Situation Monitoring Based on a Bimodal Model of the World

Aleksander A. Kharlamov [1,2,3,4], Roman M. Zharkoy [5]

[1] *RAS Institute of Higher Nervous Activity and Neurophysiology, 5A Butlerov str., Moscow, 117485, Russia*
[2] *Moscow State Linguistic University, 38 Ostozhenka str., Moscow, 119034, Russia*
[3] *Higher School of Economics University, 20 Myasnitskaya str., Moscow, 101000, Russia*
[4] *Moscow Institute of Physics and Technologies, 9 Institutskiy side str., Dolgoprudny, Moscow region, 141701, Russia*
[5] *Intelligent Security Systems, 19 Suvorovskaya str., Moscow, 10702, Russia*

Abstract

The paper describes the mechanism of a bi-modal representation of a situation as its textual description and its representation on a 2.5D model, a convenient tool to model a situation. The model of the world consisting of two parts (linguistic and extralinguistic) makes it possible to effectively use the capabilities of video analytics (including those based on artificial neural networks) and semantic analysis of textual information (using artificial neural networks as well). The co-use of these two parts of the model of the world also enables implementation of effective interaction of such a representation with user while maintaining the level of detail of representations related to the world.

Keywords

Bi-modal model of the world, language model, 2.5D model, video analytics, semantic text analysis, artificial neural networks

1. Introduction

Human manipulates the external and interoceptive world through its modeling. The model of the world in human consciousness is represented by two parts: the linguistic model of the world and the extralinguistic (multimodal) model of the world, both act as a single whole together.

Usually, the extralinguistic part of the world model is used at the lower (signal) levels of processing extralinguistic (for example, visual) information, while the language model of the world, due to the low variability of linguistic constructions, is used at the upper (symbolic) levels. The textual representation of the situation is convenient for its (situation) description in the process of interacting with the user. This form of information presentation is also convenient when the situation analysis involves textual information to represent the constraints of the situation (for example, in the form of instructions).

The corticomorphic implementation of the extralinguistic model of the world engages very complex mechanisms of structural information processing. Therefore, the so-called 2.5D representation of the situation is used as such a model. It is easily interpreted by the user and can serve as a basis for an effective understanding of the situation.

The 2.5D model of the world has another feature that makes it indispensable when used in the interpretation of extralinguistic reality: this model is physically continuous, in contrast to the sensory representation of reality observed by sensors only at some time and space points. This makes it possible to use the model for efficient tracking of real-world objects: on its basis, information obtained discretely at various points in real space and time is interpolated.

Russian Advances in Fuzzy Systems and Soft Computing: Selected Contributions to the 10th International Conference «Integrated Models and Soft Computing in Artificial Intelligence» (IMSC-2021), May 17–20, 2021, Kolomna, Russian Federation
EMAIL: kharlamov@analyst.ru (A. 1); *roman.jarkoi@iss.ru* (A. 2)
ORCID: 0000-0003-2942-5101 (A. 1)

2. Model of the human world

The goal of this work involves the creation of a model of the human world, which combines the detail of the world representation with the ease of use. For this purpose, an attempt was made to reproduce the model of the world, as formed in the human consciousness. Such reproduction, on the one hand, makes it possible to implement effective mechanisms for modeling and manipulating the model, and on the other hand, to use the model as a convenient interface for its interaction with society (and the user as its representative).

The model of the human world includes the linguistic part, which describes the world in some texts of the language, and the extralinguistic model, which represents the world in terms of other sensory modalities that help human perceive the world, primarily the visual modality.

Modeling the human world is a complex challenge [1]. And while the technology of creating language models currently allows the formation of language models more or less adequate to the task of modeling the world [2], the technology of creating extralinguistic models makes it possible to form models that are significantly inferior to linguistic ones in quality due to the complexity of describing the world in non-linguistic categories [3].

The model of the world is divided into two parts, which are not identical in terms of their participation in the formation of the user's representations of the world. The extralinguistic part of the world model forms, as a rule, the lower levels of the situation representation, but in great detail. The language part of the world model forms the upper (generalized) levels of representation.

Thus, the extralinguistic model contains all the necessary information about the situation, while the linguistic model provides an acceptable minimum, omitting the details known to the society (user), the so-called fundamental (or basic) knowledge.

2.1. 2.5D world model as an extralinguistic part of the world model

In this case, the so-called 2.5 models [4] provide interesting possibilities in terms of modeling the world. Though being comparatively simple in implementation, they include physical properties of the real world, thereby making up for the lack of taking these properties into account in language models (due to the absence of so-called basic knowledge in the texts).

And although such models differ in their mechanisms from the representation of the world in human consciousness, they easily fit as a part of the model of the world, forming common representations with the linguistic model.

2.2. The textual representation of the world is the linguistic part of the world model

The linguistic part of the world model is formed using well-known mechanisms for the formation of semantic networks [3], which makes it possible not only to visualize the situation, but also to manipulate situations, including their comparison with each other, and thus to cluster and classify them.

This representation of situations help describe the behavior of the models in order to interpret it to the user, but also to introduce restrictions into the models, presented, for example, in normative acts, which are also texts.

3. Modeling in AnyLogic

The 2.5D model can be obtained in a number of ways, for example, using the AnyLogic toolkit [5]. In this case, the model takes into account the geometry of space and some physical laws, and the special software written within its context makes it possible to naturally represent the behavior of the agent.

The agent's behavior on the model is synchronized with its behavior in the real world by means of sensors, i.e. video cameras located at various points in real space (see Figure 1).

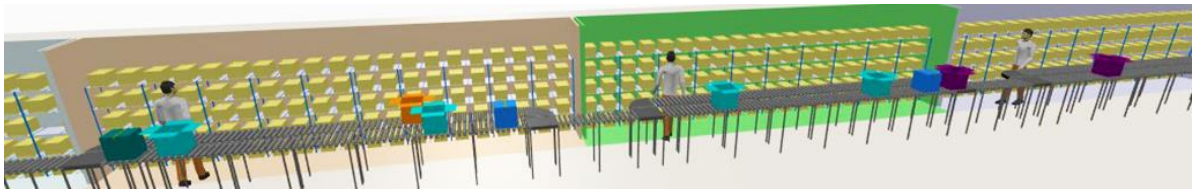


Figure 1: An example of agent behavior representation at 2.5D model

The AnyLogic software is used to build a 2.5D model of an agent moving in a given direction at a given speed, taking into account the geometry of the floor plan and the movement of other agents. The agent model allows predicting its behavior at any time by extrapolating its parameters in time within a given geometry of space constraints.

The behavioral sensorics of a real agent is discrete in time: the parameters of a real agent appear at the input of the model only at some moments of time at some points in space. At these moments in time, the real agent is verified by its synchronization with its model equivalent. Thus, the agent is effectively tracked.

Along with the physical parameters of the agent, some characteristics of the object states are evaluated, for example, the agent posture.

The agent posture turns out to be not only an effective means for describing a situation, but also a convenient tool for the subsequent classification of a situation. Thus, two situations that are indistinguishable by simple classification (for example, using convolutional artificial neural networks (see Figure 2) easily fall into two different classes of situations when using the posture categories (see Figure 3).

Figure 2: Two situations that are indistinguishable by simple classification by convolutional artificial neural network

Figure 3: Two different postures – two different situations

4. Modeling in Description of the situation by the sequence of template sentences from the “to” part of the production rules

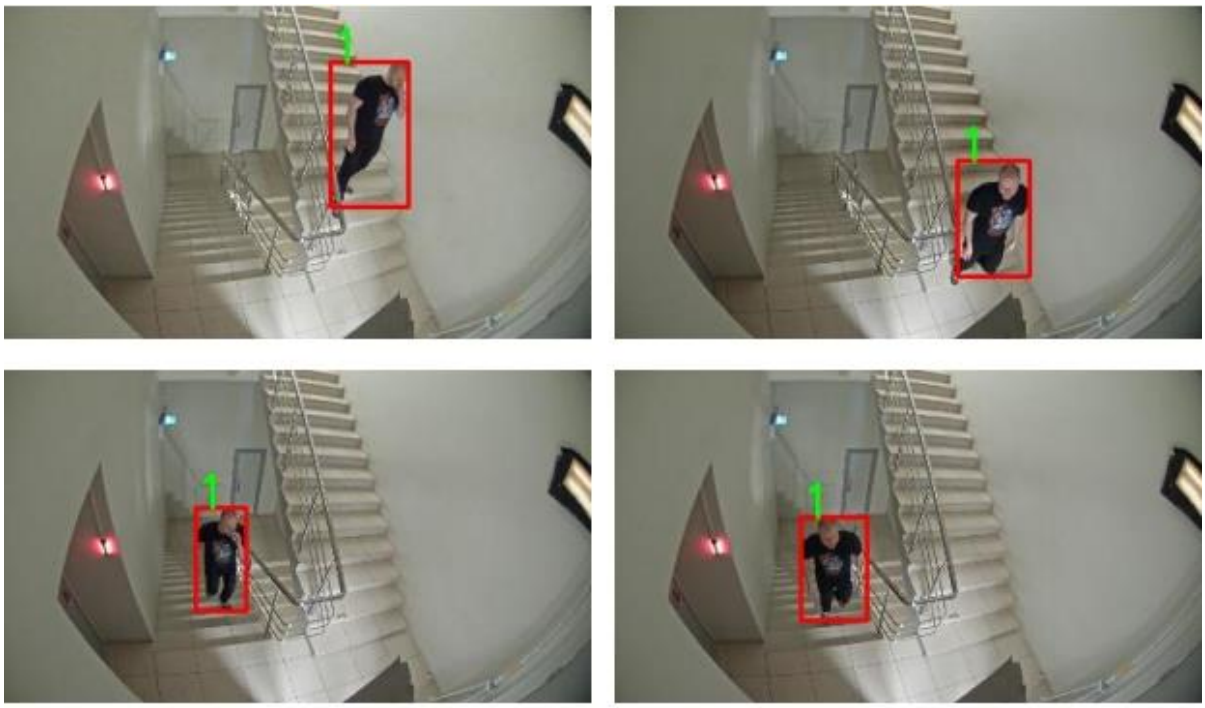
The behavior of the agent at any moment of time correlates with the behavior of the model, not only taking into account the parameters presented above, but also taking into account the agent posture at this moment of time. The use of the corresponding production rule from the knowledge base makes it possible to issue a corresponding natural language sentence describing the situation, provided that a certain posture of the agent appears on the video frame.

For example, if the agent is smoking, and the posture “agent is smoking” was recognized, the system provides the sentence “The agent is smoking”. If the agent is speaking on the phone, the sentence “The agent is speaking on the phone” is issued.

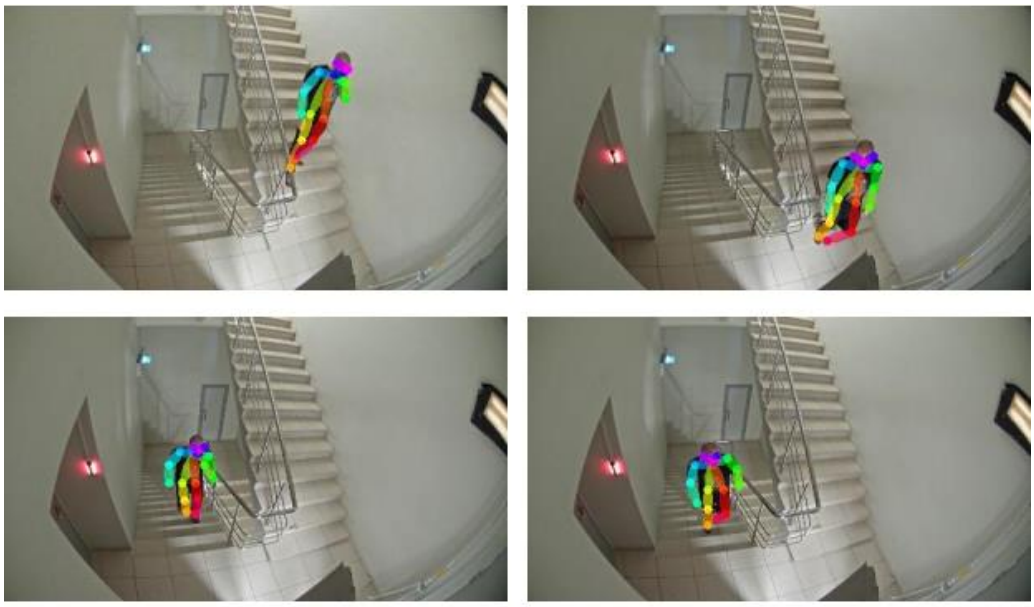
5. Text description of a sequence of simple situations as a behavior scenario

A situation, as a rule, is not only a combination of several objects in space, but also some development of this configuration in time. For example, a sequence of frames is presented in which a person’s hand is in one position for some time. In another sequence of frames, the positions of the hand at the mouth and not at the mouth alternate. One can consider other options for situation scenarios.

Situations appearing on a sequence of video frames divide it into separate scenes, which are identified by situation templates. The names of these templates in the dynamics of the video sequence are in fact a scenario of the agent’s behavior. An example of such a scenario for a video sequence looks like this: “Smoking-norm-smoking-norm” (see Figure 3 b)). The time on the corresponding video frames corresponds to the real time of the scenario development.



a)



b)

Figure 3: An analyzed sequenced of a video a). “Smoking-norm-smoking-norm” is a scenario for the video sequence b).

Two situations that are indistinguishable by simple classification easily fall into two different classes of situations when using scenario in time (see Figure 4).

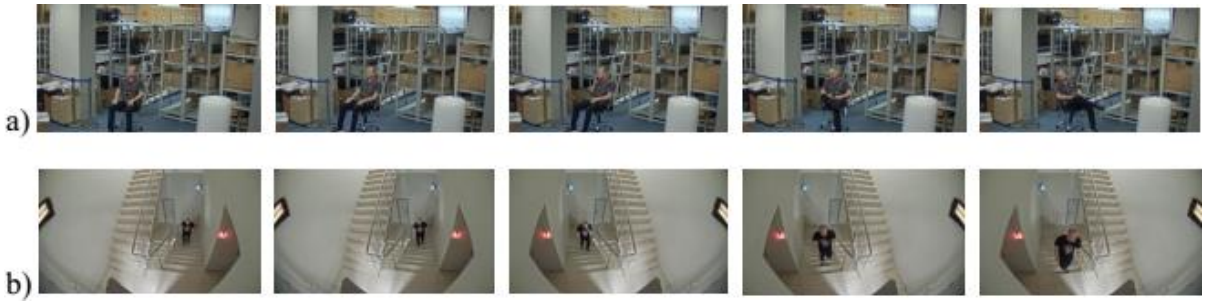


Figure 4: “Norm-norm-phoning-norm-norm” is a scenario for a video sequence a). “Norm-norm-smoking-norm-norm” is a scenario for a video sequence b).

6. Representation of the situation by the semantic network of the text describing the agent’s behavior scenario

A scenario as text is the basis for building a semantic network, which is used to recognize the situation.

Situations represented by homogeneous semantic networks are easily classified by comparing them with semantic networks of classes [4].

7. Description of the situation dynamics by the sequence of semantic networks

If it is necessary to describe a dynamically developing situation, it is represented not by one semantic network, but by their tuples. Identically named vertices of the semantic networks of a tuple are connected with each other and characterize the dynamics of objects in the situation: their appearance, existence in time and disappearance. This introduces a change in the description of the situation as its dynamics in time.

8. Conclusion

A bimodal representation of a situation – in the form of its textual description and in the form of its representation on a 2.5D model – is a convenient tool for modeling a situation. Therefore, the model of the world, consisting of two parts – linguistic and extralinguistic – makes it possible to use the capabilities of video analytics (including those based on artificial neural networks) and semantic analysis of textual information (using artificial neural networks as well). The joint use of these two parts of the world model also makes it possible to implement effective interaction of such a representation with the user while maintaining the detail of the world representations.

9. References

[1] A. Kharlamov & M. Pilgun (Ed.), Neuroinformatics and Semantic Representations. Theory and Applications., Cambridge Scholars Publishing, Tyne upon Evon, 2020.

[2] A.A. Kharlamov Assotsiativnaya pamyat' — sreda dlya formirovaniya prostranstva znaniy. Ot biologii k prilozheniyam, Palmarium [Associative memory – an environment for the formation of a knowledge space. From biology to applications.], Academic Publishing, Dyussel'dorf, 2017.

[3] A.I. Panov Odnovremennoe obuchenie i planirovanie v kognitivnoy robototekhnike. – Simultaneous learning and planning in cognitive robotics, 2001 URL: http://www.raai.org/news/pii/ppt/2021/pii2101_Panov.pdf.

[4] A.A. Kharlamov Neyrosetevoy podkhod k raspoznavaniyu situatsii po tekstu [A neural network approach to recognizing a situation by text], Speech technology 1 (2019); 22-29.

[5] Anastasiia Zhiliaeva Network port: Learning to use the Material Handling Library (Part 5), 2001. URL: https://www.anylogic.com/blog/network-port-learning-to-use-the-material-handling-library-part-5/