Detection of Violent Scenes at Public Places by Classifying Range-Doppler Maps

Hans-Günter Hirsch[*], Tobias Bolten[*], Regina Pohle-Fröhlich[*], Manfred Hägelen[**], Rainer Jetten[**], Reinhard Kulke[**]

[*]Institute for Pattern Recognition, Niederrhein University, Krefeld, Germany
email: hans-guenter.hirsch@hs-niederrhein.de

[**]IMST GmbH, Kamp-Lintfort, Germany
email: manfred.haegelen@imst.de

Abstract: *A classification system for analyzing sequences of range-Doppler maps with neural networks is presented. The system aims at the detection and classification of dangerous or violent scenes in public places. We have installed radar modules inside and outside a train station to capture radar data in these scenarios. Recognition experiments are performed on these data. Augmentation methods have been developed to artificially create additional training data. Recognition results are presented.*

1. Introduction

A radar-based approach is investigated to monitor public places with a high probability that people will get into dangerous situations or violent conflicts will occur. We use radar systems as an alternative to the visual inspection with cameras. The use of cameras can be considered as an invasion of the privacy of people present in public spaces. Moreover, according to the German legislation, police units are only allowed to observe such places with cameras when there is a high probability that violent activities will occur. Thus, the application of radar technology can avoid several problems and difficulties related to data privacy and legislation. As an activity of the research project KIRaPol.5G [1], we installed several radar systems in the entrance hall of the train station in Mönchengladbach as well as at the public space at one of the entrances to the train station.

We examine the sequence of range-Doppler maps as output of the radar modules. A detection and classification system based on neural networks uses the time sequence of these maps as input. For detection, we aim at a binary classification of scenes as either containing dangerous or violent activities or containing normal everyday situations. The detection of a dangerous or violent situation can be used in a later application to send an alarm to a central police station. Aiming at a more detailed analysis, we also investigate the classification of motion sequences as representing one of several characteristic scenarios, such as a person falling, two people hitting each other, or several people running away in a panic situation. To train the binary and the multi-class systems, we recorded radar data of motion sequences while playing dangerous or violent scenes with volunteers on the university campus or at a police training centre. Due to the need for a large amount of training data, we investigated approaches to artificially generate additional data from the captured radar data. This process is well known as data augmentation.

A good overview of research in radar-based human activity recognition (HAR) is given in [2]. Range-Doppler maps are used as a basis for HAR, e.g., in [3], but also in other application areas such as hand gesture recognition [4] or drone detection [5]. The detection of falling people as a special HAR by analyzing Doppler maps is presented, e.g., in [6]. An overview of fall detection

ISBN 978-3-00-079305-9

approaches is given in [7]. Almost all studies on radar-based detection and recognition are based on the application of neural networks [8].

In the spring and summer of 2024, four sensor nodes were installed inside the train station and outside in the public space. Several thousand hours of radar data were collected until the end of 2024. An overview of the system configuration used is given in the following section. The processes for acquisition and annotation of radar data are described. Augmentation methods are introduced to artificially generate further training data from the acquired data. We investigate different neural networks structures to detect and classify dangerous or violent activities in the captured scenes. Recognition results are presented.

2. System Structure and Setup

The project KIRaPol.5G [1] aims at the configuration shown in Figure 1.

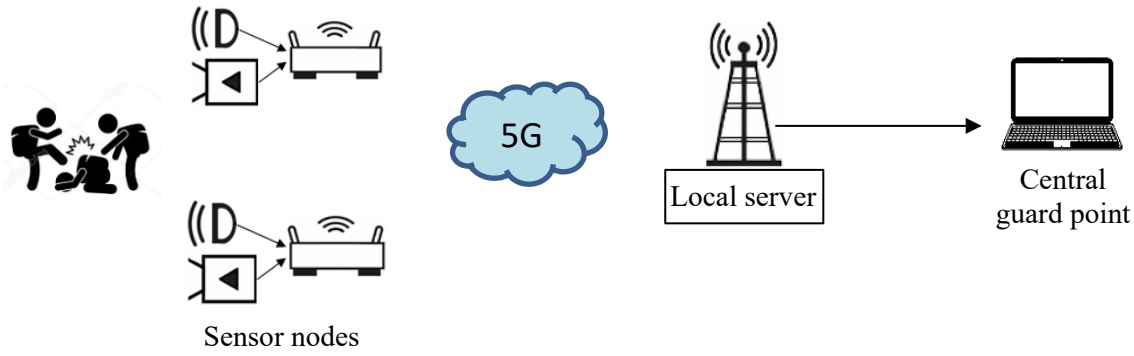

Figure 1. Structure of the observation and detection system.

We installed two sensor nodes in the entrance hall of the train station at a height of about 4 meters, and two further sensor nodes outside the train station. The orientation of the first node inside the station was towards the entrance, while the second node covered the area of the station tunnel that travelers use to access the platforms. Outside the train station, the sensor nodes were installed on the roofs of two opposite buildings at heights of about 4 and 8 meters. The sensor signals from the two nodes inside and outside the train station were transmitted and stored on two separate local servers. Only one sensor node inside the station was connnected to its server by cable. All other nodes transmitted their data to the corresponding server via a private 5G cellular network.

Each sensor node is composed of two radar modules. The two modules are positioned to cover different but slightly overlapping sectors of the observation area. Each sensor node inside the train station also contains two cameras. The field of view of these cameras is as close as possible to that of the radar modules. The video signals are recorded only for the purpose of annotating the captured radar data. Annotation includes the time information of individual scenes to define the sequence of range-Doppler maps needed for training individual classes. Outside the train station, we did not record video data because of a recommendation from the commissioner for data protection of the state of North Rhine-Westphalia. She argued that the university is not allowed to record video data in public places, even if it is only for research purposes and independent of any anonymization.

The range-Doppler maps are calculated in each radar module. Further details of the module can be found in [9]. The data of the maps are transmitted to the servers and stored there. In a later application, the sequence of range-Doppler maps would be used as input for a detection and classification algorithm running on the local server. If a dangerous scene is detected, an alarm

signal would be sent to the police so that a police officer could turn on the corresponding camera and observe the scene.

The size of the range-Doppler map is defined by a few parameters that can be set during initialization of the radar module. A single frequency ramp as part of the FMCW-based analysis covers a bandwidth of approximately 900 MHz. The number of range bins as one of the parameters is set to 512. However, we consider and transmit only 168 Doppler spectra covering the distance range between about 3 and 30 meters. Another parameter is the number of frequency ramps, which defines the size of the Doppler spectrum containing the corresponding number of velocity values. This number is set to 128. The time between two frequency ramps defines the velocity range covered by the Doppler spectrum. With respect to the expected velocity components in case of violent activities, we have chosen this third parameter so that the 128 Doppler bins cover the velocity range from about -6 m/s to +6 m/s. To avoid interferences between the two radar modules of a sensor node, a time delay is used to trigger the measurement and the calculation of the range-Doppler map in the second module. As a result of this parameterization, we can calculate 12.5 maps per second in each radar module.

3. Data Acquisition and Augmentation

The goal of the project is to detect dangerous or violent activities by analysing body movements. In this context we have focused on three scenarios:

- A hurt or helpless person is falling.
- A group of people is running away in a panic situation.
- Two or more people, approaching and attacking each other.

To train either a binary or a multi-class classification system, we need radar data for the dangerous or violent scenes. Compared to other areas of signal processing, such as speech or image processing, there are no publicly available databases with range Doppler maps for such activities. Therefore, we collected radar and video data on the university campus and at a police training centre. We used a single radar module and a single camera of the same type as the ones at the train station. The participants were instructed to behave and to move as they would in real scenarios containing such dangerous or violent activities. This is something that police officers do in a very realistic way as part of their training. Recently, we also played and recorded some dangerous scenes inside and outside the train station. In this way, we also collected a certain amount of data directly at the target locations of the project. We developed a tool based on a graphical user interface for the manual annotation of the radar data. The time information about the beginning and the end of a scene as well as the class assignment are stored in a label file.

The effort to play dangerous and violent scenes with volunteers is high. On the other hand, as much radar data as possible is needed to reliably train a classification system based on neural networks. Therefore, we investigated the artificial generation of additional data from the recorded data. This is known as data augmentation.

Two methods have been developed to generate additional data. The first method aims at shifting activities that occur in a certain distance range to larger distances. This avoids the effort of recording the same scene at several distances. Such activities cause components at the corresponding distance range in the range-Doppler map. When playing scenes on campus, the volunteers were active at 6 to 10 meters in front of the radar module. There were no other people or objects moving in the radar's field of view. We developed an algorithm to determine the distance range in the range-Doppler map where activity occurs. The Doppler spectra in this range are shifted to a greater distance. The amplitudes of each Doppler spectrum must be attenuated due to the shift to the greater distance. We estimated the attenuation factors using an empirical approach. A person carried a radar reflector and walked near the radar module from

a large distance while pointing the reflector at the module. We determined the maximum in each range-Doppler map during this movement. From the maximums of successive Doppler maps, we can estimate a characteristic for the attenuation that decreases for larger distances.
Based on the attenuation characteristics we implemented the shifting of the Doppler spectra to a larger distance by a predefined number of meters. In this way, the training includes the occurrence of violent scenes at a greater distance, as may be the case at the train station.
The second method is to consider additional people or other objects moving in the radar's field of view. Typically, many people who are not involved in the violent activity pass through the recording area. We developed a tool to overlay different parts of two range-Doppler maps. As mentioned before, we recorded the radar data for dangerous scenes where the volunteers moved at distances between 6 and 10 meters. In addition, we also recorded data where several people moved at larger distances in the background. The artificial creation of further sequences of range-Doppler maps is realized by randomly selecting a time segment of these background activities. The parts of the corresponding range-Doppler maps at larger distances are overlayed with the parts of the violent scene at smaller distances.

4. Classification

To classify a scene as containing dangerous or violent activities, we apply a neural network. The input to the neural network is the sequence of range-Doppler maps as shown in Figure 2.

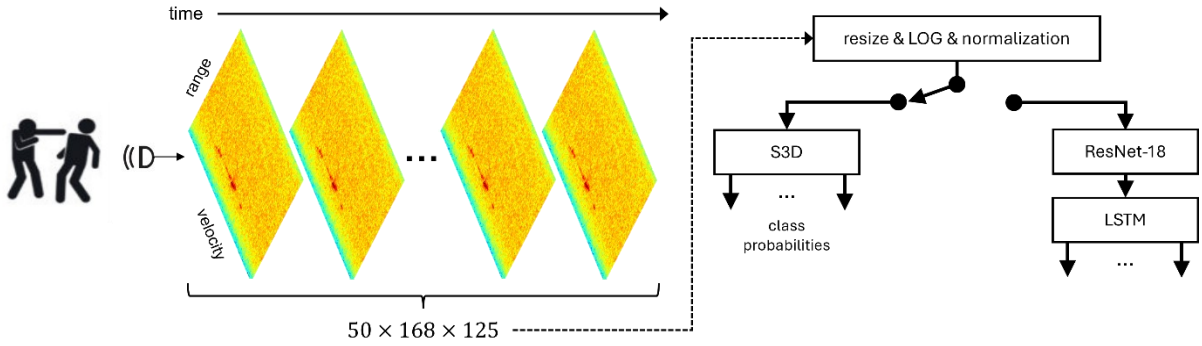

Figure 2. Structure of applied processing.

Range-Doppler maps are available from the radar module at a rate of 12,5 maps per second. By ignoring and erasing the three velocity components at and around 0 m/s, each map contains 168 times 125 values. The logarithm of the velocity amplitudes is calculated. The logarithmic values of each map are normalized. The normalization of each map is done by subtracting the mean of all 168 times 125 logarithmic values. In addition, after subtracting the mean, the values are multiplied by a fixed factor so that the input to the network consists of values in the range of approximately -1 to +1. Fifty consecutive range-Doppler maps are used as input values to the network, so that the motion within a segment of 4s is classified. We use the term snippet for such a 4s segment. In the case of a single action, such as a person falling, we extract only a single snippet from the recordings for training. In case of a repeated action like boxing or punching, we extract multiple snippets for that scene with an overlap of 2s.
We investigate two different neural networks structures. The first one combines a Resnet-18 convolutional network (CNN) [10] with three LSTM layers (long short-term memory). The Resnet-18 has been developed in the field of image recognition. We use it to analyse the content of each range-Doppler map. The successive LSTM layers analyse the time sequence at which the range-Doppler maps occur. The second approach [11] has been developed to analyse the image sequence of a video. It uses a type of 3D convolution consisting of separate spatial and

temporal convolutions. The term "S3D" is used to describe this separable 3D CNN. In our case, we use the range-Doppler maps as input to the network instead of images.

In a first experiment, we trained both network structures with about 5900 snippets. 2400 snippets were extracted from the scenes recorded at the campus and at the police training center, or were artificially generated from the recorded data using data augmentation. The remaining 3500 snippets were randomly extracted from radar data containing everyday scenes without dangerous or violent behavior. For testing, we used another set of 4750 snippets containing dangerous or violent activities or everyday scenes. The dangerous and violent scenes were also played and recorded on campus. Table 1 shows the accuracies and the F1 scores as percentages for both network structures and the binary and 5-class classification. The results are considerably better for the S3D structure. This was also observed in the later experiments, so we do not present further results for the ResNet structure. The binary distinction between dangerous and non-dangerous scenes works very well with percentages above 99%.

	S3D		ResNet18 & LSTM	
	2 classes	**5** classes	**2** classes	**5** classes
Accuracy	**99,4**	**99,2**	**94,1**	**97,1**
F1	**99,3**	**96,5**	**93,4**	**79,8**

Table 1. Accuracies and F1 scores for classification experiments with campus data.

In a second experiment, we used a set of 7650 snippets for training. These snippets were extracted from all scenes played on campus and at police training centre or were created by data augmentation. We trained the S3D network twice, once including the augmented data and once without the augmented data. Another set of 306 snippets was used for testing. These snippets were extracted from dangerous or violent scenes played in front of one of the radar modules in the entrance hall of the train station. The results are shown in Table 2. The benefit of including augmented data in the training becomes clear. The performance is worse than in the first experiment due to the different recording environment. The entrance hall has other reflection properties. Furthermore, the high number of people moving through the entrance hall parallel to the played scene is the main reason for the loss.

	with augmented data		**without** augmented data
	2 classes	**5 classes**	**5 classes**
Accuracy	**91,8**	**92,7**	**85,9**
F1	**91,6**	**73,5**	**69,4**

Table 2. Scores for classification experiments on train station data, training without and with augmented data.

Radar data recorded in the entrance hall with two different modules over several months were used for the third experiment focusing on binary classification. Assuming that no dangerous or violent activities occurred, all „dangerous" classifications would trigger a false alarm. False alarm rates are listed in Table 3 for sensor nodes SG_21 (towards the tunnel) and SG_31 (towards the entrance). The data at node SG_21 consists of 2,236,000 snippets (~1240 hours) and at node SG_31 of 4,060,000 snippets (~2250 h). The results are presented for the separate classification of each indidivual snippet and after applying a postprocessing that includes an alarm detection only in case several consecutive probabilities at the corresponding output node of the neural network exceed a predefined threshold. The considerable gain becomes evident when the alarm detection is based on the analysis of several consecutive snippets.

	Sensor nodes	
	SG_31	**SG_21**
False alarm rate (individual snippet)	**0,506 %**	**0,717 %**
False alarm rate (including postprocessing)	**0,055 %**	**0,034 %**

Table 3. False alarm rates without and with postprocessing for data recorded in the entrance hall

By training the network with additional data recorded at sensor node SG_31 containing the characteristics at this location, the false alarm rate can be further reduced from 0,055% to 0,023%.

5. Important Remarks

An approach is presented that enables the detection of violent scenes in public, based on the classification of sequences of range Doppler maps. The detection is performed by a neural network with a 3D convolutional structure consisting of separate spatial and temporal convolutions. The training of the classification system can be enhanced by artificially generating additional data from the captured radar data.

6. Acknowledgment

This research project is supported by the ministry of economic affairs, industry, climate action and energy of the state of North Rhine-Westphalia.

References

[1] KIRaPol.5G, research project, https://www.hs-niederrhein.de/auge/kirapol5g/.

[2] I. Ullmann, R.G. Guendel, N.C. Kruse, F. Fioranelli, A. Yarovoy, "A Survey on Radar-Based Continuous Human Activity Recognition*", IEEE Journal of Microwaves*, Vol. 3, 2023

[3] W.-Y. Kim, D.H. Seo, "Radar-based human activity recognition combining range–time–Doppler maps and range-distributed-convolutional neural networks", *IEEE Trans. on Geoscience and Remote Sensing*, 2022

[4] D. Auge, J. Hille, E. Mueller, A. Knoll, "Hand gesture recognition in Range-Doppler images using binary activated spiking neural networks", *IEEE Conference on Automatic Face and Gesture Recognition*, Jodpur, India, 2021

[5] I. Roldan et al., "DopplerNet: a convolutional neural network for recognising targets in real scenarios using a persistent range–Doppler radar", *IET Radar Sonar and Navigation*, Vol. 14, 2020

[6] X. Yu, C. Feng, L. Yang, M. Song, W. Zhou, "Human fall detection by FMCW radar based on time-varying range-Doppler features", *Computer and Systems Engineering*, 2022

[7] S. Hu, S. Cao, N. Toosizadeh, J. Barton, M.G. Hector, M.J. Fain, "A Survey on Radar-Based Fall Detection", *https://doi.org/10.48550/arXiv.2312.04037*, 2023

[8] W. Jiang, Y. Ren, Y. Liu, J. Leng, "Artificial neural networks and deep learning techniques applied to radar target detection: A Review", *Electronics*, 2022

[9] H.G. Hirsch, T. Bolten, F. Terstappen, R. Pohle-Fröhlich, M. Hägelen, R. Jetten, R. Kulke, "Towards the Detection of Violent Scenes at Public Places by Analyzing Range-Doppler Maps", *Int. Conference on Microwaves for Intelligent Mobility*, Boppard, Germany, 2024

[10] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition", https://doi.org/10.48550/arXiv.1512.03385, 2015

[11] S. Xie, C. Sun, J. Huang, Z. Tu, K. Murphy, "Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification", *European Conference on Computer Vision, Munich*, https://doi.org/10.1007/978-3-030-01267-0_19, 2018