GRAPHIC STORYTELLING, STORYBOARDING AND VIDEO RECORDINGS OF SURVEILLANCE CAMERAS

Leon Rothkrantz, Xiaoan Wang

Czech Technical University in Prague
e-mails: ljm.rothkrantz@gmail.com
Czech Republic

Abstract: A crisis, terroristic attack or disaster may be represented by a sequence of pictures visualizing events and actions. It is possible to sample frames from real video recordings. But not every sequence of frames tells a story. Prototypes of disasters are represented, remembered and recognized as scripts in the shared memory of humans. To test this hypothesis many aggressive scenes were played by actors and recorded. Storyboards are used to analyze the scenes. Experiments and results of analysis are presented in this paper.

Key words: Video surveillance, storyboards, aggression detection.

1. INTRODUCTION

In [1] Rothkrantz reports about projects on the development of surveillance systems with multimodal cameras in trains. The first prototypes with a focus on detection of aggressive behavior around trains have been developed and tested [1, 2, 3, 4, 5, 6, 7, 8, 9]. Such systems can be used as an alerting system to detect the onset of inappropriate behavior. To develop such systems it was needed to generate video data of aggressive scenes in a train. Because of privacy aspects no real life recordings were available. Actors were requested to play different scenes described in verbal scripts/storyboards. Operators of surveillance systems were requested to make a selection of the most prominent aggressive scenes in and around trains and to describe these scenes as a story. Every story is based on one or more ideas. Telling stories is an age-old method used to communicate ideas, recreate and preserve culture, memories and traditions.

Human observers are in many situations unable to describe events in an objective way. People have problems to handle situations which they cannot understand. They have a strong tendency to come up with possible explanations to reduce the unsure, possible threatening situations. This psychological process is described by Festinger in his Cognitive Dissonance theory [10]. According to

Abelson and Shank [11] people's knowledge of experienced events can be described by the concept of "scripts".

This is in line with the vision of Klein [12]. Human operators typically fuse all their observations and generate a hypothesis using their experience if they are confronted with relevant patterns. They basically match the situations to the pattern they have learned and experienced from past events and training. If they find a clear match, they will carry out the most typical interpretation of the situations. Human operators are able to share their observation with other operators because they have a shared world model.

Operators are able to describe different scenes as a story. The question is how to animate such a story, to use symbols and movement to sketch the different frames in a scene to effectively convey the story. This is what is called storyboards. A storyboard is a series of sketches, drawings or photographs used to plan and prepare for filming (one may consider storyboards as virtual shoots in this sense). In this paper we will validate the following scientific hypothesis:

- Is it possible to describe aggressive scenes in trains as storyboards?
- What is the general structure of a storyboard?
- Is it possible to associate with every string of video frames a story?
- Is it possible to reduce a video recording of an aggressive scene in a train to a storyboard?

In the next section we will describe related work. Then we discuss storyboards in more details. After that we show how a storyboard can be used as a script for video recording and how storyboards can generate similar stories. Then we discuss the question if the frames of video recordings can be summarized to storyboards. We present many storyboards about aggressive scenes in trains. In the section experiments we validate our scientific hypothesis and end this paper with a conclusion and list of references.

2. RELATED WORK

As early as 1977, Schank and Abelson [11] invented the concept of "script"-using scripts to describe people's knowledge of experienced events. Their theory has contributed a lot to the study of natural language and cognitive science, and it is the most important theoretical background of this thesis work as well. Well known is the restaurant script designed by Schank, as displayed in Figure 1. People who want to have a diner in a restaurant show a sequence of typical behavior. It starts with entering the restaurant and asking the waiter for a free table. Next the waiter asks for some drinks and brings the menu card. After ordering the menu the menu will be served in a fixed order first the appetizer, or starter, then the main menu and a desert at the end. Next the waiter asks if coffee is needed and finally brings the bill. Payment is done by cash or credit card and the guests leave the restaurant. It proves that people have similar experiences with having a diner in a restaurant and are able to report a list of actions.

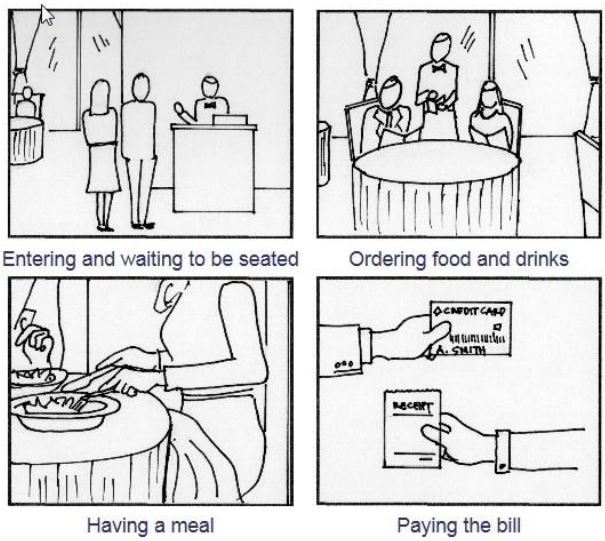

Fig. 1. Restaurant script

According to Schank and Abelson's theory "understanding is a process by which people match what they see and hear to pre-stored groupings of actions that they have already experienced. New information is understood in terms of old information." In order to understand the actions that are going on in a given situation, a person must have been that situation before. That is, understanding is knowledge-based. Scripts appear as a set of familiar scenarios in our memory and they are stored as a pattern of actions that we have been previously experienced. A script is also a standard event sequence, if a story is referenced to those frequently occurring scripts, then the story will be quite understandable to every person.

The observation process of human operators and automated systems are completely different according to Klein [12]. In expert systems many hypothesizes are considered simultaneously and after a process of reasoning the most probable is selected. According to Klein if observers receive observations a hypothesis pops up in their mind immediately. Further, he claimed that human operators are unable or not willing to consider alternatives simultaneously. They stick to the first hypothesis and give it up only if many contradictory data becomes available which supports an alternative. On the other hand, human's observations are context sensitive [1, 4, 8]. They are based on multimodal input in a given context and may be affected by emotional states due to the intense nature of the observed events.

In 2009 we published our first papers on aggression detection on railway stations [7, 8]. A multi-modal aggression detection system was built that fuses audio and video data from sensors located in a train compartment. The aggression detection system is based on many hours of observing and studying professional operators at work as they analyze and respond on surveillance data. Using classifiers, models were trained which can be used to make prediction of stress or

aggression level on new data samples. Because of privacy aspects we were not allowed to use real life video recording. We decided to make video recordings of scenes played by actors in a mock-up of a train and train station environment. Storyboards were used as a script for the video recordings.

3. MULTIMODAL VIDEO RECORDINGS OF AGGRESSION IN TRAINS

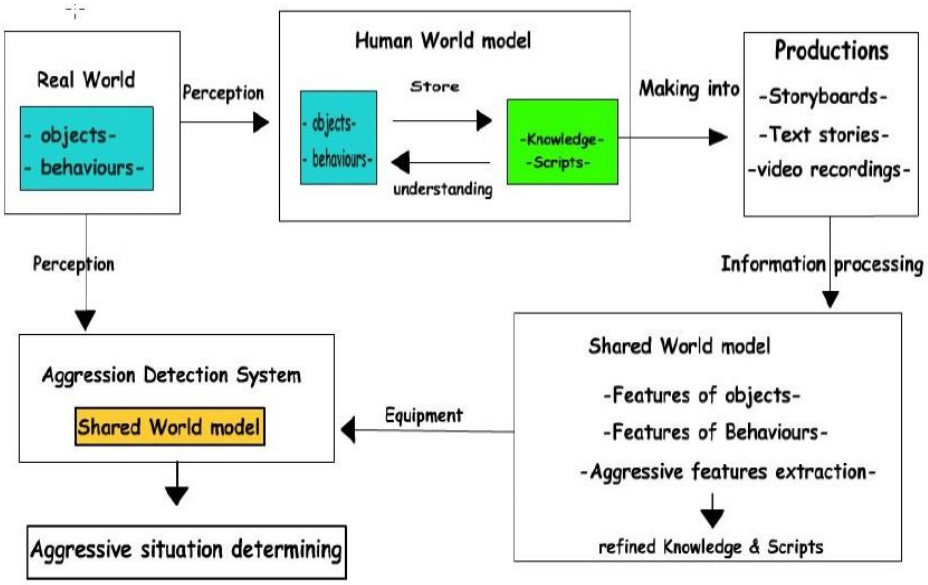

Fig. 2. Architecture of video recordings of aggression in trains

The first author of this paper was the leader of a project on automatic multimodal assessment of aggression in train. In an experiment, multimodal cameras were installed in train compartments (see Fig 3). Stand-up comedians played several aggressive scenes according to predefined scripts (see Figure 3).

Fig. 3. Picture of a train compartment and recorded cred aggressive scene

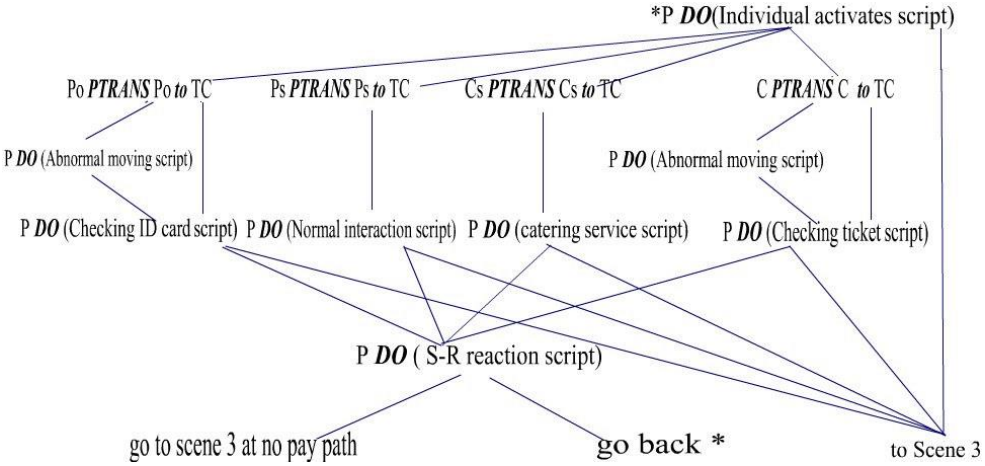

Fig.4. Annotation of recorded scenes using scripting language defined by R.Schank.

The video footage of recorded scenes was analyzed. First the recorded scenes were annotated using the scripting language as defined by R. Schank (see Figure 3). This enables automated processing of the video footage of recorded scenes.

4. DESIGN OF STOREYBOARDS

Fig. 5. Rory's Story Cubes game (www.storycubes.com)

The story cube game is composed of 9 dices. On every face of a dice there are 6 possible pictures. The dices are ordered in a box (see Figure 5). To start the game the dices are mixed in the box or thrown on a table. The six upper faces of the nine dices are supposed to tell a story. Every player writes down his story. In an experiment the story cube game was played by 9 Computer Science students.

Using their phantasy the students were able to tell stories, but the generated stories were loosely connected to each other.

5. DESIGN OF STOREYBOARDS

There is a well-known saying about the relation between language and vision: "A picture tells more than thousands of words". In this section we discuss a special scenario of aggression in and around a train. The underlying story should stimulate the phantasy of a designer, and brings an experienced scene from the past into his mind and enables the design of a corresponding storyboard. Storyboards are visual organizers, typically a series of illustraDons displayed in a sequence for the purpose of pre-visualizing a video.

A storyboard is composed of a sequence of frames corresponding to key actions. A frame is not an exact copy of the real world action. Only essential features are displayed. Usual a frame is a redesign of a video frame or just a drawing from phantasy if no video recording or pictures are available. Another option is to use augmented reality technology to annotate frames. The multimodal aspects of the real world can be taken into account by annotations of the frames, indications of movements, sound etc. It is also possible to add text balloons. In more detail a storyboard has the following characteristic structure:

- The first scene displays the context and the problem. Movements can be displayed by arrows, special behavior (running person) or a track by dotted lines.
- Every single frame is expressive and easily recognized and tells a part of the story.
- The whole story can be understood from the sequence of frames because a prototype script belongs to our shared memory.
- A new scene is generated by the onset of characteristic people (conductor, police) or representation of special behavior or actions.
- A scene can be displayed of high arousal for example by evident aggressive behavior against employee, travelers, or compartment furniture (fighting, penetration personal space, taking unusual positions on unusual places).
- The last scene displays the plot of the story.

6. STOREYBOARDS AND VIDEORECORDINGS

Operators from the control rooms of the Dutch Railway Company were interviewed about the most frequent aggressive/inappropriate behavior incidents and scenes in and around trains. It resulted in a list of 15 incidents. The most prominent incidents are aggression against personal/people, noise pollution, harassment and pickpocketing. For all the incidents a short story was written in a few lines. These stories were used by semi-professional actors to play the scenes in a train in a mock-up of a railway station. The actors were stand-up comedians used

to ply scenes in a spontaneous way. The scenes were recorded using a video camera and sound recordings.

The recorded video footage was annotated and analyzed. One of the researchers sketched a story board for each of them. In the next Figure 6 we display some storyboards. All the scenes were played two times by different group of actors without viewing to each other's recording. There was a remarkable similarity in the recorded stories. This is a strong indication for the fact that the selected scenes correspond with scripts stored in the memory of the actors. In the paper we will research the similarity of scripts and storyboards and if people tell the same story after viewing each of the storyboards.

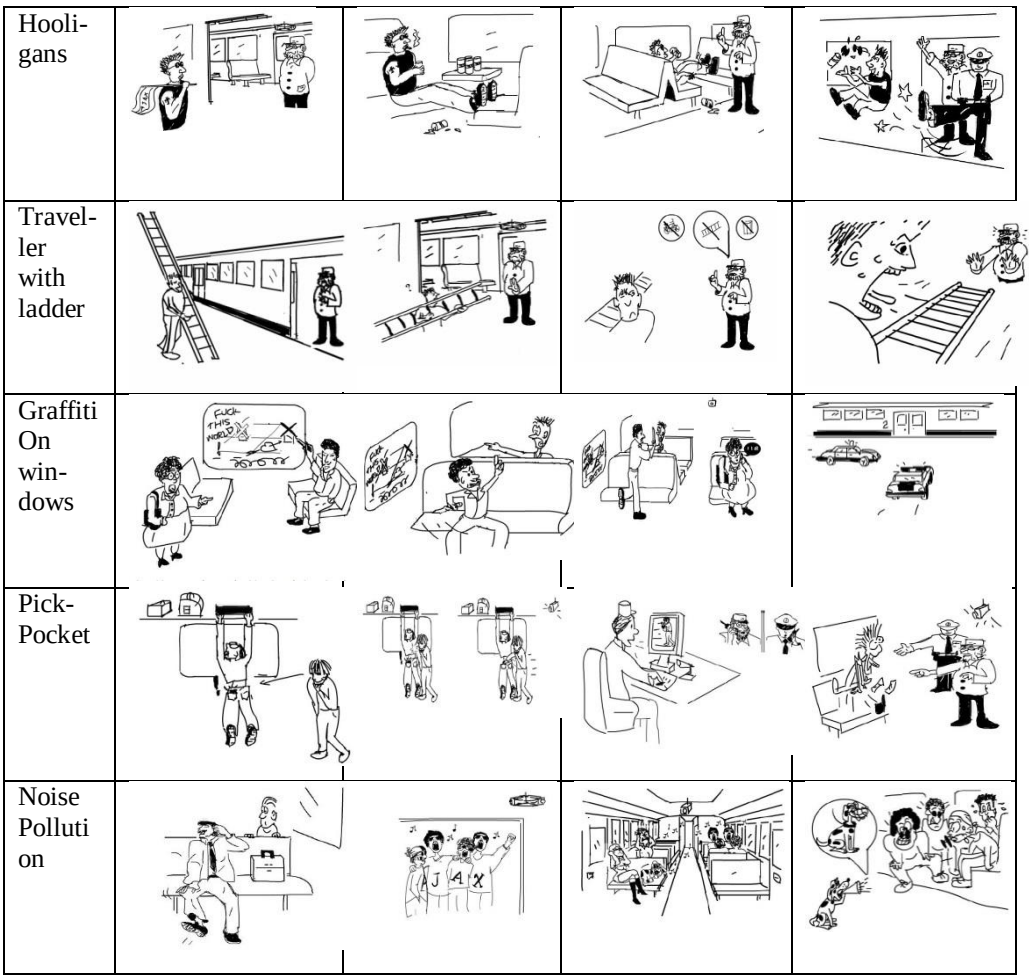

Fig. 6. Short display of storyboards.

7. VIDEORECORDINGS AND GENERATED STORIES

Designing storyboards is a bottom up process. We design isolated scenes. But these scenes are merged together by the underlying story. Human observers fill in the gap using their imagination. The art of designing storyboards is of course to design the isolated scenes in such a way that they trigger the imagination of observers to generate the same story. The question is if every designed storyboard tells the same story. We asked 25 students from a master course on multimedia to read 6 of the randomly assigned storyboards and to write down the corresponding story. Every story was annotated by 10 students. In Figure 6 we display some of the 10 stories by some frames. Most of the students gave a similar semantic interpretation of the displayed frame of the story but used different words to tell the story and focused on different details. But all understood the right plot. It proves that the generated storyboards trigger similar stories; probably the 15 scripts belong to the shared memory of all students.

To compute the similarity between different stories we used the following procedure. We take the 150 stories, 10 stories for every storyboard. We processed the text as follows:

Removal of stop words, stop words are words with low information and a minimal contribution to the content of the story such as “and”, “or”, “the”.

Stemming, reduction of every word to its root or removal of suffix (walks, walked, walking reduced to walk (we used the Webster dictionary and category database CELEX)

Removal of synonyms (WordNet has been used to detect synonyms)

All the remaining words W, are ordered in alphabetic order as a vector (W1,….Wn). We take the remaining words of the 6 versions of a story Si (i=1…15) and write Si vector Si1,…Sin. The entries Sij are natural numbers indicting the frequency of the word wj in the set Si. Finally we compute the correlation ρ(Si,Sj) between the vectors Sj. As expected the correlation ρ(Si,Si) was significant higher than ρ(Si,Sj).

Table 1. Example of the CELEX database

17194	computer	17198	computerizing
17195	computerize	17199	computers
17196	computerized	17200	computes
17197	computerizes	17201	computing

8. STORYBOARDS AND VIDEORECORDINGS

Fig. 7. Relation storyboards and video recordings.

In total we had 15 storyboards, stories and corresponding video recordings. The storyboards and corresponding stories are used as guidelines for the actors. Only a global summary of the stories were presented to the actors. The actors were split up in two groups and every group has to play all the scenes. There was a remarkable correspondence between the storyboards and the corresponding video recordings and the parallel recordings. We were able to find a frame in the video recording corresponding with a frame in the storyboard (see Fig 8)

Fig. 8. Correspondence storyboards and video recording.

9. STORYBOARDS AS SAMPLED VIDEORECORDINGS

A movie can contain thousands of frames. From the view of information convey, that is too much and there is a lot of redundancy and similar information. The question is how to sample key frames from video recordings for a relatively complete storytelling? The key frames may be composed as a storyboard. We will use a video segment experiment to show what are feasible key frames. They can be selected at a fixed rate or at random or by human selection.

Table 2: Video recording.

	Sampling	**Key-frame sequence**
File Name: [Fare dodger]	every 7 frames	482
File Size: [21MB]	every 15 frames	224
Resolution: [352] x [288]	every 30 frames	112
Duration: [00:03:44]	every 60 frames	56
Totally Frames: [3374]	every 120 frames	28
Default frame rate:	every 240 frames	14
[15frames/second]	every 480 frames	7
	every 960 frames	3

9.1. Fixed rate

First we selected a set of fixed rate fps (frames per second). After analysis of those selected frames, some interesting result has been found. If we choose the sampling rate of 1/7frames, 1/15frames and 1/30frames will have too much redundant information. If we look to 1/240frames, 1/480frames and 1/960frames important information was lost. Choosing 1/60frames and 1/120frames we got better results. Unfortunately the string of selected frames could tell a different story. If we split up the video in different topics, it proves that the different topics contain different amounts of frames. In the current movie a lot of frames describe the interaction between the traveler and conductor (around ¾ of total duration) and are very similar. A process of deletion and insertion of frames is needed to get the right story.

9.2. Random sampling

In second step, we used random sampling. It proves that the sequence of frames usually don't tell a story or a complete different story. In a storyboard the underlying script in our mind connects the key frames with key scenes. Just a sequence of frames has no underlying script in general. Storytellers with a lot of phantasy are able to generate a story, but such a story is too exceptional and not shared by other observers.

9.3. Sampling after segmentation

We segmented a video in successive scenes. We have to remember that our camera position is fixed. A new scene starts when for example a new person enters the scene drastic change in arousal, sudden movements, changing positions, shouting etc. We were able to segment a video automatically. Within a scene a frame was selected randomly. It appears that by random selection we missed a key frame and selected a frame carrying less information. Take for example the interaction between the conductor and fare dodger. This interaction takes some

minutes and the key-frames are the frames where the fare dodger starts to fight with the conductor. Most of the other frames are meaningless.

9.4. Personal selection

It was possible to select frames in such a way that they tell a story and compose a storyboard. First the video has to split up in topic and then within every topic a representative one or more representative frames have to be selected. But compared to the designed storyboard, selected frames of video recordings, usually miss important characteristics. In designing a storyboard the designer is able to fuse successive frames and add movements to it.

10. CONCLUSIONS AND FUTURE WORK

Storyboards are commonly used in the movie industry to design movies. In this paper we used storyboards to record aggressive incidents in trains played by actors. Operators in the control room of surveillance cameras usually monitor recordings. They were interviewed about the most aggressive incidents at railway stations. From those stories it was possible to design storyboards to be used as movie script. In this paper we studied the relation between storyboards, underlying stories and corresponding video recordings. It proves that for prototype incidents it was possible to design storyboards which trigger similar stories in the mind of observers. Prototype incidents belong to the shared memory of people and people were able to tell similar stories for every storyboard. More research is needed to reveal the designer characteristics of story boards.

REFERENCES

[1] Rothkrantz, L. *Surveillance angels. Neural Network World* 24, 2014, pp. 1-25.

[2] Rothkrantz, L. Smart Surveillance Systems. *Network Topology in Command and Control: Organization, Operation, and Evolution: Organization, Operation, and Evolution*, IGI. 2014, pp. 270-280.

[3] Lefter, I., L. Rothkrantz, G. Burghouts. A Comparative Study on Automatic Audio-Visual Fusion for Aggression Detection Using Meta-Information. *Pattern Recognition Letters*, volume 34 issue 15, 2014.

[4] Lefter, I. G. Burghouts, L. Rothkrantz. Automatic audio-visual fusion for aggression detection using meta-information. *Proc. Advanced Video and Signal-Based Surveillance (AVSS)*, 2012, pp. 19–24.

[5] Hameete, P., S. Leysen, T. Van der Laan, I. Lefter, L. Rothkrantz. Intelligent Multi-Camera Video Surveillance. *International Journal on Information Technologies* & Security 4 (4), 2012.

[6] Lefter, I., L. Rothkrantz, G. Burghouts, Z. Yang, P. Wiggers. Addressing multimodality in overt aggression detection. *International Conference on Text, Speech and Dialogue*, 2011, pp. 25-32.

[7] Yang, Z., S Fitrianie, D. Datcu, L. Rothkrantz. An aggression detection system for the train compartment. *Advances in artificial intelligence for privacy protection and security*, 2011, pp. 249-286.

[8] Yang, Z., L. Rothkrantz. Automatic aggression detection inside trains. *IEEE International conference on Systems Man and Cybernetics.* 2010, pp. 2364-2372

[9] Datcu, D., Z Yang, L. Rothkrantz. Multimodal workbench for automatic surveillance applications. *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, vol 0, pp 1-2.

[10] Festinger, L. A theory of cognitive dissonance. Stanford University Press, 1962.

[11] Schank, R., R. Abelson. *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*. Hillsdale, New Jersey, 1977.

[12] Klein, G. (1988*). Sources of power: How people make decisions.* Cambridge, MA: MIT Press, 1988.

[13] Dorner, R., P. Grimm, D.F. Abawi. Synergies between interactive training simulations and digital storytelling: A component-based framework. *Computers & Graphics*, 26, 2002, pp.45-55.

[14] Miller, E. (2009) *Digital storytelling. Master of Arts*, University of Northern Iowa.

[15] Digital storytelling, resources, instructions and guidelines in the practice of digital storytelling, retrieved from http://libguides.mercy.edu/digitalstorytelling.

Information about the author:

Leon Rothkrantz was appointed as Associate Professor at Delft University of Technology and as a full professor at The Netherlands Defence Academy. At this moment he is visiting Professor at Czech Technical University in Prague. His main research topic is Artificial Intelligence.

Manuscript received on 04 October 2018