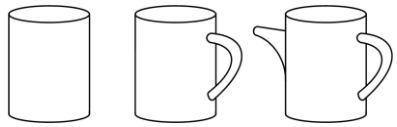


Figure 13: Objects with the same main shape feature (a – glass (cylinder), b – mug (cylinder + handle), c – kettle (cylinder + handle + cone)).

In multiple classification tasks, we often have to deal with false-positive and false-negative object identification (for example, objects can have features which can be associated with two or more classes or important object elements are not visible on the image). In this regard, specific algorithms for object recognition are being developed, which include:

- multi-step recognition process (the number of classes is reduced, the information about the object is gradually clarified) (e.g., changing a view angle [4]);
- object part recognition (objects are classified by the presence or absence of some predefined parts and their relative position; information about the position of object parts allows, among other things, to make assumptions about the position of the object in the image).

At the moment, among the methods which allow to work with objects as complex elements of the environment there are classification models and methods [30] (classifiers such as Naïve Bayes, SVM, kNN, decision trees), associative rule learning methods (Apriori, Eclat, FP-growth, OPUS, SlopeOne [31]), expert systems, predicate logic, neural networks (see [17], [18], [27], [28]) (see the Table 2).

Table 2: The peculiarities of different groups of methods used for object recognition with features.

A method	Tasks	Recognition approach
1. Classifiers	Object classification	Detection of object features and relations between them in an image
2. Association rule learning	The analysis of relations between the components of the object	Pairwise comparison of relations between features
3. Expert systems	Hypothesis testing based on information about the object	The detection of object features
4. Predicate logic	The description of objects as a set of elements (parts) and relations between them	The detection of object features and relations between them in an image
5. Neural networks	The detection of objects and their configuration (position, pose)	Feature extraction from an image using a trained neural network

Holistic objects (consisting of parts) can also be considered as elements. Thus, information about the objects position in the image allows us to estimate the relative positions of objects in the environment.

3.3 Building a Base of Knowledge of the Environment

The task is related to the tools used to obtain initial information about the environment (photo/video images, images from stereo cameras, 3-D scans, etc.), to form maps of the environment, and to store information about it (classes of objects, relations between objects, etc.).

Environment mapping is used in robotic systems' positioning and navigation (obstacle detection and shortest path search).

The SLAM methods [32] focus on collecting information about the environment in order to build a map of an area within which it is possible to move the manipulator. This approach does not provide information about the content of the environment, which includes a number of objects. The information about obstacles is enough for the robotic system navigation, but in the tasks related to the object manipulation it becomes insufficient, because it is necessary to identify objects in order to interact with them. Information just about the classes of objects and their position in the image (or in point cloud) can also be insufficient. A more detailed analysis of the object, its parts and its position can be required for object manipulations as it allows to determine the list of operations that can be performed with an object. Thus, in addition to accumulating information about the environment, it is necessary to accumulate information about its content.

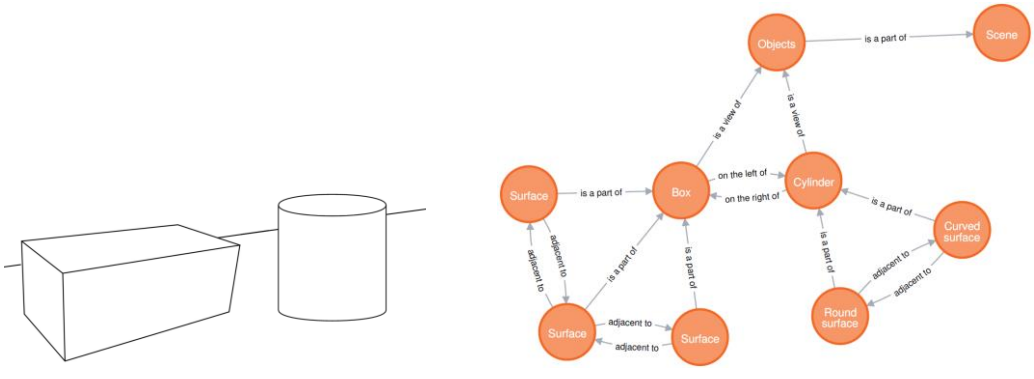
Therefore, for object manipulation, the recognition task includes the recognition of object features, the recognition of object parts, the identification of elements with which it is possible to interact and the relations between them, as well as the recognition of the scene or map of the environment. For this purpose, a knowledge base about the environment is built.

To describe the environment it is necessary to describe what objects it contains, as well as the relations between the objects. Objects can be represented as nodes with a set of some properties (color, material, etc.), and object relations can be represented as edges between nodes (left, above, etc.). Semantic networks, frame-type expert systems, network data models [33] have been used to describe such information (see an example in the Figure 14).

These approaches can also be applied to describe the object structure (for example, the representation of objects as spatial combinations geometric primitives). Primitives can be represented by nodes, and spatial relations between these primitives can be described by edges between nodes. A set of primitives, in this case, defines a list of operations that can be performed with an object.

4 CONCLUSIONS

The paper considers a method of semantic task description (metalanguage) for robotic systems and its graphical representation. The paper investigates approaches to object identification using features as well as approaches to describe the environment. It allows to formulate tasks, taking into account the context of the operation and the object features, that allows the following:



a) *б)*

Figure 14: An example of applying a semantic network for scene description: a) the scene, b) a semantic network.

- to bring the task assignment for robotic systems closer to the operations description in natural language;
- to bring invariance to the task execution, depending on the algorithm efficiency;
- to not impose algorithmic constraints to the practical implementation of basic operations.

ACKNOWLEDGMENTS

The reported study was supported by the Government of Perm Krai, research project No. C-26/692.

REFERENCES

[1] Fanuc Europe, "Robot industrial applications." [Online]. Available: https://www.fanuc.eu/de/en/industrial-applications. [Accessed: 01-Jun-2021].

[2] A. Delfanti and B. Frey, "Humanly Extended Automation or the Future of Work Seen through Amazon Patents," Sci. Technol. Hum. Values, vol. 46, no. 3, pp. 655–682, 2021.

[3] P. Slivnitsin, A. Bachurin, and L. Mylnikov, "Robotic system position control algorithm based on target object recognition," in Proceedings of International Conference on Applied Innovation in IT, 2020, vol. 8, no. 1, pp. 87–94.

[4] A. Zeng et al., "Multi-view self-supervised deep learning for 6D pose estimation in the Amazon Picking Challenge," Proc. - IEEE Int. Conf. Robot. Autom., pp. 1386–1393, 2017.

[5] A. Zeng, S. Song, J. Lee, A. Rodriguez, and T. Funkhouser, "TossingBot: Learning to Throw Arbitrary Objects with Residual Physics," IEEE Trans. Robot., vol. 36, no. 4, pp. 1307–1319, 2020.

[6] A. Novikova, "Direct Machine Translation and Formalization Issues of Language Structures and Their Matches by Automated Machine Translation for the Russian-English Language Pair," in Proceedings of International Conference on Applied Innovation in IT, 2018, pp. 85–92.

[7] P. Thompson, "Margaret Thatcher: A New Illusion," Perception, vol. 9, no. 4, pp. 483–484, Aug. 1980.

[8] R. M. Nosofsky, "The generalized context model: an exemplar model of classification," Form. Approaches Categ., pp. 18–39, 2012.

[9] E. Rosch, "Cognitive representations of semantic categories.," J. Exp. Psychol. Gen., vol. 104, no. 3, pp. 192–233, Sep. 1975.

[10] P. G. Neumann, "Visual prototype formation with discontinuous representation of dimensions of variability," Mem. Cognit., vol. 5, no. 2, pp. 187–197, Mar. 1977.

[11] D. G. Lowe, "Object Recognition from Local Scale-Invariant Features," in Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999, p. 8.

[12] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 3951 LNCS, no. July 2006, pp. 404–417, 2006.

[13] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," Psychol. Rev., vol. 65, no. 6, pp. 386–408, 1958.

[14] O. G. Selfridge, "Pandemonium: a paradigm for learning," in Proceedings on the Symposium on Mechanisation of Thought Processe, 1959, pp. 511–529.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," Proc. - 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, CVPR 2005, vol. I, no. 16, pp. 886–893, 2005.

[16] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," in Proceedings IEEE Conf. on Computer Vision and Pattern Recognition, 2001, pp. 511–518.

[17] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[18] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Unsupervised learning of hierarchical representations with convolutional deep belief networks," Commun. ACM, vol. 54, no. 10, pp. 95–103, 2011.

[19] A. E. Kononyuk, Obshchaya teoriya raspoznavaniya. Matematicheskiye sredstva opisaniya raspoznavayemykh obyektov i raspoznayushchikh protsessov. Kiyev. 2012.

[20] P. J. Diggle and J. Serra, “Image Analysis and Mathematical Morphology.,” Biometrics, vol. 39, no. 2, p. 536, Jun. 1983.
[21] Y. V. Vizilter, Y. P. Pyt’ev, A. I. Chulichkov, and L. M. Mestetskiy, “Morphological Image Analysis for Computer Vision Applications,” 2015, pp. 9–58.
[22] D. Marr and T. Poggio, “A computational theory of human stereo vision,” Proc. R. Soc. London - Biol. Sci., vol. 204, no. 1156, pp. 301–328, 1979.
[23] D. Marr and H. K. Nishihara, “Representation and recognition of the spatial organization of three-dimensional shapes,” Proc. R. Soc. London. Ser. B. Biol. Sci., vol. 200, no. 1140, pp. 269–294, Feb. 1978.
[24] D. Marr and L. Vaina, “Representation and recognition of the movements of shapes,” Proc. R. Soc. London. Ser. B. Biol. Sci., vol. 214, no. 1197, pp. 501–524, Mar. 1982.
[25] I. Biederman, “Recognition-by-Components: A Theory of Human Image Understanding,” Psychol. Rev., vol. 94, no. 2, pp. 115–147, 1987.
[26] I. Biederman, “Matching Image Edges To Object Memory.” pp. 384–392, 1987.
[27] S. Jin et al., “Whole-Body Human Pose Estimation in the Wild,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12354 LNCS, pp. 196–214, Jul. 2020.
[28] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, “Hand keypoint detection in single images using multiview bootstrapping,” Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017-Janua, pp. 4645–4653, 2017.
[29] L. Jiao et al., “A survey of deep learning-based object detection,” IEEE Access, vol. 7, no. 3, pp. 128837–128868, 2019.
[30] L. A. Mylnikov, Statisticheskiye metody intellekttsualnogo analiza dannykh. SPb.: BKhV-Peterburg. 2021.
[31] D. Lemire and A. Maclachlan, “Slope {One} {Predictors} for {Online} {Rating}-{Based} {Collaborative} {Filtering},” SIAM Data Min. (SDM’05), Newport Beach, California, April 21-23, 2005.
[32] D. Vershinin and L. Mylnikov, “A review and comparison of mapping and trajectory selection algorithms,” Proc. Int. Conf. Appl. Innov. IT, vol. 9, no. 1, pp. 85–92, 2021.
[33] G. F. Luger, Artificial Intelligence: Structures and Strategies for Complex Problem Solving, vol. 5th. 2005.