Hybrid imaging-AI approach for handling critical situations in a fast-changing environment: preliminary study

Adam Surówka[1, *]

[1]Cracow University of Technology, Warszawska 24, 31-155 Cracow, Poland

Abstract. The purpose of this study is to explore the possibility of using selected imaging technologies in automated video surveillance systems. The main goal of this project is to handle events that may lead to security risks, injuries, etc in various environments without relaying on more conventional sensors such as infrared photocells. For this purpose it is necessary to perform a thorough analysis of the events to be interpreted as situations of interest. It is also important to consider the hardware requirements and restrictions for developing such system. The project requires defining a hardware as well as software platform(s) and their integration into an automated tool. This paper describes the implementation of the famous Microsoft Kinect 2.0 depth sensor (well known in gaming and recreational applications) for shape/skeleton detection, and its integration into an artificial intelligence based platform utilizing selected machine learning methods. The author reveals the system implementation details, and then demonstrates its shape detection capabilities while in operation.

1 Introduction

The use of vision techniques in surveillance systems of industrial facilities and of public use began in the 1940s. These systems were called Closed-Circuit Television (CCTV) [1, 2]. Attempts to develop effective recognition of human activities and behavior based on the image from CCTV monitoring have been ongoing since 1980. Human activity recognition (HAR) has revolutionized the area of computer vision research in a wide spectrum of applications [3]. Systems based on HAR enable among others the implementation of tasks related to recognizing life threatening situations [4], preventing crime and vandalism [5, 6], supervision of the sick and elderly [7], biometric face identification [8-11] and analysis and classification of all forms of human activity that may be of interest in a given situation [12-17]. To achieve full efficiency, it is required to develop optimal decision algorithms. Decision models are most often based on machine learning techniques such as artificial neural networks, fuzzy logic and classifiers [18-21]. Figure 1 shows the hierarchical breakdown of methodologies found in HAR systems.

* Corresponding author: asurowka6@gmail.com

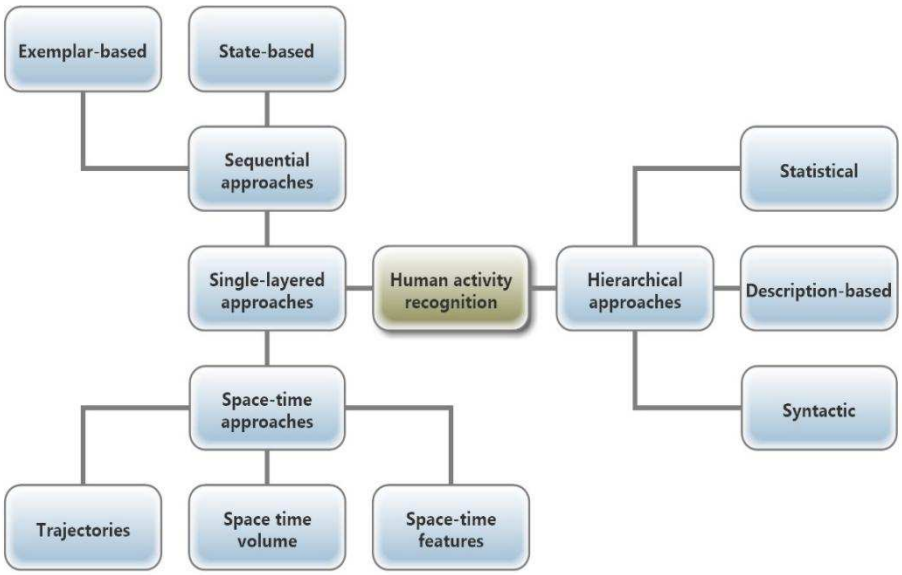

Fig. 1. Hierarchical taxonomy of HAR methodologies.

The synergistic combination of HAR vision techniques and advanced decision models enables the development of optimal and efficient automated video surveillance systems (AVSS). A characteristic feature of AVSS systems is the ability to autonomously detect and report situations that have been saved in the device's memory and are interpreted by them as critical events. This activity significantly improves supervision and enables much faster identification of potential emergency situations [22].

2 Research methodology

The main purpose of this project is to identify events that can lead to safety hazards, injuries, etc. Preliminary research will focus on developing an automated video surveillance system to report potential emergencies. The study uses the results of the author's previous research in the field of in-depth analysis of events, which are interpreted as critical situations. Databases and artificial neural networks specially developed for this purpose were also used. They can be located in previous publications of the author. To achieve this goal, it was decided to use the implementation of the famous Microsoft Kinect 2.0 depth sensor, well known in games and recreational applications. The legitimacy of this choice boils down to the benefits offered by accessing the built-in functionality to detect people's shapes / skeletons. Another aspect behind choosing a Kinect device is the ability to create software in the popular environment: Matlab. In addition, the environment offers a wide spectrum of tools enabling the use of various machine learning methods in the programs code. Thanks to this, creating decision models will be easy to implement.

3 Proposed AVSS structure

The proposed system is based on three main assumptions: avoiding data redundancy, developing health and life risk patterns and implementing machine learning in decision algorithms.

3.1 Microsoft Kinect 2.0 depth sensor

The Kinect 2.0 device can track up to six different people within the depth sensor's field of view. This field is 0.5 - 4.5 meters from the lens, while the highest accuracy is obtained in the range of 0.8 - 3.5 meters from the device. The horizontal angle of view of the sensor is 70 degrees, vertical 60 degrees. In order to avoid data redundancy the proposed AVSS system will be using these sensor capabilities [23, 24]:

- recording a color image from an RGB camera with a resolution of 1920 x 1080 pixels (based on the colorFrameData matrix),
- recognising the number of people being tracked in the field of view of the sensor (based on the IsBodyTracked matrix),
- mapping twenty-five characteristic points each of the six possible skeletons, whose two-dimensional coordinates are transposed to the coordinates of the color image, enabling graphical determination of the sketched outlines (based on the ColorJointIndices matrix),
- mapping twenty-five characteristic points of each of the six possible skeletons, whose three-dimensional coordinates allow not only to locate a person in the room, but also to determine the correlation of individual body parts (based on the JointPositions matrix),

3.2 Analysis of a critical situation

Following the example of human cognitive abilities, a set of behavior patterns has been developed that can largely suggest the occurrence of expected critical situations. In order to achieve the highest efficiency of the system, it was decided to develop three main groups of human behavior. The first of them was named *Attacker*. This group includes posture patterns that indicate aggressive behavior. Among them were characteristic models of a man who performs a kick, punch, holds a dangerous tool such as a firearm, knife, club, etc. The second group was called *Victim*. Distinguished here are the characteristic postures of a man experiencing pain in the chest and stomach area, who is lying on the floor, standing with his arms raised, kneeling, etc. The third group was called *Normal*. Postures of people standing and sitting are presented here, whose behavior does not arouse suspicion. It was decided that critical situations should be identified in two cases: when identifying the behavioral pattern of at least one aggressor or at least one victim in the analyzed video image. This is due to the multitude of reasons that may initiate situations of threat to human health or life.

3.3 Software implementation of the algorithm

The decision algorithm used in this project was developed using an artificial neural network. To determine the network architecture, it was decided to use the environment tool: Pattern Recognition app. The database consists of 9,500 samples obtained from the depth sensor, which were collected during several measurement sessions. The use of the Pattern Recognition app toolbox enabled the generation of a trained neural network function whose code was implemented in the main AVSS program. Fig. 2 shows the workflow of the designed system. In the first phase, the Kinect device samples the signals from the depth sensor and RGB camera in real time. In the initial processing of images, detection of people, tracking of postures and mapping characteristic points of the skeletons for all tracked bodies is carried out. The set of information obtained this way is subjected to the data extraction process. This process enables the development of data matrices that are compatible with the arguments adopted by decision-making algorithms. Detection of body posture and interpretation of behavior is carried out using the function of an artificial neural network specially trained for this purpose. The emergency procedure is activated when a

potential threat to health or life is detected. The system user is immediately informed of this by a graphic message displayed in the event preview window.

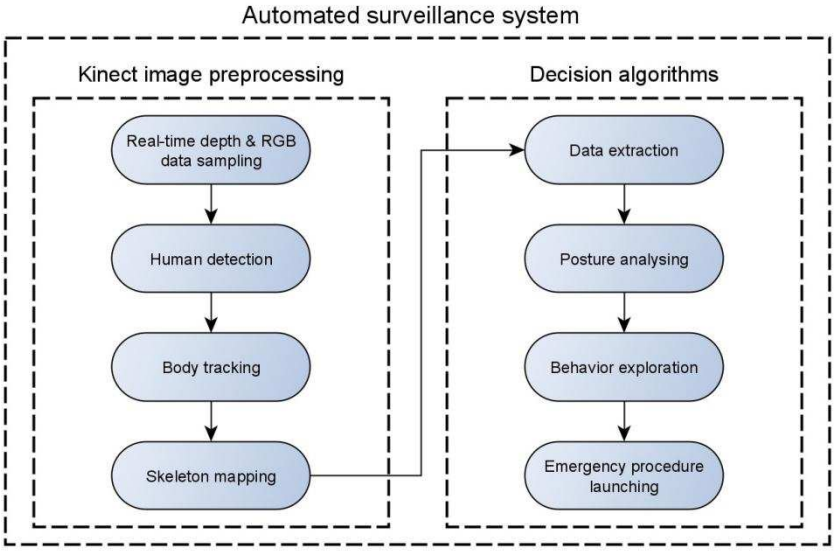

Fig. 2. Automated video surveillance systems workflow steps.

The input layer of the neural network consists of seventy-five neurons. The cause of that is data matrix from Kinect, which provide 25 characteristic points of the skeleton in a three-dimensional coordinate system. The number of thirty hidden layers was set by an experimental method. The number of neurons in the output layer is obligatory to be ten. At the output we get 1x10 array, which shows probability of occurrence each of the classified features (10 different body positions). The process of analyzing the emergency situation as well as creating the database and machine learning were discussed in detail in previous papers of the author.

4 Test of the developed AVSS system

4.1 Parameters

Tests of the AVSS system were carried out in a room measuring 5.5m x 4.7m x 2.65m. Kinect 2.0 has been placed in the center of a smaller wall. According to the manufacturer's recommendations, the sensor was installed at a height of about 1.6m from the floor level. The main parameters of the computer used to perform the measurements: AMD FX6 processor, 16GB RAM memory, 500GB HDD disk, Windows 10 Professional 64-bit operating system, Matlab R2019a programming environment. A group of men and women with various anthropometric characteristics was invited to conduct the tests.

4.2 Test process

The program launched in the environment carried out data capture from the Kinect 2.0 device in real time. The software user received an image from the RGB camera with a resolution of 1920x1080 pixels and a refresh rate oscillating around fifteen frames per second. This value depends primarily on the computing power of the host on which the program is running. This was confirmed during later tests on other machines, where the

result was about 23 fps. If a human figure is detected, the programs algorithm applies colored skeletal markers to the image. In addition, each skeleton receives a label that was generated as a result of behavior recognition. To trigger an alarm procedure, it is sufficient for only one of the detected positions to be classified into the *Attacker* or *Victim* group. This is due to the variety of postures suggesting a potential threat. In preliminary studies, it was decided that the alarm procedure should be reported in the form of an alert appearing in the upper left corner of the screen.

Fig. 3. Examples of AVSS test images.

Figure 3 presents four examples of simulated situations. The first presents two people who do not display behaviors interpreted as dangerous. The system algorithm detects two positions: *Sitting* and *OK* from the *Normal* behavior group. The second case presents an attacker kicking a victim. The system algorithm detects two positions: *Kicking* from the *Attacker* behavior group and *Ill* from the *Victim* group. The third case presents a situation in which the attacker points a firearm at the victim. The system algorithm detects two positions: *Pistol* from the *Attacker* group of behaviors and *Hostage* from the *Victim* group. The last case presents a situation in which an attacker holds a dangerous tool near a victim lying on the floor. The system algorithm detects two positions: *Baseball bat* from the *Attacker* behavior group and *Lying* from the *Victim* group. According to the assumptions, case 1 does not trigger the alarm procedure. However, in cases 2, 3 and 4, an alert appears in the upper part of the screen informing about the possibility of a potential threat to health or life.

5 Summary and conclusion

Based on the tests, it was found that the developed automated video surveillance system correctly implements the project assumptions. The algorithm for recognizing and classifying threat situations works in most cases properly, as evidenced by the rapid and logically justified responsiveness noticeable from the user interface. The system presented can be designed to work in small closed rooms, in which the presence of more than six

people at the same time is not expected. This is due to the way the Kinect 2.0 device works, which according to the specification can correctly map up to 6 people in a specific field of view. This is the main limitation of the proposed solution, because this fact is independent of the developed software implementation. The possible solution for this might be using several depth sensors integrated in the structure of zonal monitoring. In order to increase the efficiency of the system, it is planned to enlarge the database with further reference positions in a larger number of variances. In further work on the system, the implementation of an additional algorithm is also considered, which is based on the structure of the convolutional neural network, allowing the recognition of potentialy dangerous tools (firearms, clubs, knives, etc.). The synergistic operation of the posture classifier based on data from the Kinect 2.0 sensor in conjunction with the recognition of threatening tools will significantly increase the identification of potential critical events as well as its accuracy.

References

1. G. F. Shidik, E. Noersasongko, A. Nugraha, P. N. Andono, J. Jumanto, E. J. Kusuma, IEEE Access, **7**, 170457-170473 (2019)
2. A. Jodełka, A. Rosiński, Inżynieria Bezpieczeństwa Obiektów Antropogenicznych, **3-4**, 21-29 (2018) (in Polish)
3. A. B.Sargano, P. Angelov, Z. Habib, Appl. Sci. **7**, 1-37 (2017)
4. A. Costin, *Proceedings of the 6th International Workshop on Trustworthy Embedded Devices*, 45–54 (2016)
5. R. Xu, Y. Guan, and Y. Huang, Multimed. Tools. Appl. **74**, 729–742 (2015)
6. H. M. Moon, S.-H. Chae, D. Moon, Y. Chung, and S. B. Pan, Telecommun. Syst. **52**, 2249–2257 (2013)
7. W. Y. Shieh and J. C. Huang, Med. Eng. Phys, **34**, 954–963 (2012)
8. M. Włodarczyk, D. Kacperski, W. Sankowski, K. Grabowski, Comput. Sci. **20**, 3-25 (2019)
9. F. Ahmed, H. Kabir, Int. J. Ap. Mat. Com-Pol. **28**, 399–409 (2018)
10. T. Marciniak, A. Dąbrowski, Przegląd Elektrotechniczny, **9**, 137-140 (2016) (in Polish)
11. M. Selianinau, Edukacja Techniczna i Informatyczna, **6**, 563–574 (2018) (in Polish)
12. P. Yugendar, K. V. R. Ravishankar, Journal of KONBiN, 46, 5-20 (2018)
13. G. Baliniskite, E. Lavendelis, M. Pudane, Appl. Comput. Sci. **24**, 134–140 (2019)
14. M. Lotfi, S. A. Motamedi, and S. Sharifian, J. Real-Time Image Pr. **16**, 1301–1316 (2019)
15. A. S. Murugan, K. S. Devi, A. Sivaranjani, and P. Srinivasan, Multimed. Tools. Appl. **77**, 23273–23290 (2018)
16. L. Zhao, Z. He, W. Cao, and D. Zhao, IEEE Trans. Circuits Syst. Video Technol. **28**, 1346–1357 (2018)
17. S. Zhang, D. Cheng, Y. Gong, D. Shi, X. Qiu, Y. Xia, and Y. Zhang, Neurocomputing, **283** 120–128, (2018)
18. G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, B. P. Buckles, arXiv preprint arXiv:1501.05964, 1-30 (2015)
19. S. Sani, N. Wiratunga, S. Massie, *CEUR Workshop Proceedings*, 95-103 (2017)

20. S. Abdelhedi, A. Wali, A. M. Alimi, *Proceedings of the Second International Afro-European Conference for Industrial Advancement,* 227-235 (2016)
21. Z. Uddin, J. Kim, KSII T. Internet Inf. **10**, 2767-2790 (2016)
22. M. B. Ayed, M. Abid, Int. J. Adv. Comput. Sci. Appl. **8**, 59-66, (2017)
23. http://download.microsoft.com/download/6/7/6/676611b4-1982-47a4-a42e-4cf84e1095a8/kinecthig.2.0.pdf (2020)
24. https://www.mathworks.com/help/supportpkg/kinectforwindowsruntime/ug/acquire-image-and-body-data-using-kinect-v2.html (2020)