Vision and Crowdsensing Technology for an Optimal Response in Physical-Security

Fernando Enríquez[1], Luis Miguel Soria[1], Juan Antonio Álvarez-García[1(✉)], Fernando Sancho Caparrini[1], Francisco Velasco[1], Oscar Deniz[2], and Noelia Vallez[2]

[1] Universidad de Sevilla, Seville, Spain
{fenros,lsoria,jaalvarez,fsancho,velasco}@us.es
[2] VISILAB, E.T.S.I.I, University of Castilla-La Mancha, 13071 Ciudad Real, Spain
{oscar.deniz,noelia.vallez}@uclm.es

Abstract. Law enforcement agencies and private security companies work to prevent, detect and counteract any threat with the resources they have, including alarms and video surveillance. Even so, there are still terrorist attacks or shootings in schools in which armed people move around a venue exercising violence and generating victims, showing the limitations of current systems. For example, they force security agents to monitor continuously all the images coming from the installed cameras, and potential victims nearby are not aware of the danger until someone triggers a general alarm, which also does not give them information on what to do to protect themselves. In this article we present a project that is being developed to apply the latest technologies in early threat detection and optimal response. The system is based on the automatic processing of video surveillance images to detect weapons and a mobile app that serves both for detection through the analysis of mobile device sensors, and to send users personalised and dynamic indications. The objective is to react in the shortest possible time and minimise the damage suffered.

Keywords: Computer vision · Weapon detection · Crowdsensing

1 Introduction

Every city in the world suffers events of diverse nature that endanger the life of its citizens. Security specialists protect public and private institutions with professionalism, but the great diversity of possible events and the size and structure of the area to monitor make it very difficult to prevent them, and above all, to plan an optimal response for each threat. Unfortunately, the current global alert situation due to the proliferation of terrorist acts, directly oriented towards citizenship, has only emphasised the need to evolve current security systems to deal with threats in the best possible way.

The project presented in this paper, called VICTORY, aims to provide a next-generation security system, more intelligent, agile and effective, that significantly

improves the reaction time and the response to a threat in a predefined area (for example a building) which we will call 'security zone'. Therefore, the main objectives of the system are twofold:

1. Global solution to detect, analyse and classify security threats using new technologies and generating progress in the state of the art, complementing and improving current solutions.
2. Provide quick and personalised information to potential victims within the security zone, as well as to security personnel to facilitate an optimised management of the threat.

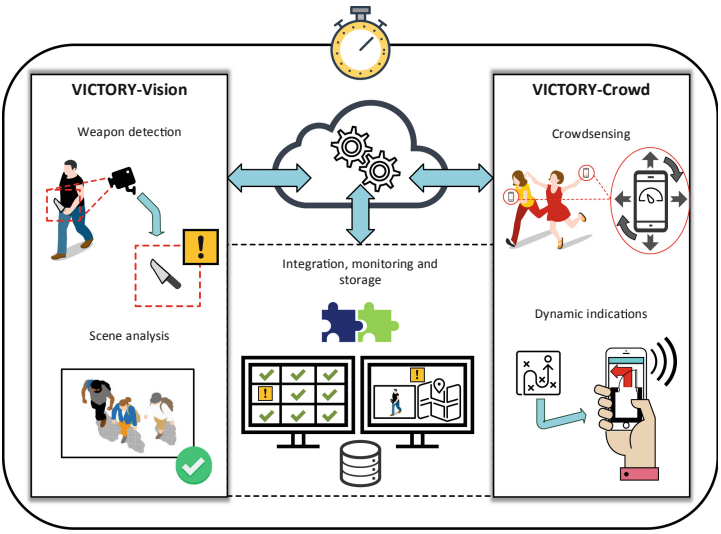

Fig. 1. System components

For this purpose, the design of the system has been divided into several components as can be seen in Fig. 1:

- A computer vision subsystem that automatically analyses the images captured by the security cameras installed in the security zone using Deep Learning techniques. The analysis contemplates two different and complementary approaches that will provide robustness to the system. On the one hand, the generation of detectors specific to relevant patterns (such as different types of weapons), and on the other hand the use of autoencoders that alert when anomalous patterns deviate from the usual scenes captured by the cameras.
- An efficient crowdsensing subsystem with indoor positioning for the recognition of falls and/or stampedes among other relevant physical activities of the occupants of the security zone using the inertial units of smartphones and machine learning techniques. These events allow detecting anomalous situations and generate specific indications to the potential victims.

– An integration framework that analyses the indications that are received from the vision and crowdsensing systems, and processes them to establish levels of priority, probability percentages and actions to be carried out. The results will be shown to the users according to their role:
  - Security agents: a central console shall be provided with all information relating to the indications received, giving the possibility of confirming the threat or discarding it and offering a list of possible actions to be taken in response. The main objective in this case will be to optimise the precision in all the components involved so that the system is of the maximum utility for the security personnel.
  - All other users in the security zone: they will receive indications to follow through a mobile application (for example an escape route or a recommendation to lock the door and hide) based on their position relative to the threat, the nature of the threat and the user profile (e.g. reduced mobility). This functionality flows from the integration framework to the crowdsensing subsystem communicating with individual devices and the biggest challenge will be to achieve a method capable of adapting and reacting quickly in a changing scenario considering a large number of parameters.

The place where the system is being tested is the School of Computer Engineering at the University of Seville, Spain. The building takes up more than 9500 m^2 and can be seen in Fig. 2. It has a closed circuit television (CCTV) composed by more than 50 cameras where five of them have been shared with the research team. It also has a infrastructure of WiFi with more than 400 access points.

Fig. 2. School of Computer Engineering building

In the next sections we will show the approaches that are being studied to implement these components, specifically the vision subsystem in Sect. 2 and the crowdsensing subsystem in Sect. 3, on which work is already underway. Finally, we draw our conclusions in Sect. 4.

2 Computer Vision-Based Weapon Detection

2.1 Object Detection

From the International Joint Conference on Neural Networks in 2011 where IDSIA team [6] won the German traffic sign recognition benchmark and ImageNet Large Scale Visual Recognition Challenge 2012 where AlexNet won [14], object detection is evolved by leaps and bounds. Competitions such as Imagenet [21] or COCO [17] have promoted these advances.

In video surveillance, the detection of dangerous objects as weapons has been studied with the Deep Learning methodology, but only very recently. In [19], Faster-RCNN was shown as the best detector in this task using a dataset generated from violent movies, also giving a very low response time, 5 frames per second.

Unfortunately, these images, normally foreground images of pistols and several kind of guns, differ from fixed CCTV cameras that obtain a wide shot, transforming the problem into a hardest one: small object detection.

2.2 Autoencoder

The development of a detection system is normally driven to achieve good detection and false positive rates on a certain dataset. Ideally, training data would contain representative instances from all possible application scenarios. In practice, obtaining such a huge amount of data is not feasible in terms of time and resources. This problem forces data scientists to be cautious about overfitting and poor generalization when training new models [27]. Some techniques such as dataset partitioning, L1 and L2 regularizations or early stopping are applied to alleviate them [22]. However, misclassification of samples in new scenarios must be addressed. Thus, it is conceivable that a weapon detector could be trained using a dataset containing instances from all possible weapons that provides accurate detections and a small number of false positives. Then, when put into a surveillance system in a real scenario the result is generally an unbearable rate of false alarms [23]. This means that the system will almost certainly be switched off, specially in cases where the incidence of the event of interest is very low. In this context we propose to add an additional step that models and filters the typical false alarms of the new scenario while maintaining the ability to detect the objects it was trained for (Fig. 3).

In the first step of the process, the detector runs in the new scenario over a period of time, saving all the detections. Those detections, that with high probability will be false positives, are then used to train an autoencoder. Autoencoder networks learn how to compress input data into a short code and then reconstruct that code into something as close as possible to its original input and are commonly used for anomaly detection [10]. In this case the autoencoder is trained to model one class: the typical false positives of the new scenario. Finally, it is applied to reconstruct images from a test dataset that contains also instances from the searched objects. If the reconstruction error is compared between both

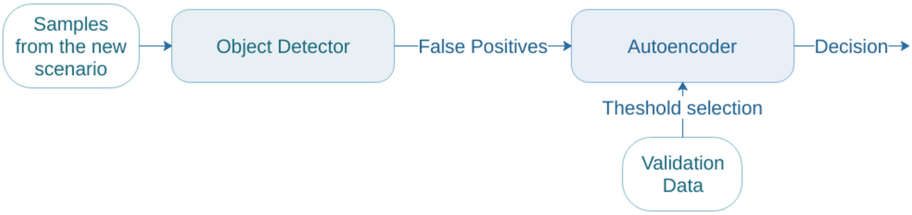

Fig. 3. Autoencoder training phase

classes, it is lower for the class used to train the autoencoder and thus, a threshold could be established to separate them. Figure 4 illustrates this difference by computing the reconstruction error as the mean squared error (MSE).

We have applied the proposed methodology with a handgun detector. A previously available detector has been used for this purpose [1]. For the dataset we downloaded 4 videos from YouTube where people appear testing handguns in the countryside. In this scenario most false positives of the detector are caused by the trees in the background (Fig. 5). Three of the videos were used to train and adjust the threshold of the autoencoder, following a 3-fold cross-validation approach. The remaining video was used only for testing both the detector and the detector+autoencoder configurations.

The experimental results show an improvement in the false positive rate at roughly the same detection rates (Fig. 6). To obtain the detector+autoencoder ROC curve the detector operational point is fixed (at the optimal threshold selected during training) and the autoencoder threshold is made to vary from the lowest to the highest value.

3 Crowdsensing and Action Policy

Mobile crowdsensing consists in the extraction and sharing of data coming from mobile device sensors carried by a group of users. The information can be analysed from all sources and draw conclusions about a common interest. Crowdsensing has been used in various security-related scenarios. We can highlight the systems for detecting incidents related to traffic [20], as well as natural disasters such as fires [18] or earthquakes [8]. Generally, an analysis is made of messages published in various social networks that allows monitoring of a specific emergency event [26], thereby making use of "social media" as a crowdsourcing mechanism [25] for the contribution and dissemination of information on the effects of the emergency on the population. There are also studies based on crowdsensing, using sensors from mobile devices such as GPS [5] to develop organizational strategies that avoid crowding and the risk of incidents. Indoors, localisation from WiFi footprint [11] or Bluetooth [4] is also used. For the detection of physical activities of individuals [15], inertial sensors are used, where the accelerometer is the most used device because of its low energy consumption. In this context, thanks to the use of data from inertial sensors and through the

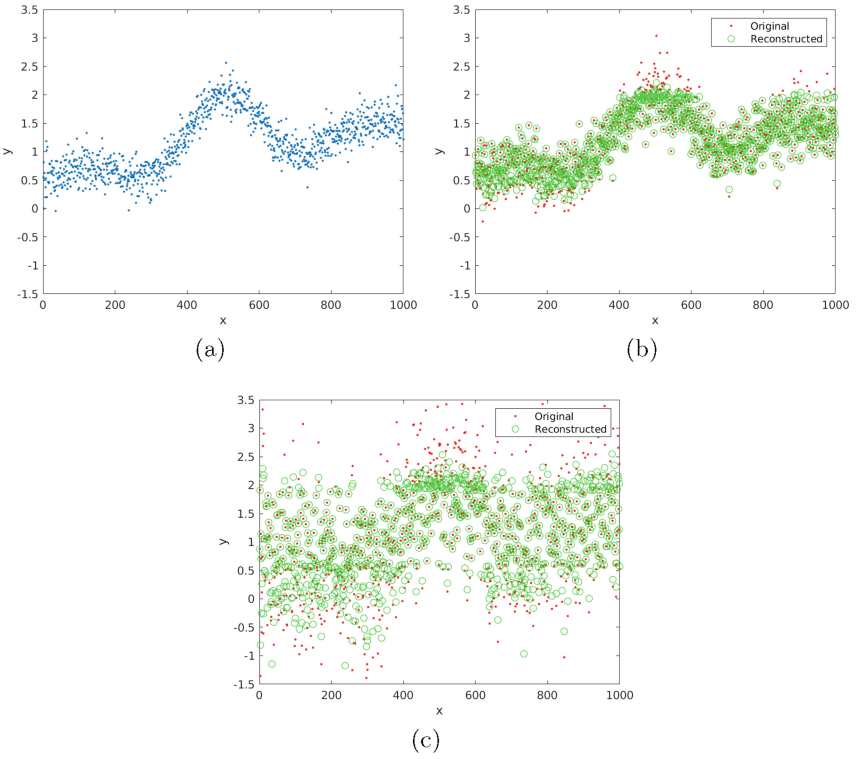

Fig. 4. Reconstruction error comparison. The more the data differs from the training set the higher the error is. (a) Autoencoder training data from the function $y = 1 + 0.05x + sin(x)/x + 0.2 * r$ where r is a normally distributed random number. (b) Reconstructed data from $y = 1 + 0.05x + sin(x)/x + 0.3 * r$ with $MSE = 0.533$. (c) Reconstructed data from $y = 1 + 0.05x + sin(x)/x + 0.8 * r$ with $MSE = 1.099$.

Fig. 5. Sample frame from one of the videos used and a typical false positive in this scenario (enlarged).

application of different classification algorithms, it is possible to recognize the physical activity that the user is doing at each moment. In the field of security, there are systems [2] capable of recognizing the steps and length of the user's stride to design an efficient evacuation system. Falls and other daily live activities [9] are other events that have been studied in depth based on accelerometers [7]. However, the vast majority of these works focus on the detection of falls in a context of Ambient Assisted Living [3], and there are no systems evaluated for the detection of falls in emergency contexts.

3.1 Indoor Positioning and Activity Recognition

As previously commented, one of the objectives of this work is to obtain the location of the different users of the system inside the building. This functionality allows to determine, in case of a threat, the exact position of the users and to propose, if so required, an escape route adapted according to its location and the characteristics of the route.

There exist several commercial solutions that require a high initial investment in infrastructure but due to the more than 400 WiFi access points (APs) allocated within the building a fingerprint method based on [13] has been developed. A fingerprint is composed of several access points, identified by its MAC address and the received signal strength (RSS) value observed by a mobile phone.

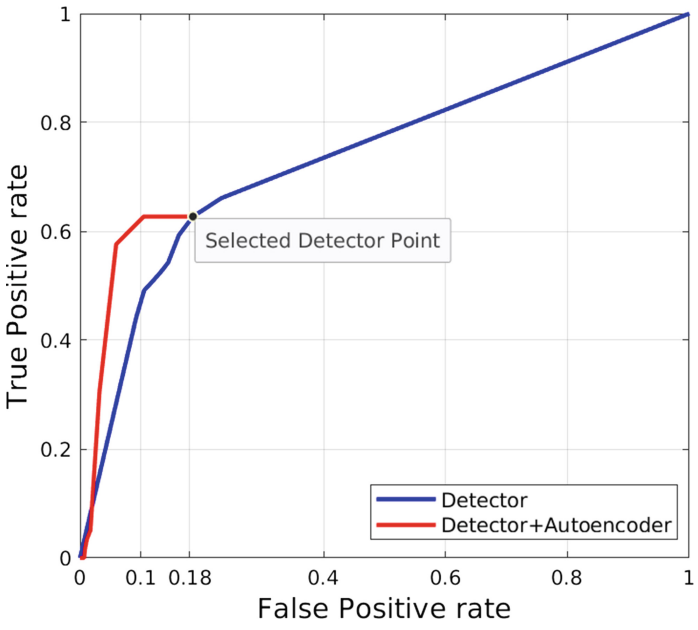

Fig. 6. ROC curves. The autoencoder shows a reduction of 8% in the number of false positives while maintaining the detection rate of the selected detector operational point.

To train the system hundreds of fingerprints have been collected in the middle of corridors every 2 m and in several parts of each classroom.

To test the indoor positioning system an Android app has been developed, as it can be seen in Fig. 7. The app obtains fingerprints every 30 s and compares them with the training set using overlapping of APs and the correlation with their RSS. Results show that the mean average error is 9.1 m. The accuracy can be improved using other methods such as dead reckoning or including deep learning but the system must be used by hundreds of students and lecturers in the building so a trade-off between accuracy and battery has been obtained simplifying the positioning technique and including some improvements: the accelerometer is used to detect steps, when no steps are detected during 3 min, the app stops the fingerprint gathering up to the next step, minimizing the battery draining.

Another important feature from the app is the use of the accelerometer to detect falls and run activities (stampedes). A simple approach is used to detect them using acceleration thresholds. Although false positives occur, crowdsensing here is primary to filter them: when two or more users are experimenting one of these activities in close locations, an emergency event is transmitted to the security personnel and they can check the CCTV and the location of all the application users to confirm or rebut the alarm. Isolated events can normally be avoided if there are people close to the event but nothing changes in their behaviour.

3.2 Risk Analysis and Security Policies

In order to provide adequate responses to the users (members of the security team and regular users of the monitored area) it is necessary to carry out a previous risk analysis of the recognized event as well as a study of the structural characteristics of the area.

In relation to the response that must be provided to the security team, previous sections have shown how the solutions designed must reduce the number of false positives in order to maximize user confidence in the alarms generated. A correct implementation of this type of solutions involves working collaboratively with the control centers, since a filtering process is required that highlights the events detected by the sub-systems and that must be evaluated by the security personnel to decide on the suitability of activate the response protocols.

To avoid unnecessary moments of confusion and panic, only when an alarm has been recognised as a real hazard should some action be taken to inform and guide the other users in the area. To this end, the project proposes channels of communication in addition to the standards (loudspeaker systems and light signals). Specifically, the developed mobile application allows to send personalized information in real time in order to guide users on the safest actions during the valuable initial minutes of the hazard situation.

Taking into account the risk analysis of the area (prior to the implementation of the system and that only needs to be updated when structural changes occur in the topology of the area), and the features of the detected threat, the platform will inform the user about the main actions to be carried out considering their

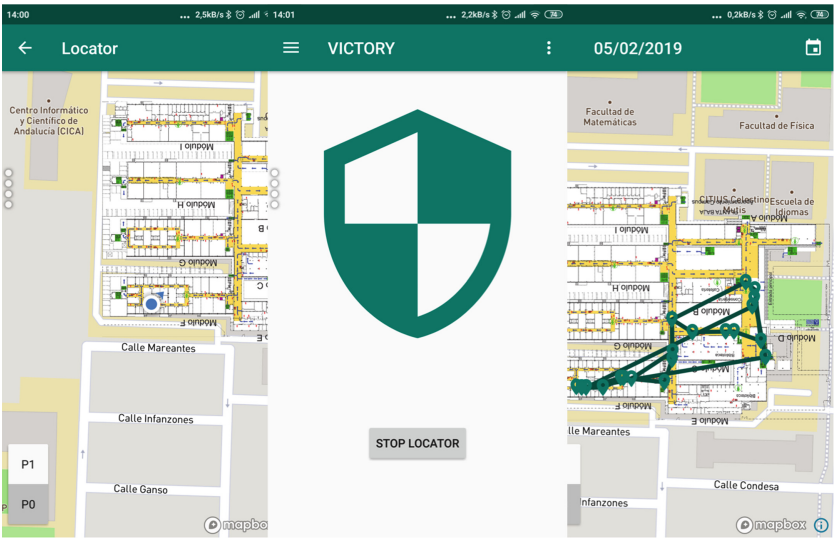

Fig. 7. User interface for Victory App

current position within the area and personal traits. Thus, the main actions that the platform will suggest, depending on the classification of the threat, are:

- Evacuation Route: on a map of the area, indications to follow a safe route to a nearby exit.
- Assurance Tips: actions that increase the user security during the first few minutes of the threat (for example, which accessible area is safe and what procedures to perform to maintain and improve the security).
- Collective Behaviour Tips: basic rules to increase the safety environment of other users in the area.

The main problem we encounter both in the risk assessment process and the generation of security protocols is the high dependence of these on the particular parameters of the area under control and the complexity of modelling and predicting human behaviour under panic situations. Although in the literature we can find some work about similar topics, most of the studies focus on the interactions between humans and their surroundings during evacuation in general emergencies [12], but there are very few studies about other threats (such as violent assaults, including terrorist attacks with planning) [24]. If it is normally hard to model the reaction of a group under a panic situation when the focus of danger is clear and inert, it is even harder when the focus requires additional modelling due to human causes [16].

As a first approach to this problem, we have decided to use multi-agent modelling that allows to parameterize in an open and independent way the possible dangerous conditions (fire, attack of mobile agents with weapons, bombs, attack in group, etc.) and a collection of diverse users that have to react to the

detected threat. We aim to provide an experimental platform on which to test various response strategies and measure their effectiveness under a controlled environment (Fig. 8).

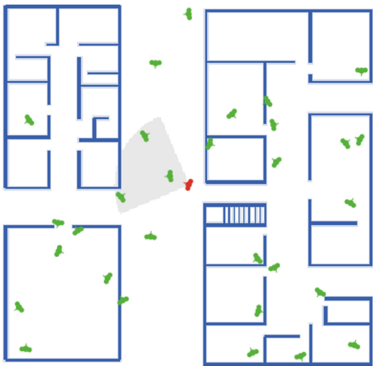

Fig. 8. Hazard Simulation environment.

4 Conclusions

We have introduced the VICTORY security system, which seeks to improve current security systems by providing better reaction time and the ability to generate an optimal response through automation and the use of mobile devices to help potential victims.

The system is mainly divided into a computer vision component and a crowd-sensing component, joined by an integration framework that aggregates all the information and generates the response. The vision component uses deep learning techniques to detect threats such as firearms or knives automatically. The crowd-sensing component analyses the data from the sensors of the mobile devices of the users who are in the security zone, detecting falls, stampedes or other signs of violence. These devices also provide users with personalised and dynamic guidance to help keep them safe.

The VICTORY system is still an ongoing project in which there is still a long way to go, but we are confident that the application of new technological advances to security systems can in the near future significantly reduce the damage caused by violent events that occur periodically worldwide.

References

1. Gun detector. https://github.com/swatz10/Weapon-Detection-Final-Year-Project. Accessed 09 Feb 2019
2. Ahn, J., Han, R.: An indoor augmented-reality evacuation system for the smartphone using personalized pedometry. Hum. Centric Comput. Inf. Sci. **2**(1), 18 (2012)

3. Álvarez-García, J.A., Barsocchi, P., Chessa, S., Salvi, D.: Evaluation of localization and activity recognition systems for ambient assisted living: the experience of the 2012 evaal competition. J. Ambient Intell. Smart Environ. **5**(1), 119–132 (2013)
4. Basalamah, A.: Sensing the crowds using bluetooth low energy tags. IEEE Access **4**, 4225–4233 (2016)
5. Blanke, U., Troster, G., Franke, T., Lukowicz, P.: Capturing crowd dynamics at large scale events using participatory GPS-localization. In: 2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), pp. 1–7. IEEE (2014)
6. Cireşan, D., Meier, U., Masci, J., Schmidhuber, J.: Multi-column deep neural network for traffic sign classification. Neural Netw. **32**, 333– 338 (2012). https://doi.org/10.1016/j.neunet.2012.02.023, http://www.sciencedirect.com/science/article/pii/S0893608012000524. Selected Papers from IJCNN 2011
7. de la Concepción, M.Á.Á., Morillo, L.M.S., García, J.A.Á., González-Abril, L.: Mobile activity recognition and fall detection system for elderly people using ameva algorithm. Pervasive Mob. Comput. **34**, 3–13 (2017)
8. Crooks, A., Croitoru, A., Stefanidis, A., Radzikowski, J.: # earthquake: Twitter as a distributed sensor system. Trans. GIS **17**(1), 124–147 (2013)
9. Gjoreski, H., et al.: Competitive live evaluations of activity-recognition systems. IEEE Pervasive Comput. **14**(1), 70–77 (2015)
10. Gutoski, M., Ribeiro, M., Aquino, N.M.R., Lazzaretti, A.E., Lopes, H.S.: A clustering-based deep autoencoder for one-class image classification. In: 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI), pp. 1–6 (2017)
11. He, S., Chan, S.H.G.: Wi-fi fingerprint-based indoor positioning: recent advances and comparisons. IEEE Commun. Surv. Tutorials **18**(1), 466–490 (2016)
12. Helbing, D., Farkas, I., Vicsek, T.: Simulating dynamical features of escape panic. Nature **407**, 487–490 (2000). https://doi.org/10.1038/35035023
13. Jiang, Y., et al.: Ariel: automatic wi-fi based room fingerprinting for indoor localization. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 441–450. ACM (2012)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
15. Lara, O.D., Labrador, M.A., et al.: A survey on human activity recognition using wearable sensors. IEEE Commun. Surv. Tutorials **15**(3), 1192–1209 (2013)
16. Li, S., Zhuang, J., Shen, S.: A three-stage evacuation decision-making and behavior model for the onset of an attack. Transp. Res. Part C: Emerg. Technol. **79**, 119–135 (2017). http://www.sciencedirect.com/science/article/pii/S0968090X17300840
17. Lin, T., et al.: Microsoft COCO: common objects in context. CoRR abs/1405.0312 (2014). http://arxiv.org/abs/1405.0312
18. Nunavath, V., Prinz, A.: Liferescue: a web based application for emergency responders during fire emergency response. In: 2016 3rd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM), pp. 1–8. IEEE (2016)
19. Olmos, R., Tabik, S., Herrera, F.: Automatic handgun detection alarm in videos using deep learning. Neurocomputing **275**, 66–72 (2018)
20. Pan, B., Zheng, Y., Wilkie, D., Shahabi, C.: Crowd sensing of traffic anomalies based on human mobility and social media. In: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 344–353. ACM (2013)

21. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
22. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
23. Vállez, N., Bueno, G., Déniz, O.: False positive reduction in detector implantation. In: Peek, N., Marín Morales, R., Peleg, M. (eds.) AIME 2013. LNCS (LNAI), vol. 7885, pp. 181–185. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38326-7_28
24. Wang, H., Mostafizi, A., Cramer, L.A., Cox, D., Park, H.: An agent-based model of a multimodal near-field tsunami evacuation: decision-making and life safety. Transp. Res. Part C: Emerg. Technol. **64**, 86–100 (2016). http://www.sciencedirect.com/science/article/pii/S0968090X15004106
25. Xu, Z., et al.: Crowdsourcing based description of urban emergency events using social media big data. IEEE Trans. Cloud Comput. **10**, 1109 (2016)
26. Xu, Z., Liu, Y., Zhang, H., Luo, X., Mei, L., Hu, C.: Building the multi-modal storytelling of urban emergency events based on crowdsensing of social media analytics. Mob. Netw. Appl. **22**(2), 218–227 (2017)
27. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization (2016)