Sign-based image criteria for social interaction visual question answering

ANFISA A. CHUGANSKAYA*, *Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 60th October Anniversary prospect 9, 117312, Moscow, Russia.*

ALEXEY K. KOVALEV**, *Artificial Intelligence Research Institute, Kutuzovsky ave. 32, b.1, 121170, Moscow, Russia; Moscow Institute of Physics and Technology, Kerchenskaya str. 1 A, b.1, 117303, Moscow, Russia.*

ALEKSANDR I. PANOV†, *Artificial Intelligence Research Institute, Kutuzovsky ave. 32, b.1, 121170, Moscow, Russia; Moscow Institute of Physics and Technology, Kerchenskaya str. 1 A, b.1, 117303, Moscow, Russia.*

Abstract

The multi-modal tasks have started to play a significant role in the research on artificial intelligence. A particular example of that domain is visual–linguistic tasks, such as visual question answering. The progress of modern machine learning systems is determined, among other things, by the data on which these systems are trained. Most modern visual question answering data sets contain limited type questions that can be answered either by directly accessing the image itself or by using external data. At the same time, insufficient attention is paid to the issues of social interactions between people, which limits the scope of visual question answering systems. In this paper, we propose criteria by which images suitable for social interaction visual question answering can be selected for composing such questions, based on psychological research. We believe this should serve the progress of visual question answering systems.

Keywords: Visual question answering, social interaction, perception, action, scenario, sign-based world model

1. Introduction

In recent years, multimodal tasks have attracted increased attention. Visual question answering (VQA) [2], which combines visual and linguistic modalities, could be considered an example of such tasks. In VQA, a system is queried by an image and a question to that image and should output an answer to the given question. This task serves as the starting point for more advanced problems, such as visual commonsense reasoning (VCR) [40] and visual dialog (VD) [6]. VCR is testing the system's ability to justify its answer by forcing it to choose rationale for it. Visual dialog presents an example of a possible scenario of interaction between an intellectual assistant and a user. The data set has been collected during the interaction of two agents (Amazon Mechanical Turk workers) with each other. One agent (answerer) is exposed to an image and its caption and its role to answer questions asked by another agent (questioner) who does not see the image but the caption. Thus, the

*E-mail: anfisa.makh@gmail.com
**E-mail: alexeykkov@gmail.com
†E-mail: panov.ai@mipt.ru

Vol. 00, No. 00, © The Author(s) 2024. Published by Oxford University Press.
https://doi.org/10.1093/jigpal/jzae026

questioner implicitly solves the task of refining the representation of a scene depicted on the image. Explicitly, the collected data are used in a situation where the answerer is replaced with a computer system and asked to answer the last question about the image in the dialog considering the dialog history. Thereby, the problem statement determines the types of questions asked. They are very clear-cut, as the questioner tries to refine the scene understanding. Most questions ask about external features (color, shape, size, appearance, etc.) and the existence/counting of objects. Questions of this kind (about color, shape, quantity, spatial relationship between objects, etc.) characterize most of the existing data sets for the VQA task, such as VQA v2.0 [8], CLEVR [15], GQA [14], and visual genome [17]. The symbol grounding problem [2, 5, 11] plays a decisive role in the effectiveness of such systems, as the system should relate the information from the image and the textual data with its internal representation of concepts. However, this problem statement limits the number of question types and does not allow for modeling complex cognitive functions as answers are based directly on the system's sensory input.

Other works complicate the task by adding questions that are not answered directly in the image, which leads to the need to use external knowledge bases [22, 28, 36, 37]. The questions from the VCR data set [40] look the most interesting. However, the format of the answer by choosing from multiple options allows us not to refer directly to the reasoning about what is happening in the image, but to act using the elimination method.

In this paper, we propose to turn to developments in the field of studying perception, in particular social perception, which can be used to collect data to advance research in the field of VQA. The main purpose of our research is to conduct a methodological study of the correlation between visual information and the compilation of a verbal description of situations of a social scenario. This is solved in the tasks of compiling algorithms for collecting the 'gold standard' data set that could be used for training of VQA systems. The contribution of our work is the classification of image types that can be used to build data sets on social interactions, and the formed criteria for selecting such images based on psychological research.

The paper is structured as follows: section 2 provides an overview of existing data sets for the visual question answering task. Section 3 considers the psychological aspects of visual perception, offers a classification of images and forms the main criteria for selecting images for visual question answering. Section 4 provides an example stage of selection of visual information, and section 5 concludes our work.

2. Visual question answering data sets

For a long time, the VQA [2] data set and its modification VQA v2.0 [8] served as standard benchmarks for the visual question answering task. The data set consists of both real-life images and animated scenes and covers a wide range of open questions. Questions vary in complexity: determining some property of an object (e.g., color, 'What color is the hydrant'?, 'What color is the ground'?), determining the type of object in an image ('What animal is in the window'?, 'What type of meat do you see'?), including requiring the ability to recognize text on an image ('What brand is the pickup truck'?), counting entities ('How many people are in the image'?, identifying the entity state ('Is the water still'?), determining actions ('What is the kid doing'?, 'Why are the men jumping'?), and other types of questions ('Is this a color photo'?). A distinctive feature of such questions is that all of them can be answered using only the information from the image. Modern models have already outperformed the human result (80.78 overall accuracy) on this data set (84.03—BEiT [39], 82.00—OFA [38]); however, the task of visual question answering cannot be considered completely solved.

Other similar data sets [14, 15, 17] have the same limitations, although they complicate some aspects of the problem, e.g. answers to CLEVR [15] questions require longer chains of reasoning.

An interesting approach to further development of the visual question answering task is the development of adversarial data sets [21, 31], which are collected on the basis of error analysis of existing state-of-the-art solutions. This approach advances research, but does not fundamentally solve the problem with the presence of the answer directly on the image.

Another direction is connected with the addition of a question for the answer to which it is necessary to attract external knowledge from the knowledge base or Wikipedia [22, 28, 36, 37]. However, in most cases, the answers to such questions come down to fact-checking and do not require reasoning, e.g. 'What is this device used for'?, 'What is the name of the man who built the arc in the bible due to this natural catastrophe'?, 'What is the popular brand of soda'?

Separately, there are data sets serving specialized visual question answering tasks, such as assisting visually impaired people [9, 10] and analysis of medical radiology images [1, 12, 16, 18] and images of pathologies [13].

Thus, developing criteria by which images could be selected for questions about social interactions is an important task.

3. Psychological aspects of visual cognition

One of the most popular and relevant areas in the field of machine learning is the task of analyzing visual information and drawing conclusions based on it [33]. There are two levels of image analysis: perceptual and semantic. First of all, the analysis of visual information for determining objects in psychology, their location and the construction of a perceived image refers to the field of perception (primarily visual) [34]. Such task is reduced to the classification of objects and analysis based on correlation with a reference (sensory standard). Within the framework of perception, psychologists identify up to 30 signs that can describe images and their mutual location [25]. Perceptual classification of objects is now solvable in machine learning and help to distinguish an image [40]. But another model of recognition and construction of learning methods is needed for the class of social actions. What is the difference between social perception and sensory perception? What models can be implemented? It warrants a closer look.

Social perception is the process of people's perception of social relations, actions, and assessments as part of interaction and communication [26]. At the same time, the process of perception becomes closely related to the sphere of meanings and understanding of behavioral scenarios. In machine learning, the task of understanding social relationships arises within the framework of visual common reasoning [2].

A particularly important part of the problem of reasoning in everyday logic is the task of using visual experience. Perceptual information within the framework of ontogenetic development serves as the basis for the construction of the main categories, their connection with cultural patterns, and the construction of a single construct: 'image-nominative designation—social significance' [35]. In Piaget's works [24], the ideas of socially significant phenomena became possible based on the sensorimotor stage of intellectual development, where the initial analysis and synthesis of information are understood from actions with objects, and subsequently move to more complex categories.

Vygotsky [35] attributed such a process of forming the internal structure of the mentality to external social activity (above all, through communication). The semantic-symbolic function of consciousness develops during this process. It transfers mentality into a special dimension of perception of the surrounding reality in a system of meanings. The concept of a 'scenario' refers to the level of social values.

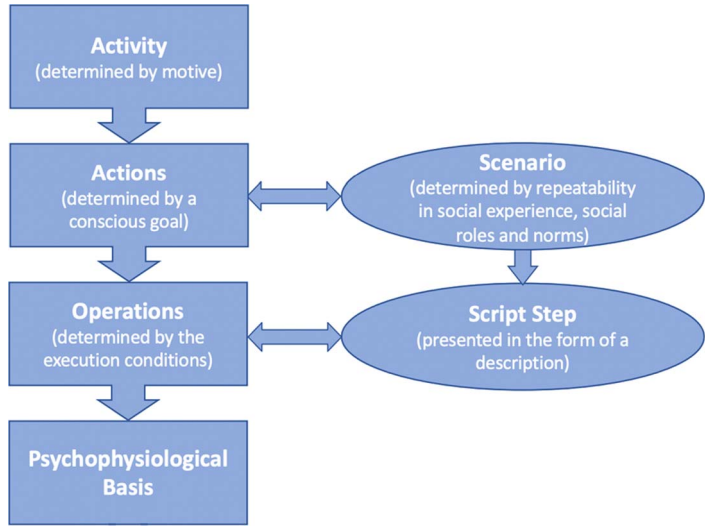

FIGURE 1. Ratio of the concepts of 'activity' and 'scenario'.

In the 1920s and 1930s, Piaget [24] and Bartlett [3] described the importance of subjective interpretation processes. They introduced the concept of 'schema' or 'knowledge structures', summarizing knowledge and accumulated experience in relation to events classes. The schema becomes the basis of expectation associated with the same stimuli and events in the future, thereby acting as the basic mechanism of scenario behavior.

The term 'scenario' came into use in the works of Schanck and Abelson [27]. It was introduced to explain how people, getting into many familiar situations, perform strictly defined roles, making a choice from a set of behaviors. The concept of scenarios is based on the idea that people enter into predictable, almost ritualistic interactions to meet their needs with little social stress and cognitive effort [26].

Schanck and Abelson subsequently proposed identifying the terms of the frame and the scenario with the most characteristic issues related to this situation. The answers to these questions are useful for understanding this situation. In order to understand the action that is being described or observed, a person is often forced to ask such questions: 'Who performs the action (agent)'?, 'What is the purpose of the action (intention)'?, 'What are the consequences (effect)'?, 'Who is affected by this action (recipient)'?, 'How is it produced (tool)'? [28]. The construct of the 'scenario' takes into account the appeal to visual information.

These areas of research in the cognitive and behavioral fields can be operationally represented in the form of a scheme. It can be considered in both contexts: activity as an objective reality [19] and the scenario as a mental construct. The ratio of these two methodological sections of the review is shown in Figure 1.

Activity is determined by a motive. They exist objectively and are determined by a conscious goal. Actions are behavioral acts. They connected with the objective and social world. Operations are methods of performing the action which are set by the execution conditions. On the other hand, a scenario exists mentally as a cognitive scheme. It is a field of knowledge and is represented in linguistic terms, determined by replicability in social experience, social roles and norms. A scenario consists of steps: sequential elements of the script, presented in the form of a description [30].

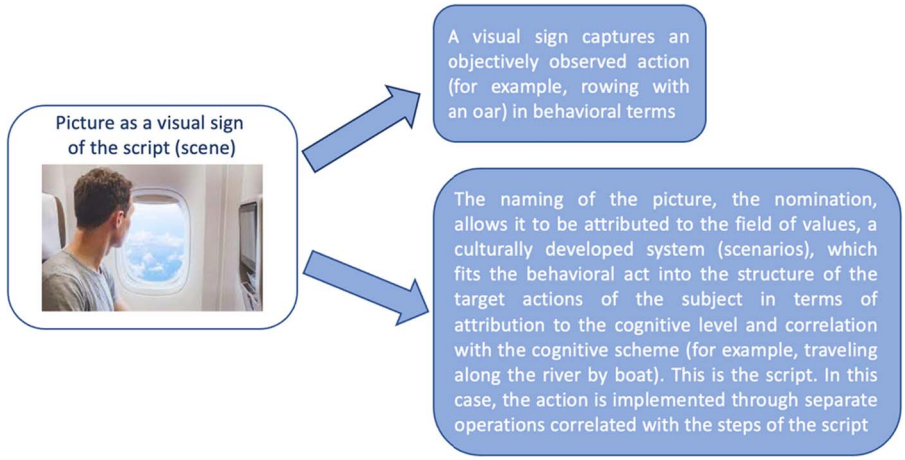

FIGURE 2. The model of the script structure with the visual-verbal structure of the sign.

At the heart of the construction of visual data sets with the theme of social interaction, it is necessary to take into account the ratio of the image as perceptual information and a system of meanings that are revealed in the field of the subject's linguistic consciousness and presented in the form of scenarios. At the same time, the verbal description of the image can be carried out at the operational level, for example, the description of the picture as a form, type, color of objects, and not revealing the social significance of the situation. The social function of the depicted is revealed by the correlation with the step of the scenario, where the general meaning of social interaction is determined through the place of what is happening, the role of subjects, objects performing a function in the activity (e.g. labor or household). There is a methodological problem of establishing a connection between an image clearly attributed to a certain type of social interaction and a set of those significant features that allow it to be attributed to this class. To solve this problem, we proposed using the concept of a visual sign, which allows one to combine significant performance indicators with motivational and needs components and scenarios as a mental scheme of social interactions and representations. We show it in Figure 2.

The scenario is expressed by a visual sign (picture), which captures some separate behavioral action (which is essentially a step of the scenario). But the visual sign itself turns to the cognitive level, in which the action and the scheme underlying its target basis, consisting of separate steps operations, are revealed. The visual sign-image appears in two planes: the perceptual reflection of the scenario step as a behavioral act and the actualization of the scenario as an action, part of the scheme in cognitive terms (the concept of 'schema in the sense of Bartlett's works').

The visual sign of the scenario cannot be considered outside the cognitive level and nomination, because it does not allow one to determine, see the action and activity behind it, but can only be defined as a separate behavioral act (operation). We propose a scheme of sequential steps for compiling a data set collection algorithm for this task, it is shown in Figure 3.

Scenario chains are built on the basis of motivational-target compliance, and they can be built by asking questions about what happened before and after.

The classification of action types and the process of their formation were proposed by Nikolai Bernstein. He identified five levels of actions that are characteristic of human activity. Each level was designated with a Latin letter: A, B, C, D and E [4]. According to Bernstein, movement levels

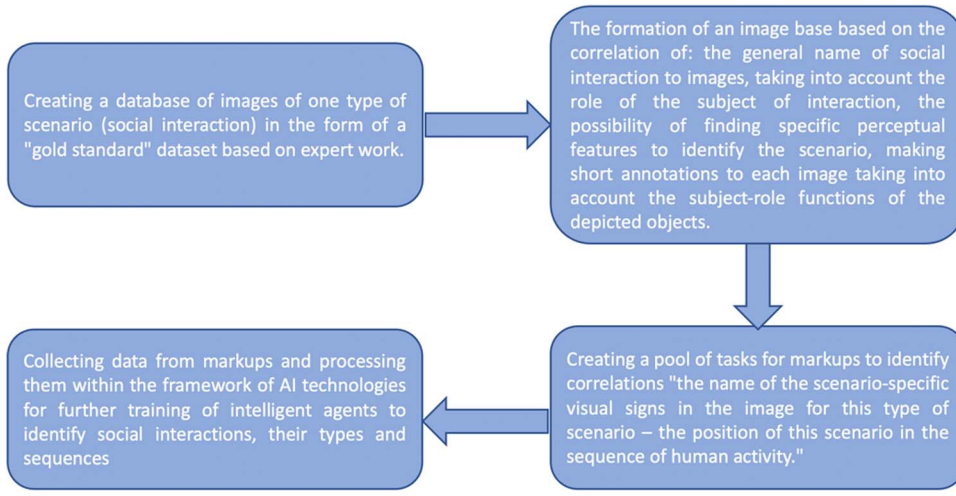

FIGURE 3. The model of the data set collection scheme.

have their neurophysiological organization. The higher the goal of the action, the higher the level of the corresponding anatomical and physiological organization.

A is the lowest level. It includes tonic movements, e.g. trembling. They are regulated by simple neurophysiological reactions which are similar both for humans and animals. **B** is the level of coordinated movements. These are coordinated actions without the need for spatial orientation, e.g. hand movements while lying on the surface. The **C** level needs movement and orientation in space. For example, you need to go around some obstacles. **D** is the subject level that is typical for a person. At this stage, the movements are built according to the logic of the subject. For example, if a person needs to take a cup without a handle, they can find an action consistent with this goal. E is the level of the speech muscle movements. This level is activated when we speak, express our thoughts or in symbolic movements, e.g. dancing [4]. According to the organization of the levels of action, Bernstein proposed the concept of 'models of the necessary future' The higher the motive of the activity, the higher the levels connected to its implementation. Levels C and D are significant for the construction of perceptual images. Figure 4 shows a diagram of the levels and their brain representation.

The paper [25] identified the four stages of perceptual actions of a preschool child to build a holistic image of the subject. At the first stage, the subject is perceived as a whole. It can be claimed that at this stage, there is a comparison of the general characteristics of the object with sensory standards. There is a primary categorization of the object into a certain class. At the second stage, the main parts of the object are isolated and their properties (shape, size, color, etc.) are determined. At the same time, the signs that will relate to the main ones will be updated depending on the perceptual attitude, which updates the field of attention and the appropriate signs. At the third stage, the spatial relationships of the parts are distinguished relative to each other (above, below, right, left) and to the entire context. At the fourth stage, an examination is carried out by repeated holistic perception of the object. All the data obtained about the object properties is analyzed. The results of the performed perceptual actions are synthesized into a single image.

There are many classifications of images [7]. However, in the tasks of psychological diagnostics and use in psychological experimental schemes, four main types can be distinguished. Here is a list of image types:

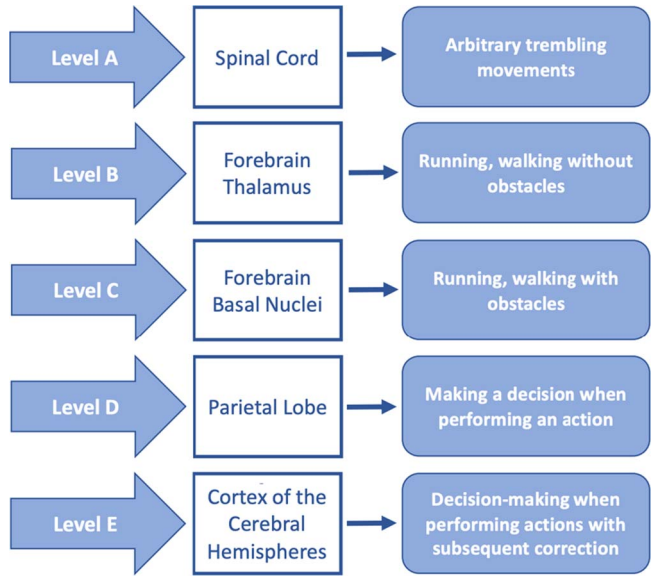

FIGURE 4. Table of levels of movement construction according to N. Bernstein with correlation of the type of action and the level of its brain organization

1. The first type is realistic images. This class includes photo and video frames from real life. Examples of this type of images are shown in Figure 5 a, b from the VCR data set [40].
2. The second type encompasses photos and video scenes that curate certain image properties for better semantic attribution to a certain category. This class includes scenes from movies or photos with processing. Examples of this type of images are shown in Figure 5 c, d from the VCR data set [40].
3. The third type is schematic images (Figure 6a) [32].
4. The fourth type involves sequences of schematic images (Figure 6b) as a set of visual reference signs for composing a story [29].

The theme of the connection of a visual image and the disclosure of a verbal scheme of actions and meanings is based on the problem of composing a story based on a picture. Child psychologists and speech therapists offer algorithms for composing stories based on a picture and a system for checking the depth of proficiency at the speech level [29]. To develop a story based on a picture, ready-made question diagrams are even used, which help to understand the aspect of the meaning of social actions in the picture based on the allocation of significant perceptual components.

For example, there are such variants of leading questions depending on the number of drawn characters [29]:

– Who is depicted?—a subject of activity;
– What is he/she doing?—action;
– What is his/her face expression?—emotions;
– What is he/she wearing? What color is it?—the role of the character;
– What is the time of year in the picture and the place?—context.

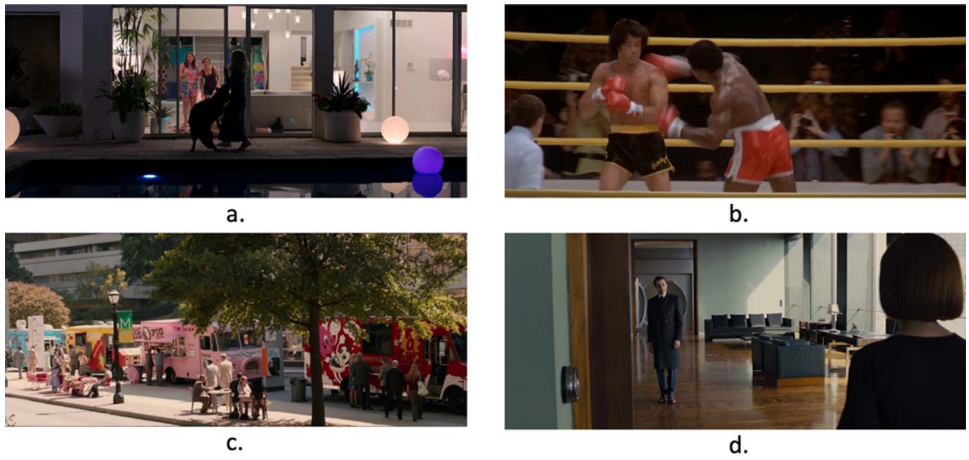

FIGURE 5. Realistic images and photos and video scenes. Examples are taken from the VCR data set [40].

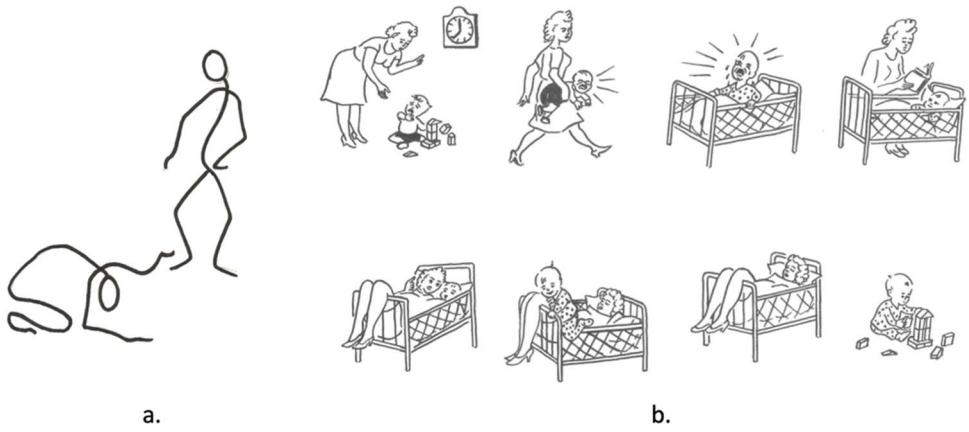

FIGURE 6. Examples of schematic images (a) [32] and sequences of schematic story-images (b) [29].

Based on a review of psychological research [7, 29, 33–35], we have identified the main criteria for selecting images of all types for a data set on social interaction:

1. It must contain visualized external action that is perceptually observable and can be distinguished by processing visual information from similar actions.
2. Three planes of the image can be highlighted: the social roles of the actor, the subject of the action (allowing one to determine the tool component of the activity, its purpose and meaning), the circumstances of the action (the surrounding space clarifying the circumstances of the action and clarifying the semantic field of the actualizing situation).
3. In social interaction we must pay attention to the direction of the subject's gaze and movement, hand positions and facial expressions.

Here is an example of creating a data set of social interactions based on the features of visual information and types of images. The schema of the algorithm preliminary stage of preparing the data set for subsequent modeling of reasoning is presented in Figure 7.

At the next stage we use the work of the trial markup in three stages (Figure 8).

4. Example of visual information selection stages and discussion

As an example of applying criteria and types to the selection of images for data sets with social interactions, let us consider the algorithm of the work of researchers and markups.

An action without defining the target side and component in the structure of the activity actually becomes an operation. In complex social actions, scenarios are constructed according to the type of instruction (a sequence of actions in the psychological sense, having a goal and replacing each other) and ritualized (nuclear, the goal is in the main core of the action). Therefore, the work begins with the selection of the scenario of the created data set. For practical purposes, realistic images are the most relevant, the first type of their classification, presented in Section 3 of this paper. However, other types of images allow you to solve narrower tasks. The analysis of schematic images leaves only significant elements of the situation, thereby forming a visual prototype-standard. Image stories allow you to take into account the dynamic component and understand the semantic part of the images.

Preparation for the presentation of tasks for the text markups consists of several stages:

1. Selecting from the list the appropriate name of the situation (options from 3 to 7–9). Most likely, the name will include the subject, action and place/circumstances, or can be presented in generalized form. For example, for a situation of pack for a trip, this could be the following list of names: take a taxi, get on the plane, pack for a trip, pass an airport security check.
2. Marking the perceptual side of the situation with the cursor on the image. Defining significant objects for the scene.
3. Drawing up a description of the situation/scene in the image by the marketer.
4. The marketer's answer to questions about the purpose of actions and the pre- diction of further steps. The first part of the questions and descriptions from stages.3 is related to what is depicted and given, including in the perceptual plan. The second part is related to the target questions (e.g. why, for what purpose). Questions are asked before the respondent is out of possible answers. Further, clarifying questions are asked to the answers of the second level (target), allowing the machine to make predictions in the future.

So, for the 'Journey' scenario [20] in the trial version, you can imagine five steps that can be visualized:

- Scenario 1: 'Pack a suitcase';
- Scenario 2: 'Take a taxi';
- Scenario 3: 'Check in at the airport';
- Scenario 4: 'To fly on an airplane';
- Scenario 5: 'To meet the arriving'.

Images that meet the criteria for the data set for images with social interaction are selected for each scenario. The first stage of the study with the work of the markup should include several tasks with appropriate instructions for each image of the corresponding scenario. For example, for scenario 1: 'Packing for a trip'—choose one of the appropriate names from the list (Figure 9).

Next, mark the objects with a marker cursor on the image, which help visually refer it to the selected name of the situation, thus establishing a correspondence between specific visual features and their semantic nomination (Figure 10).

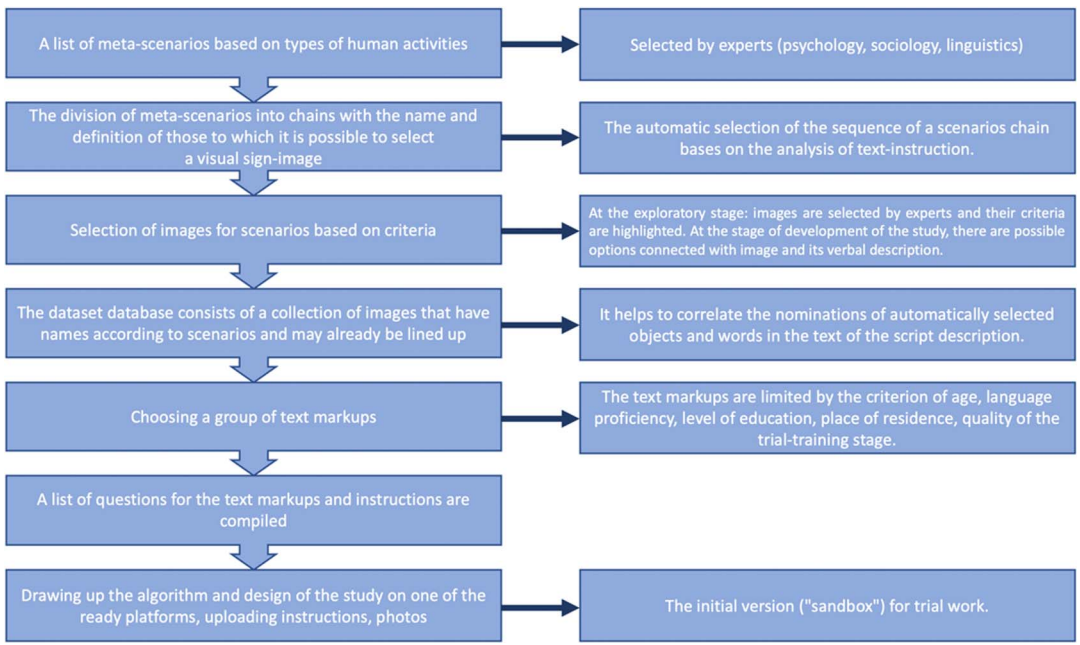

FIGURE 7. The scheme of the study of the correlation of a visual sign and a scenario.

FIGURE 8. The trial work with images and verbal information for data set.

FIGURE 9. Example of the stage one for the 'Packing for a trip' scenario. 'Choose a suitable name for the image: take a taxi, get on the plane, pack for a trip, pass an airport security check'.

FIGURE 10. Example of the task in which the marker must mark with the cursor those parts of the image (objects, part of the situation, the hero of the scene, etc.) that help to understand what kind of situation it is. 'In the drawing, select the objects/circle with a marker cursor those parts of the image that help you relate it to the selected name of the situation.'

One of the directions of modern research of visual perception is the identification of those visual parameters that can act in the construction of a complex (semantic) image. For example, when investigating the problem of increasing the number of informative points used by the visual system in comparison with the number that directly falls on the retina of the eye, Shapoval [30] shows that when an image is perceived, a 'field picture' is formed, structured not only by the visible (actually indicated in the image itself) but also by the imaginary or invisible axes. Such results determine the prospects of research in the field of the extra-sensory basis of human perceptual activity and the need to focus on it when modeling perception by means of artificial intelligence.

5. Conclusion

The correlation of two methodological constructions within the frameworks of the activity approach and scenario concepts is possible when building a connection at the level of an action, its target component and its reflection in the scenario as a semantic representation scheme. There are two

subtasks of the connection of the visual image and its nomination in the connection through the nomination with the structure of the script and reasoning.

Objectively observed actions in the psychological sense within the framework of a particular activity include the basis of visual information (external actions). The name of the action allows us to define it as belonging to the scenario level. The description of the action reveals the level of the cognitive scheme and the actual script that exists in the language. It is solved on the basis of compiling a verbal description of the action and finding out its target side. Methodology-wise, it helps us go to the stage of reasoning based on the clarification of cause-and-effect relationships. This leads to modeling the reasoning process in artificial intelligence [23].

References

[1] A. B. Abacha, S. A. Hasan, V. Datla, J. Liu, D. Demner-Fushman and H. Muller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*, Lugano, Switzerland, September 9-12, CEUR Workshop Proceedings, vol. 2380, 2019. CEUR-WS.org.

[2] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra and D. Parikh. VQA: visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433. IEEE Computer Society, Santiago, Chile, 7-13 December 2015. https://doi.org/10.1109/ICCV.2015.279.

[3] F. C. Bartlett. Remembering: a study in experimental and social psychology. *Philosophy*, **8**, 374–376, 1932.

[4] A. N. Bernstein. *On Dexterity and its Development [in Russian]*. Publishing House "Physical Culture and Sport", Moscow, 1991.

[5] T. R. Besold and K. U. Kuhnberger. Towards integrated neural–symbolic systems for human-level ai: two research programs helping to bridge the gaps. *Biologically Inspired Cognitive Architectures*, **14**, 97–110, 2015.

[6] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh and D. Batra. Visual dialog. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21-26 July 2017, pp. 1080–1089. IEEE Computer Society, 2017. https://doi.org/10.1109/CVPR.2017.121.

[7] J. Gibson. *The Perception of the Visual World*, J. Gibson ed. Houghton Mifflin, Boston, 1950.

[8] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra and D. Parikh. Making the V in VQA matter: elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21-26 July 2017, pp. 6325–6334. IEEE Computer Society, 2017. https://doi.org/10.1109/CVPR.2017.670.

[9] D. Gurari, Q. Li, C. Lin, Y. Zhao, A. Guo, A. Stangl and J. P. Bigham. Vizwiz-priv: a dataset for recognizing the presence and purpose of private visual in- formation in images taken by blind people. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 15-20 June 2019, pp. 939–948. IEEE Computer Society, 2019. https://doi.org/10.1109/CVPR.2019.00103.

[10] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo and J. P. Bigham. Vizwiz grand challenge: answering visual questions from blind people. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 18-23 June 2018, pp. 3608–3617. IEEE Computer Society, 2018. https://doi.org/10.1109/CVPR.2018.00380.

[11] S. Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, **42**, 335–346, 1990.

[12] S. A. Hasan, Y. Ling, O. Farri, J. Liu, H. Muller and M. P. Lungren. Overview of imageclef 2018 medical domain visual question answering task. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, Avignon, France,10-14 September 2018. CEUR Workshop Proceedings, vol. 2125, 2018. CEUR-WS.org.

[13] X. He, X. Zhang, S. Ren and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 27-30 June 2016, pp. 770–778. IEEE Computer Society, 2016. https://doi.org/10.1109/CVPR.2016.90.

[14] D. A. Hudson and C. D. Manning. GQA: a new dataset for compositional question answering over real-world images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 15-20 June 2019, pp. 6693–6702. IEEE Computer Society, 2019. https://doi.org/10.1109/CVPR.2019.00686.

[15] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21-26 July 2017, pp. 1988–1997. IEEE Computer Society, 2017. https://doi.org/10.1109/CVPR.2017.215.

[16] O. Kovaleva, C. P. Shivade, S. Kashyap, K. Kanjaria, J. T. Wu, D. Ballah, A. Coy, A. Karargyris, Y. Guo, D. J. Beymer, A. Rumshisky and V. V. Mukherjee. To-wards visual dialog for radiology. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, Online, July 9, 2020, pp. 60–69. Association for Computational Linguistics, 2020.

[17] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. J. Li, D. A. Shamma, M. Bernstein and L. Fei-Fei. Visual genome. *Computer vision*, **123**, 32–73, 2017.

[18] J. J. Lau, S. Gayen, A. B. Abacha and D. Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, **5**, 180251, 2018.

[19] A. N. Leontiev. Psychology of the image [in Russian]. *Vestn. Mosk. un-ta. Ser. 14, Psychology*, vol. 2, pp. 3–13, 1979.

[20] E.M. Lesnichaya. Frame structure of the phraseological meaning of the concept "JOURNEY". In *Bulletin of the YUrGGPU*, Vol. 9, pp. 236–244, 2008. URL: https://cyberleninka.ru/article/n/freymovaya-struktura-frazeologicheskogo-znacheniya-kontsepta-puteshestvie (accessed 28.08.2022).

[21] L. Li, J. Lei, Z. Gan and J. Liu. Adversarial VQA: a new benchmark for evaluating the robustness of VQA models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 10-17 Oct. 2021, pp. 2022–2031. IEEE Computer Society, 2021. https://doi.org/10.1109/ICCV48922.2021.00205.

[22] K. Marino, M. Rastegari, A. Farhadi and R. Mottaghi. Ok-vqa: a visual question answering benchmark requiring external knowledge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 15-20 June 2019, pp. 3190–3199. IEEE Computer Society, 2019. https://doi.org/10.1109/CVPR.2019.00688.

[23] G. S. Osipov. Signs-based vs. symbolic models. In *Advances in Artificial Intelligence and Soft Computing*, G. Sidorov and S. N. Galicia-Haro, eds, pp. 3–11. Springer International Publishing, Cham, 2015.

[24] J. Piaget. *Les M'Ecanismes Perceptifs [in French]*. Presses universitaires de France, Paris, 1961.

[25] N. N. Poddyakov. *Features of Mental Development of Preschool Children [in Russian]*. Professional Education Publishing House, Moscow, 1996.

[26] L. Ross and E. Nisbett. *The Person and the Situation: Perspectives of Social Psy- Chology*, 2nd revised edition. Pinter and Martin Publishers, London, 2011.

[27] R. C. Schank and R. P. Abelson. *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Lawrence Erlbaum, New York, 1977.

[28] D. Schwenk, A. Khandelwal, C. Clark, K. Marino and R. Mottaghi. A-okvqa: a benchmark for visual question answering using world knowledge. In *Computer Vision – ECCV 2022. ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, T. Hassner eds., Lecture Notes in Computer Science, vol. 13668, pp. 146–162. Springer, 2022.

[29] N. Y. Semago and M. M. Semago. Theory and practice of assessing the mental development of a child. In *Preschool and Primary School Age*, N. Y. Semago, M. M. Semago, eds. Speech, Saint-Peterburg, 2005.

[30] A. V. Shapoval. Description of the image structure in modern art criticism analysis [in Russian]. *Izvestiya Samarskogo nauchnogo tsentra Rossiyskoy akademii nauk*, **13**, 240–246, 2011.

[31] S. Sheng, A. Singh, V. Goswami, J. A. L. Magana, W. Galuba, D. Parikh and D. Kiela. Human-Adversarial Visual Question Answering. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, J. Wortman Vaughan, eds., pp. 20346–20359. Curran Associates, 2021.

[32] L. N. Sobchik. *Hand-Drawn Apperceptive Test RAT*. Speech, Saint-Petersburg, 2002.

[33] R. Vedantam, X. Lin, T. Batra, C. L. Zitnick and D. Parikh. Learning common sense through visual abstraction. In *IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 7-13 Dec. 2015, pp. 2542–2550. IEEE Computer Society, 2015. https://doi.org/10.1109/ICCV.2015.292.

[34] B. M. Velichkovsky. Cognitive science: Fundamentals of the psychology of cognition. In 2 Volumes [in Russian], Smysl/Akademiya, Moscow, 2006.

[35] L. Vygotsky. *Thinking and Speaking*. The M.I.T. Press, Cambridge, Mass., 1962.

[36] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som and F. Wei. Image as a Foreign Language: Beit Pretraining for all Vision and Vision-Language Tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 17-24 June 2023, pp. 19175–19186. IEEE Computer Society, 2023. https://doi.org/10.1109/CVPR52729.2023.01838.

[37] P. Wang, Q. Wu, C. Shen, A. Dick and A. van den Hengel. Fvqa: fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 2413–2427, 2018.

[38] P. Wang, Q. Wu, C. Shen, A. van den Hengel and A. R. Dick. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, C. Sierra ed., Melbourne, Australia 19-25 August 2017, pp. 1290–1296. IJCAI, 2017. https://doi.org/10.24963/ijcai.2017/179.

[39] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou and H. Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Proceedings of the 39 the International Conference on Machine Learning*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato eds., Baltimore, Maryland, USA, 17-23 July 2022, Vol. 162, pp. 23318–23340. PMLR, 2022.

[40] R. Zellers, Y. Bisk, A. Farhadi and Y. Choi. From recognition to cognition: visual commonsense reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 15-20 June 2019, pp. 6713–6724. IEEE Computer Society, 2019. https://doi.org/10.1109/CVPR.2019.00688.

Received 20 May 2022