Classification of human activity using radar and video multimodal learning

de Jong, Richard J.; de Wit, Jacco J.M.; Uysal, Faruk

DOI
10.1049/rsn2.12064

Publication date
2021

Document Version
Final published version

Published in
IET Radar, Sonar and Navigation

Received: 15 July 2020 | Revised: 21 September 2020 | Accepted: 13 October 2020

IET Radar, Sonar & Navigation

DOI: 10.1049/rsn2.12064

ORIGINAL RESEARCH PAPER

Classification of human activity using radar and video multimodal learning

Richard J. de Jong[1] | Jacco J.M. de Wit[1] | Faruk Uysal[2]

[1]Department of Radar Technology, TNO, The Hague, The Netherlands

[2]Microwave Sensing, Signals & Systems, Delft University of Technology, Delft, The Netherlands

Correspondence

Richard J. de Jong, Department of Radar Technology, TNO, The Hague, The Netherlands.
Email: richard.dejong@tno.nl

Abstract

In the defence and security domain, camera systems are widely used for surveillance. The major advantage of using camera systems for surveillance is that they provide high-resolution imagery, which is easy to interpret. However, the use of camera systems and optical imagery has some drawbacks, especially for application in the military domain. In poor lighting conditions, dust or smoke the image quality degrades and, additionally, cameras cannot provide range information too. These drawbacks can be mitigated by exploiting the strengths of radar. Radar performance can be largely maintained during the night, in various weather conditions and in dust and smoke. Moreover, radar provides the distance to detected objects. Since, the strongpoints and weaknesses of radar and camera systems seem complementary, a natural question is: can radar and camera systems learn from each other? Here the potential of radar/video multimodal learning is evaluated for human activity classification. The novelty of this work is the use of radar spectrograms and related video frames for classification with a multimodal neural network. Radar spectrograms and video frames are both two-dimensional images, but the information they contain is of different nature. This approach was adopted to limit the required preprocessing load, while maintaining the complementary nature of the sensor data.

1 | INTRODUCTION

Human activity classification is a major asset in the defence and security domain. The activity or behaviour a person exhibits may (partly) reveal their intent. A person wandering around a parking lot may be just on the way to a car or he or she may be scanning the cars for possible valuable items to steal. The behaviour of a person, for example, walking speed and walking pattern, may reveal the actual intent.

In the security domain, cameras are widely used for surveillance; camera systems can be found in cities, in shopping centres, in parking garages, in public transportation, on airports etc. This widespread use of cameras in the civil domain is motivated by their ease of use and the fact that optical images are easy to interpret for humans, avoiding the need for extended operator training. Moreover, optical imagery allows the application of facial recognition. This is a crucial asset regarding the prosecution of possible offenders, although it may arouse privacy issues in some situations.

Camera systems do have some drawbacks, especially when considering applications in the defence domain. The quality of (daylight) camera imagery degrades in poor lighting conditions, smoke and dust. Furthermore, a standard camera cannot provide information about the range to a subject.

These issues relate directly to the strongpoints of radar. Radar systems provide the range and velocity of detected subjects, have all-weather capability and maintain their performance in darkness, dust or smoke. Radar imagery is, however, typically unsuited for recognition and is difficult to interpret by a human operator.

Since their strengths and weaknesses seem to complement each other, a natural question seems to be: *Can radar and camera systems learn from each other?*

An example of such multimodal learning is the fusion of video data with laser range measurements for autonomous navigation [1]. The aim of this study was to exploit the complementary nature of the individual sensors. Indeed, the fusion of these complementary measurements leads to better performance, as in general data fusion leads to more consistent and accurate information. However, the crux of multimodal learning is that: the improvement is maintained even if one of the sensor modalities delivers data of degraded quality or is absent all together [2, 3]. An extreme example is the fusion of video data of a speaking person and the related audio signal [4–6]. The ultimate purpose of this multimodal learning application is to achieve robust speech recognition using the video data only, even in absence of the audio signal. This notion formed the basis for the current study.

Here, it was investigated if human activity classification can be improved when feeding corresponding video and radar data to a classifier based on a multimodal convolutional neural network (CNN) instead of a unimodal classifier using only video or radar data. If indeed there is some performance improvement using a multimodal CNN, the next question to be addressed is whether this improvement is maintained when one of the sensor modalities is absent or delivers data of degraded quality (e.g. in darkness when a daylight camera cannot provide suitable data, whereas the quality of the radar data is preserved). To gain insight in the process of multimodal video/radar learning, visualization techniques have been applied to identify the pixels in the images that are exploited by the CNN for classification. The application of CNNs for classification of persons or objects in pictures and video frames is already well-established [7]. Recently, CNNs have also been successfully applied to classify human activity based on radar micro-Doppler signatures [8–11]. The idea of fusing video and radar data using CNNs in a multimodal setup is, however, relatively novel. Neural networks have been applied to register synthetic aperture radar (SAR) images and optical imagery [12, 13]. For registration, the CNN is trained to find related features in the SAR and optical images and exploit these features to align the images. In this application, the complementary nature of radar and optical data is not used per se (the image registration is based on features that are present in both modalities). Only in recent years, methods exploiting the complementary nature of radar and one or more other modalities have been reported [14–18]. Some of these methods are particularly focussed on data fusion and less on cross learning, that is, all modalities are always assumed to be present [14, 15]. In other cases, extended preprocessing is applied to the data, such that the data of the different modalities can be presented in similar format [16–18], for example, by transforming video frames into two-dimensional (2D) heat maps or radar data into SAR images. The novelty of the presented work is that radar spectrograms and video frames are fed to the CNN-based classifier. This approach was adopted to limit the preprocessing. In essence spectrograms and video frames are both 2D images in which pixel colour or intensity has some meaning, but the type of information they represent is very different. The CNN-based classifiers and their performance are discussed in Sections 3 and 4. More details on this work can be found in [19] and it has been presented earlier in [20].

One disadvantage of using a CNN-based classifier is that the temporal relation between successive measurements cannot be exploited. Since successive radar measurements and video frames are correlated sequences, applying recurrent neural networks (RNNs) may improve classification accuracy. RNNs contain a feedback loop and are therefore capable of managing problems with a temporal nature. Several types of RNNs exist, in the current study the long short-term memory (LSTM) architecture is applied. LSTM architectures are used for radar-based human activity classification with success [21–24], including a multimodal LSTM architecture [25]. The assessment of an LSTM network applied to radar and video data for human activity classification is presented in Sections 5 and 6.

For the evaluation of the different architectures, a data set of simultaneous radar measurements and video recordings of walking people was used. For these measurements the test subjects performed three ‘activities’: strolling, walking while carrying an object and walking with a backpack. These three cases are assumed representative for different types of human intent, as persons carrying heavy items may be regarded suspect in specific situations (such as a person carrying a crowbar on a parking lot). In Section 2, the measurements and data sets are discussed in detail.

2 | RADAR AND VIDEO MEASUREMENTS

To evaluate the potential of multimodal learning for classification of human activity, experiments were conducted with a compact radar and a high-definition camera (see Figure 1). Prior to the start of each measurement, a test subject positioned itself at approximately 40 m range from the sensors. At the start of each measurement the test subject began walking towards the sensor systems. With an average walking speed of the test subjects of approximately 1.5 m/s each measurement run was limited to 20 s. The experiments were conducted on two days both during the morning and during the afternoon. The circumstances were similar during the experiments, but the lighting conditions varied depending on the time of day. The test subjects were asked to walk (normally) towards the sensor systems so some natural variation in walking speeds and patterns could be observed.

Measurements were made of test subjects strolling, that is, walking without objects in their hands (class N), walking while carrying a rifle-like object in both hands (class R), and walking with a relatively heavy backpack (class B). It should be noted that the test subjects inclined to swing their arms when not carrying the object. Thirty-five test subjects took part in the experiments. Each test subject performed the defined activities twice; as a result 210 measurements are available for training and validation.

2.1 | Radar measurements

For the experiments, the X-band AMBER frequency-modulated continuous wave radar was used [26]. For these experiments, the sweep repetition frequency was set to 2.5 kHz and the bandwidth was set to 100 MHz. The range resolution (1.5 m) was relatively coarse to capture most of a test subject's

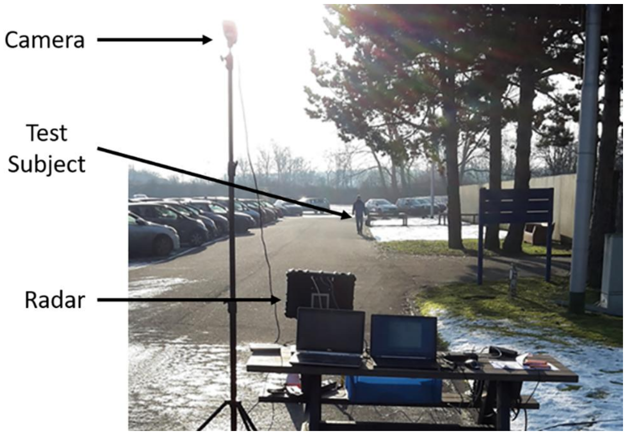

FIGURE 1 Experimental setup with a high-definition camera and a compact radar

micro-Doppler signature in a single resolution cell (including swinging legs and arms). The micro-Doppler signatures were visualised by means of the spectrogram. The spectrograms were obtained by taking the squared magnitude of the short-time Fourier transforms (STFTs) computed from successive, but overlapping sequences of the radar signal. Before computing the related STFT, a Blackman window was applied to a radar sequence. The integration length of each STFT was 0.1 s and the overlap between successive STFTs was 80%. A spectrogram of each class of activity is presented in Figure 2. Each spectrogram shows 1.28 s of radar data, corresponding with at least a single human gait cycle.

It is assumed that the gait and torso motion change if a test subject carries a heavy object. If this difference in gait and torso motion can be recognized in the spectrograms, it can be determined whether the person carries a heavy item or hefty backpack. However, from a previous study it was already concluded that the swinging arms or the lack thereof (if the test subject carries an object with both hands) is the most distinctive feature [27]. This is observable in the spectrograms in Figure 2; if the test subject is strolling or carrying the backpack, the spectrogram exhibits an 'arc' related to the (lower) arm's motion (indicated by the [dashed] arc and arrow). When both hands of the test subject are engaged, this arc is absent. The results of this earlier study also showed that the effect of carrying a backpack (of around 10 kg or a little less) on the micro-Doppler signature is minor, as the classes N and B could not be distinguished [27]. It was also clear from visual observation, that the impact of the backpack on the cadence was negligible for most test subjects. For those test subjects, a heavier load would have been required to impact their movements [28].

2.2 | Video recordings

The video recordings were made using a high-definition daylight camera (1920 × 1080 pixels) with an average frame rate of 13.5 frames per second. As preprocessing step a single-shot detector (SSD) was used [29]. An SSD is a deep neural network for detecting objects in images and providing their bounding box at the same time. In the first stage of an SSD, feature maps are extracted of the input image using a pretrained network. In the second stage, multiple default bounding boxes, of varying size and aspect ratio, are applied to each feature map location. For each default bounding box, a score is generated for the presence of each object class in that box and the box is adapted to better fit the object's size and shape. The results from multiple feature maps with different resolutions are combined, such that objects of varying sizes can be detected. Here, a pretrained SSD was used to detect persons in the video frames. After detection, the area within the defined bounding box, that is, the pixels related to the detected person, is then extracted from the frame. Due to the test subjects walking towards the sensor systems, starting from an initial range of about 40 m, the bounding boxes differ in size in successive frames. As the test subjects come closer to the sensors, the size of the bounding box increases. However, for the chosen CNN implementation all inputs must be of the same size. Therefore, all extracted subimages were resized to a width of 64 pixels and a height of 128 pixels.

The SSD performed relatively well on this data set. A high detection probability (of the order of 90%) and a low false alarm rate was obtained. Nonetheless, in some video frames the test subject was missed, or a random object in the background was detected as a person. These video frames were deleted from the data set by hand.

3 | CNN-BASED UNIMODAL LEARNING

First two CNNs were trained and validated for the video and radar modalities separately. The classification performance of the individual networks is evaluated and the main features contributing to the classification accuracy are assessed. This assessment provides insight in the potentially added value of multimodal learning.

3.1 | CNN background

A CNN is a specialised kind of neural network that excels in processing data with a known grid like topology [30]. A CNN employs a convolutional operation in one or several of its layers. The weights of the kernels are learnt form the data. This allows a CNN to learn increasingly more complex representations as the depth of the architecture increases. A typical CNN architecture for classification consists of a number of convolutional stages that consist of convolutional filters, non-linear activation functions and pooling. The final layers consist of fully connected layers and at last a softmax function to assess the probability of each class.

3.2 | Individual CNN architectures

Keras and TensorFlow were used to implement the neural networks in this work. The hyperparameters of the models

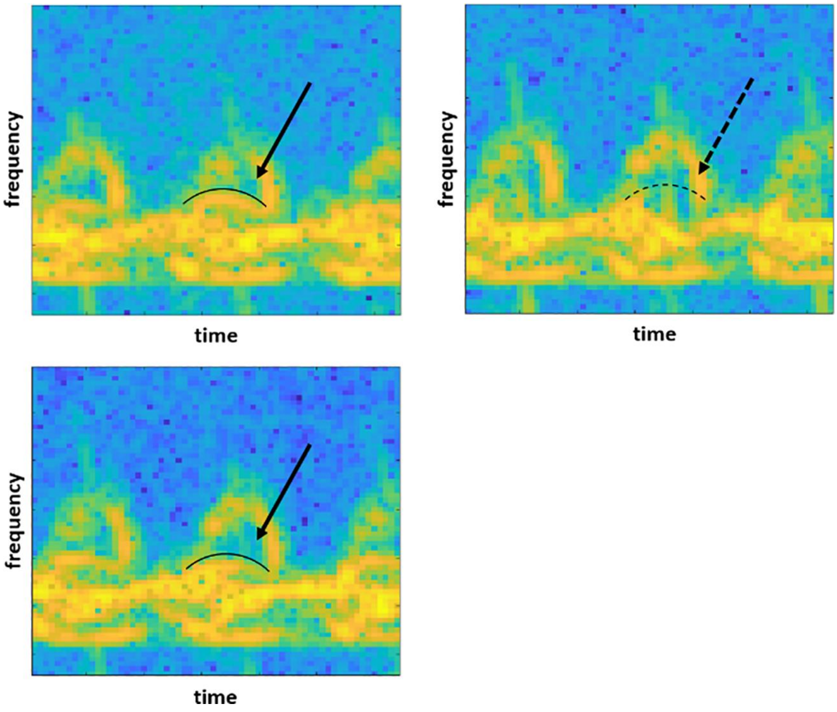

FIGURE 2 Measured spectrograms of a test subject strolling (class N) (top left), holding an object in both hands (class R) (top right) and carrying the backpack (class B) (bottom)

were defined by performing a grid search. The optimised (unimodal) CNN architecture for the radar data is presented in Table 1 and the unimodal CNN architecture for the video data is presented in Table 2.

After each convolution and fully connected layer a rectified linear unit (ReLU) activation function is used.

The last part of both classifiers consists of two fully connected layers with a size of 500 neurons and as output a softmax activation function is applied. The number of outputs is two for the two class classification and three for the three class classification problem.

TABLE 1 Unimodal radar CNN architecture

Layer	Number of filters	Kernel size	Dimensions
Radar input	-	-	64 × 64 × 1
Convolution	20	5 × 5	64 × 64 × 20
Max pooling	-	2 × 2	32 × 32 × 20
Convolution	30	5 × 5	32 × 32 × 30
Max pooling	-	2 × 2	16 × 16 × 30
Convolution	40	5 × 5	16 × 16 × 40
Max pooling	-	2 × 2	8 × 8 × 40
Convolution	50	5 × 5	8 × 8 × 50
Max pooling	-	2 × 2	4 × 4 × 50
(Flatten)	-	-	800
Fully connected	-	-	500
Fully connected	-	-	500
Softmax	-	-	2/3

Abbreviation: CNN, convolutional neural network.

3.3 | Training and validation sets

As stated in Section 2.1, spectrograms were generated from the radar measurements. From each spectrogram, a 1.28 s long excerpt was extracted, such that at least a single human gait cycle is included. This excerpt was paired with the video frame corresponding to the start time of the excerpt. The maximum synchronization error between the start time of the excerpt and the actual time of the video frame is 0.01 s. More than 40,000 of such video/radar image pairs were available for training and validation. The image pairs of 28 randomly chosen test subjects were used for training and the image pairs of the remaining seven test subjects were used for validation. Thus, the training and validation sets were mutually exclusive in terms of the test subjects.

3.4 | Respective unimodal performance

The classification accuracy of the unimodal CNNs is presented in Table 3, for the two-class (N and R), and the three-class (N, R, and B) classification problems respectively. For both problems, the video-based classifier outperforms the radar-based classifier.

As was stated in Section 2.1, the radar-based classifier has difficulties to discriminate the N and B classes; the overall classification accuracy for the three-class (N, R and B) problem is only 62.6%. The confusion matrices in Figure 3 illustrate this. As can be seen, the radar-based classifier is able to distinguish the R class from the other two classes with high probability, constituting the major contribution to the overall accuracy. The confusion between the R and B classes is partly due to the test subjects not moving their arms when walking (recall that the lack of arm motion is the most distinguishing feature of class R). Due to the cold weather during the experiments some test subjects walked with their arms stiffly besides their body. Because of the reduced arm motion this behaviour can result in the absence of the earlier discussed 'arc' feature in the spectrograms. In addition, some test subjects grabbed the straps of the backpack whilst walking, thus reducing the arm motion.

The video-based classifier can distinguish the N and B classes, see the corresponding confusion matrix in Figure 3. This means the video-based classifier must exploit different features than the radar-based classifier. Most likely the clear visibility of the object or the straps of the backpack in optical images is the main feature exploited by the video-based classifier.

If indeed the video and radar-based classifiers exploit complementary information, there might be added value in video/radar multimodal learning. To evaluate the potential added value of multimodal learning, it is examined what information is used for feature extraction by the radar and video CNNs. For this evaluation, the gradient-weighted class activation mapping (Grad-CAM++) visualization technique is applied [31]. A class activation map (CAM) highlights the image pixels that are actually used by a CNN for feature extraction and classification, for example [10, 27]. The generation of such a CAM or saliency map alleviates the black-box nature of CNNs. The Grad-CAM++ is a generalisation of the CAM that can be broadly applied to any CNN network architecture. In order to obtain the Grad-CAM++, the gradient of the class score is computed with respect to the feature maps of a convolutional layer. The Grad-CAM++ is the weighted combination of the feature maps followed by a ReLU.

TABLE 2 Unimodal video CNN architecture

Layer	Number of filters	Kernel size	Dimensions
Video input	-	-	128 × 64 × 3
Convolution	16	3 × 3	128 × 64 × 16
Convolution	16	3 × 3	128 × 64 × 16
Max pooling	-	2 × 2	64 × 32 × 16
Convolution	32	3 × 3	64 × 32 × 32
Convolution	32	3 × 3	64 × 32 × 32
Max pooling	-	2 × 2	32 × 16 × 32
Convolution	64	3 × 3	32 × 16 × 64
Convolution	64	3 × 3	32 × 16 × 64
Max pooling	-	2 × 2	16 × 8 × 64
Convolution	128	3 × 3	16 × 8 × 128
Convolution	128	3 × 3	16 × 8 × 128
Max pooling	-	2 × 2	8 × 4 × 128
(Flatten)	-	-	4096
Fully connected	-	-	500
Fully connected	-	-	500
Softmax	-	-	2/3

Abbreviation: CNN, convolutional neural network.

3.5 | Video recordings saliency maps

In Figure 4, detected video frames are presented with the Grad-CAM++ saliency maps displayed on top. The frames are correctly classified with more than 99% certainty. The examples in Figure 4 are a fair representation of saliency maps observed for this data set. The first row of maps illustrates the general findings; the second row displays some saliency maps highlighting undesirable features.

As illustrated by the top-left saliency map, in case of the test subject strolling, the pixels around the lower arms and hands have high saliency. This indicates that arms hanging loose next to the torso are a key feature to classify a person strolling. The top-middle saliency map shows that in case of the test subject carrying the object, the pixels around the hands and the object have high saliency. This suggests that the presence of an object in front of the torso is the main feature. Finally, the top-right saliency map shows that for the test subject carrying the backpack the pixels around the backpack straps have high saliency, thus correctly contributing to the classification. The free hands also have high saliency, but this is not a discriminative feature for the activity.

The bottom row of detected video frames in Figure 4 is also classified correctly, although the saliency maps highlight undesirable features. In these video frames, the test subject's feet seem an important feature. It is uncertain why the feet should contribute to the activity classification and whether this is a wanted property. Some video frames have been cropped

		{N,R}	{N,R,B}
Single Modality classification	Radar	88.9	62.6
	Video	95.2	87.0
Video/radar multimodal classification	Data-level fusion	95.5	87.0
	Feature-level fusion	96.3	86.9
	Decision-level fusion	96.4	87.3

TABLE 3 Classification accuracy (%) of the individual unimodal classifiers and the different implementations of the multimodal classifier. The results are shown for the two-class (N and R) and three-class (N, R and B) classification problems

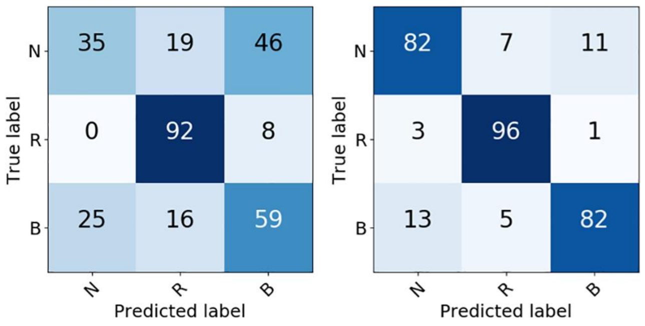

FIGURE 3 Confusion matrices for the radar-based (left) and video-based (right) classifiers. N refers to a person strolling, R to a person holding an object and B to a person carrying a backpack. Classification results are in percentage

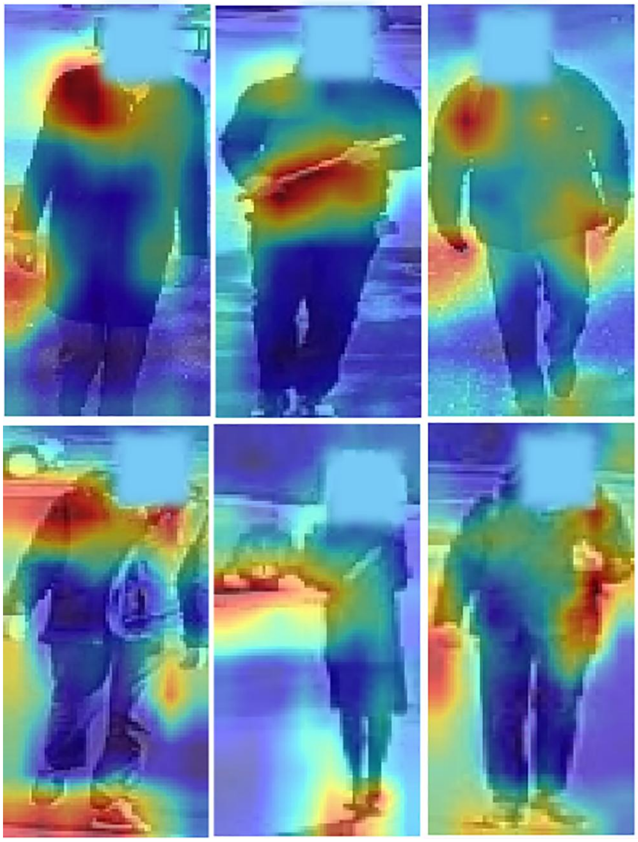

FIGURE 4 Saliency maps with a correct classification (with at least 99% certainty) displayed on top of the related video frame, for a test subject strolling (left column), a test subject holding an object (middle column) and a test subject carrying a backpack (right column)

(by the SSD) such that the feet are cut off, this can cause the feet to contain some information about the position of the test subject and the objects in question. The saliency map displayed on the video frame of the test subject carrying an object is a more pronounced example of undesired features. This saliency map indicates that the surroundings have high saliency. However, the surroundings of the test subject do not contain information about the activity and is therefore an unwanted feature.

3.6 | Radar spectrograms saliency maps

In Figure 5, spectrograms and related maps are presented of a test subject strolling with his/her arms swinging and a test subject holding an object in both hands. These results were produced with a CNN trained to separate the two classes N and R for the unimodal CNN architecture as stated in Section 3.1 (right two columns) and a CNN architecture with a larger last convolutional layer (left two columns). As it is challenging to classify a test subject carrying a backpack based on radar spectrograms, class B was omitted from this evaluation (including activity class B in the evaluation led to noise-like saliency maps). In case of the test subject strolling, the region where the arc of the moving arms is (cf. Figure 2) has high saliency, which is best visible in the top-left image. In case of the test subject holding the object, the response to the torso has high saliency. This evaluation confirms the notion that the arm motion or lack thereof is the major feature to distinguish a strolling person from a person holding an object in both hands in radar spectrograms (given that a person just strolling typically swings his/her arms). Due to the rescaling of the saliency maps with the dimensions of the last convolutional layer, however, the outcome is not always clear, as is illustrated by the two columns on the right.

3.7 | Discussion

Considering the unimodal classifiers, the video-based classifier outperforms the radar-based classifier. This is for a part due to the biased data sets. All measurements were performed during daytime, with lighting conditions well-suited for video recordings. Preferably, measurements should be performed in varying weather and lighting conditions, such that the SSD might fail to recognise humans in the video imagery. In adverse lighting conditions, the performance of the video-based classifier might be degraded more severely than the performance of the radar-based classifier. In such varying conditions, the complementary nature of video and radar measurements will emerge, and a multimodal approach might improve the overall classification performance.

Here the RGB-based imagery was used for the video input, this might be extended to other types of imaging, such as multispectral/hyperspectral imaging. In principle the additional information can be included by increasing the number of channels in the input. However, this would require retraining and validation of the neural networks.

A multimodal approach has added value if the radar-based and video-based classifiers exploit different features for classification. To assess the main features used by the individual classifiers, Grad-CAM++ saliency maps were generated. For both classifiers, the position of the arms is a key feature. If the arms swing loose next to the torso, is a strong indication that

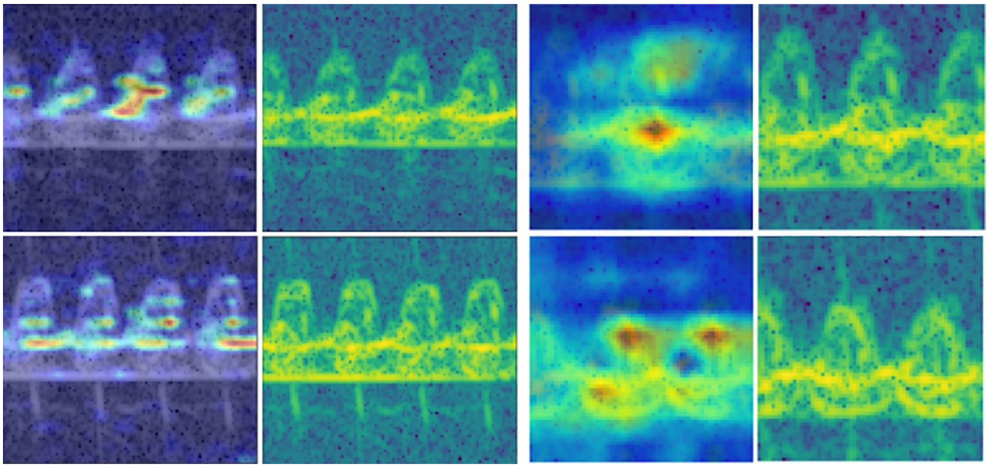

FIGURE 5 Measured spectrograms and related saliency maps. The spectrograms and saliency maps are shown for the current CNN architecture (right two columns) and for a CNN with a larger last convolutional layer (left two columns). The results are shown for a test subject strolling (top row) and a test subject holding an object (bottom row). CNN, convolutional neural network

the person is just strolling. For the video-based classifier, in addition, the presence of an object in front of the torso or the straps of a backpack are important features. The Grad-CAM++ saliency maps provide insight in the pixels (i.e. features) relevant for classification. However, they do not explain why certain pixels are important for classification.

4 | CNN-BASED MULTIMODAL LEARNING

Modality refers to the way the environment or events are perceived. Different types of sensors, for example, acoustic, optical and RF sensors, perceive the environment in different ways and are thus referred to as different modalities. Multimodal CNNs are neural networks that can jointly interpret data from various modalities [32]. Within a neural network architecture, the integration of data from different modalities may be done at different levels. These different levels of integration are displayed in Figure 6.

The architecture on the left illustrates decision-level fusion. The radar and video data are essentially considered independently resulting in two classification results. A joint final stage aggregates the individual result in some way to obtain the overall classification. By using this decision-fusion architecture, the individual CNNs for the video and radar data are trained and validated separately and therefore cannot learn from each other.

The architecture in the middle shows feature-level fusion. In this case CNNs are applied independently to the radar and video data to extract the desired features for each modality. The combined features are then past to a neural network with one or several fully connected layers and a softmax layer to perform the final classification on the shared feature space. By using this architecture, the individual radar and video CNNs may learn from each other in the back-propagation stage, as the weights can be adapted based on the overall classification result.

Finally, the architecture on the right depicts data-level fusion. By using this data-level fusion architecture, the video and radar data are simultaneously input to a single CNN. Potentially, data-level fusion allows deep exploitation of the correlation (or complementarity) between the various modalities. A disadvantage of this approach is that the video and radar images are forced to have the same dimensions (expressed in image pixels). Another potential disadvantage of this architecture is that the modalities use convolutional stages that need to be optimised for the combined input, this does not necessarily lead to optimal feature extraction for the individual modalities.

The classification results of the three multimodal fusion strategies are presented and discussed in the following subsections, in the order of the fusion depth.

4.1 | Data-level fusion

The data-level fusion architecture was implemented by adding the spectrogram excerpt as an extra channel to the video CNN from Table 2. To do so the spectrogram excerpts was upscaled to have the same dimensions as the video CNN input. The combined input dimensions are 128 × 64 × 4. The classification accuracy of the data-level fusion is listed in Table 1. There is no significant difference in performance, as compared to the unimodal video-based classifier. The radar input either just introduces noise into the feature extraction process or the model seemingly learns to ignore the radar input.

4.2 | Feature-level fusion

The feature-level fusion architecture was implemented by concatenating the features obtained after the convolutional stages from the unimodal architectures. The features from the 'Flatten' layers from Tables 1 and 2 were concatenated. The last layers of the combined architecture were the same as the layers

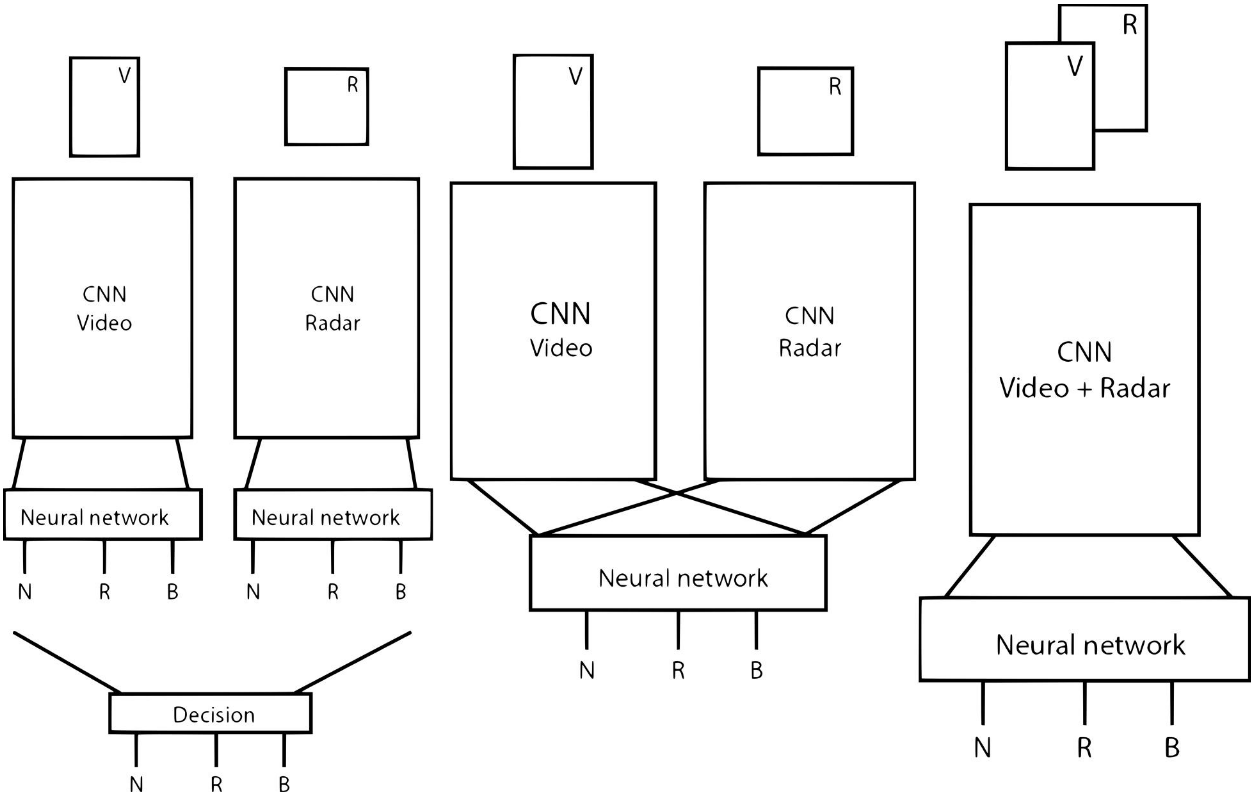

FIGURE 6 Diagrams of three multimodal fusion approaches. Here CNN indicates the convolutional and pooling layers of the CNN, used for feature extraction, whereas neural network refers to the fully connected layer(s), including a softmax layer, used for the overall classification. The three output classes are denoted by N, R and B. CNN, convolutional neural network

after the 'Flatten' layer from the unimodal architecture in Table 2. This model was then trained and validated from scratch. The classification accuracy achieved with feature-level fusion is also presented in Table 3. Considering the two-class (N and R) classification problem, the classification accuracy improves slightly as compared to the accuracy of the unimodal video-based classifier.

4.3 | Decision-level fusion

As aggregation strategy for decision-level fusion, the average of the individual classifiers was used. The CNN implementations discussed in Section 3.1 were used for the single modalities. The class predictions from the softmax activations from Tables 1 and 2 were averaged. The classification accuracy obtained with decision-level fusion is given in Table 3. Considering the two-class (N and R) classification problem, the accuracy is again slightly improved as compared to the accuracy of the unimodal video-based classifier.

Considering the three-class (N, R, and B) classification problem, the radar data does not seem to have much added value. The classification accuracy of the different multimodal approaches is similar to the accuracy of the unimodal video-based classifier. As stated earlier, the radar-based classifier is unable to distinguish the N and B classes. In the multimodal approach, the radar-based classifier does, however, aid more robust classification of the R class. This is illustrated by the confusion matrices in Figure 7. Compared to the unimodal video-based classifier, the multimodal approach performs similar on the N and B classes. The classification of the R class is, on the other hand, slightly improved from 96% to 99%.

4.4 | Missing modality

The results presented in the previous sections were obtained after training and validation with both modalities. However, one of the starting points for this study was the notion that the overall classification performance can be maintained if one of the modalities is absent or delivers data of degraded quality. To evaluate the robustness against a missing modality, the feature-level fusion classifier was also trained with an incomplete training set, that is, a training set in which one of the modalities is occasionally missing. During training either the spectrogram or the video frame was removed with a one-third chance. Validation has been performed with a complete validation set, but also with unimodal validation sets. The overall classification results are shown in Table 4.

As can be seen, the overall classification performance when validated with both video and radar data is similar, irrespective of the training set used. What is, however, remarkable, is that the overall classification performance improves drastically after training with an incomplete training set if only radar data are

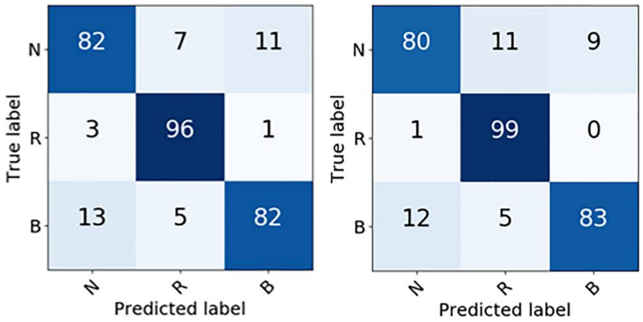

FIGURE 7 Confusion matrices for the unimodal video-based classifier (left) and for decision-level fusion (right). The classes are N for test subject strolling, R for a test subject holding an object and B for a test subject carrying a backpack. Results are in %

used for validation. This improvement comes at the cost of a somewhat degraded classification performance if only video data are used for validation. At the same time, these results should be put in perspective. The classification accuracies, when using only radar data for validation (86.7% and 62.2%), are similar to the performance obtained with the unimodal radar-based classifier (cf. Table 3: 88.9% and 62.6%). Thus, if one of the modalities is occasionally absent, a multimodal approach has the potential to improve overall performance. If, on the other hand, one of the modalities is structurally absent, for example, at night when the video camera delivers imagery of degraded quality, it is better to apply a unimodal classifier.

4.5 | Discussion

Several options related to the fusion depth of radar and video data were investigated. Both feature-level and decision-level fusion show the possibility to improve the classification accuracy. The feature-level fusion approach showed that the models can be made more robust against missing modalities by adapting the training phase (e.g. randomly dropping samples of one of the modalities from the training set). Feature-level fusion is expected to improve the classification performance when the activities are better resolved using the correlation of the individual modalities. This was not found to be the case for a person carrying a backpack.

Overall, the data from the video recordings were found to be leading in the classification process, which is probably also related to the architecture design. The feature vector obtained from the video data is larger than the vector for the radar data this is likely to introduce a bias towards the video data. It is expected that the classification performance can be improved by further optimising the multimodal architecture. As the N and B class spectrograms were quite similar it might be beneficial to use for instance a decision tree to first resolve the N and R classes and subsequently, in case of class N, just use the video data to resolve between the N and B classes.

Moreover, the current approach uses only single video frames that are associated with 1.28 s of radar data. During this time, however, multiple video frames are available. Ideally all data collected up to a certain point in time are used for classification simultaneously. This is the topic of the next sections.

TABLE 4 Feature-level fusion classification performance (%), for training and validation with complete and incomplete sets

	Validation set	{N,R}	{N,R,B}
Complete training set	Radar only	74.0	46.2
	Video only	96.2	87.2
	Radar + video	96.3	86.9
Incomplete training set	Radar only	86.7	62.2
	Video only	92.0	86.8
	Radar + video	95.3	87.2

5 | LSTM-BASED UNIMODAL LEARNING

The downsides of the CNN-based classification are: the fixed size of the inputs, the fact that sequential data are considered independently, and the delay in the classification process. Ideally once new data are obtained, that is, the result from applying an STFT on a new segment of radar data or a new video frame, these data are classified instantly. A prediction needs to be made for each segment of input data. In principle LSTM neural network architectures can do this. For each input a synchronous output is produced. This setup allows to make predictions on variable length sequences while still obtaining a prediction at each time step. The accuracy depends on how well each time step is classified. As soon as the first STFT result is extracted a classification is performed of the current activity. As time progresses and more information becomes available, the classification should become more accurate.

5.1 | LSTM background

RNNs are a type of neural networks that are designed to process sequential data. RNNs make use of an internal state (memory) to process sequences of input data. The state is an accumulation of all data that the network has seen. LSTM is a type of RNN that bypasses the vanishing gradient problem, allowing it to use and remember relevant information over a long duration.

5.2 | CNN–LSTM architectures

Inspired by [24], a hybrid CNN–LSTM architecture is used as presented in Figure 8. First the CNN–LSTM-based models for the radar data and video data are optimised separately. The radar CNN–LSTM and the video CNN–LSTM are presented in Tables 5 and 6, respectively. The tables indicate the architecture for a single timestep; however, the models are trained such that a variable length input sequence can be presented and a prediction is made for each timestep. The convolutional layers for the video

CNN–LSTM are identical to the video unimodal CNN architecture in Table 2 up to the 'Flatten' layer. For the combined architecture, as presented in Figure 8, the 'Flatten' layers from the unimodal CNN–LSTM architectures are concatenated. The last layers are an LSTM layer with 512 neurons and a softmax activation function for the prediction of the class probabilities.

The video-based CNN-LSTM classifier is depicted by the upper feed of the multimodal CNN-LSTM architecture. The radar-based CNN-LSTM classifier is represented by the lower feed of the multimodal architecture. Note that the input to the radar-based CNN is now the result of the STFT, that is, a vector of length 64.

The same measurements were used for training and validation of the CNN–LSTM architectures as for the CNN architectures discussed in Section 3. However, for training and validation of the CNN architectures, the data set was limited to just the cases where both a radar spectrogram and a related video frame were available. For the CNN–LSTM approach all data were used. Missing video frames, that is, video frames in which the SSD did not detect a person, were replaced by black frames. During training randomly selected time sequences of 1.28 s were extracted and fed to the CNN-LSTM model. Validation is performed for all sequences in the validation set from the start to the end of a measurement. Each input is classified instantaneously, but past information is taken into account for the classification. The classification accuracy is the percentage of inputs that is correctly classified.

5.3 | Respective unimodal performance

The classification accuracy of the unimodal CNN–LSTM classifiers is listed in Table 7. As compared to the performance of the unimodal CNN classifiers, discussed in Section 3.3, the performance of the radar-based classifier is better, in particular for the three-class (N, R, and B) classification problem.

Figure 9 shows a complete radar measurement with the spectrogram and the related softmax score over time. As the measurement starts the CNN–LSTM-based classifier can quickly predict the activity correctly. Within half a gait cycle an accurate prediction is made of the activity.After this short initialisation, the spectrogram is classified almost correctly throughout time, apart from some interruptions between 13 s and 15 s where the correct class score is momentarily very low. It is not entirely clear why this happens, but at this point the clutter return is a bit stronger compared to the other regions which might interfere with the classification.

Other errors that are observed mainly seem to originate from people walking with their hands in their pockets and/or barely moving their arms while walking, these cases are then classified as the R class as no arm motion is observed. For most persons the classification is accurate throughout the measurement although for some persons the fluctuation of the class label is worse than the example in Figure 9.

As compared to performance of the unimodal CNN classifiers, the performance of the video-based classifier is worse. This is due to the video frames at the start of the measurements in which the SSD could not detect a moving person. These video frames were replaced by black frames, resulting in low classification accuracy for those frames. Recall that these bad video frames were deleted from the training set used for the CNN-based classifiers.

Similar to Figure 9, the classification score over time is shown in Figure 10 for the video-based CNN–LSTM classifier. The figure shows an uncertain classification at the start of the measurement, the softmax classification score is around 0.5, this is due to the fact that up to that point no detections have been made in the video feed and therefore only black frames are presented up to this point. As soon as a frame is detected around 2.7 s a correct classification is made of the activity,

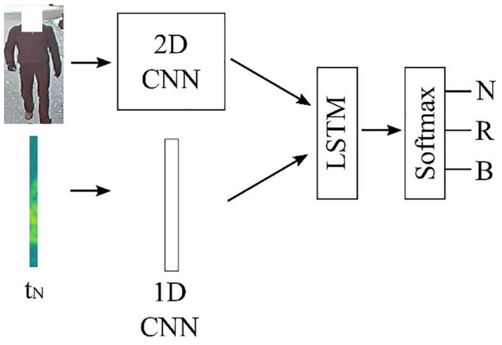

FIGURE 8 Multimodal CNN–LSTM architecture. The input on the left: the video frames and the result of a single STFT window. The output classes are indicated by N, R and B: CNN, convolutional neural network; LSTM, long short-term memory, STFT, short-time Fourier transform

TABLE 5 Radar CNN-LSTM architecture

Layer	Number of filters	Kernel size	Dimensions
Radar input	-	-	64 × 1
Convolution	8	5 × 1	64 × 8
Max pooling	-	2 × 1	32 × 8
Convolution	16	3 × 1	32 × 16
Max pooling	-	2 × 1	16 × 16
(Flatten)	-	-	256
LSTM	-	-	512
Softmax	-	-	2/3

Abbreviations: CNN, convolutional neural network; LSTM, long short-term memory.

TABLE 6 Video CNN–LSTM architecture

Layer	Number of filters	Kernel size	Dimensions
Video input	-	-	128 × 64 × 3
Video CNN stages	Layers from input up to 'Flatten' layer from Table 2.		
(Flatten)	-	-	4096
LSTM	-	-	512
Softmax	-	-	2/3

Abbreviations: CNN, convolutional neural network; LSTM, long short-term memory.

		{N, R}	{N, R, B}
Single modality classification	Radar	91.2	73.0
	Video	91.5	81.0
Video/radar multimodal classification	Validation with radar data	78.6	65.8
	Validation with video data	86.5	75.2
	Multimodal validation	94.3	86.5

Abbreviations: CNN, convolutional neural network; LSTM, long short-term memory.

TABLE 7 CNN–LSTM-based classification performance (%), for validation with complete and incomplete sets

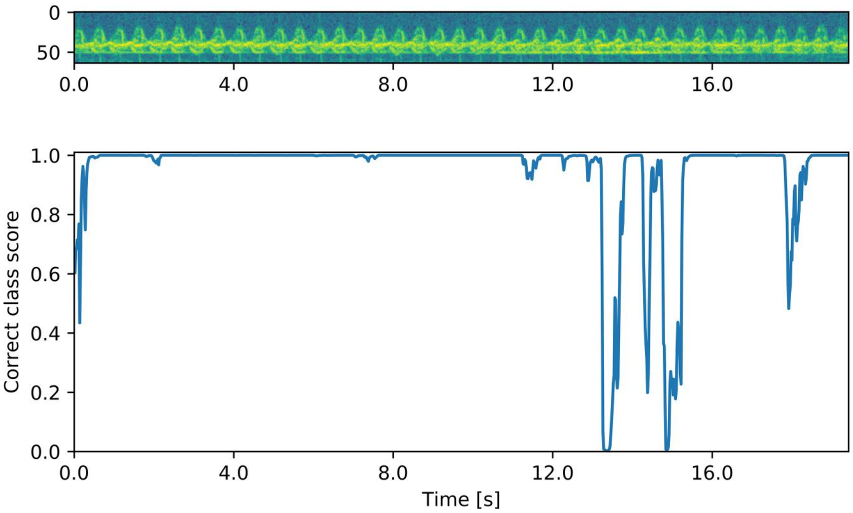

FIGURE 9 Classification performance (correct class score) of the radar-based CNN–LSTM architecture for a single radar measurement. On top the spectrogram of the complete measurement, at the bottom the correct class score. CNN, convolutional neural network; LSTM, long short-term memory

although beyond 2.7 s there are also missed detections in the feed (which are substituted by black frames) the classification is accurate throughout the rest of the observation. As the persons walked from 40 m towards the radar and video sensors during the measurements, the detection rate on the video frames worsen at larger distances. Although quite often, once a detection is made the classification is also accurate. The main source of error for this video classification approach is therefore the missing frames at the start of the measurements.

5.4 | Discussion

The hybrid CNN–LSTM architectures show potential to outperform the CNN-based classifiers. However, a direct performance comparison is difficult since different training and validation sets have been used (bad video frames were deleted from the sets used for the CNN-based classifier). For a part the classification accuracy might be improved further by optimising the CNN–LSTM architecture.

6 | LSTM-BASED MULTIMODAL LEARNING

The proposed multimodal CNN–LSTM architecture is depicted in Figure 8. The individual architectures are the same as discussed in the previous section. Each new STFT result is fed to the radar-based CNN and then to the LSTM network. In parallel, the most recent detected video frame, related to the time window of the STFT, is fed to the video-based CNN and subsequently to the LSTM network. This procedure is illustrated in Figure 11. The red window at time t_N illustrates the combined video and STFT result. In this way the most recent data are used for the classification. Video frames are not always available, for those cases a black frame is fed to the video feed, as illustrated at time t_M. The LSTM model should recognise these frames and is able to remember the necessary features from past frames. This way not only the information about objects can be used but also the sequence of the video feed can be exploited.

6.1 | Multimodal CNN–LSTM analysis

The classification performance of the multimodal CNN–LSTM architecture is also listed in Table 7. When comparing these results with the unimodal classification results, it can be observed that the classification performance for the classification of the N, R and B classes is improved by a few percent. The main contribution in this improvement is the more accurate classification at the start of the measurements, that is, when the SSD cannot detect the moving person in the video frames due to the unfavourable lighting conditions.

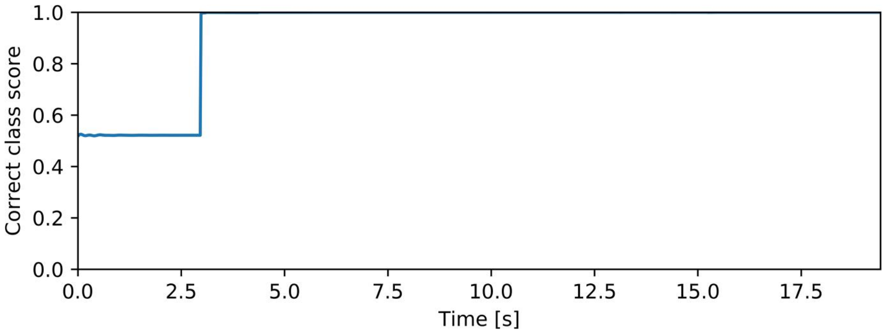

FIGURE 10 Classification performance (correct class score) of the video-based CNN-LSTM classifier as function of time for a complete measurement. CNN, convolutional neural network; LSTM, long short-term memory

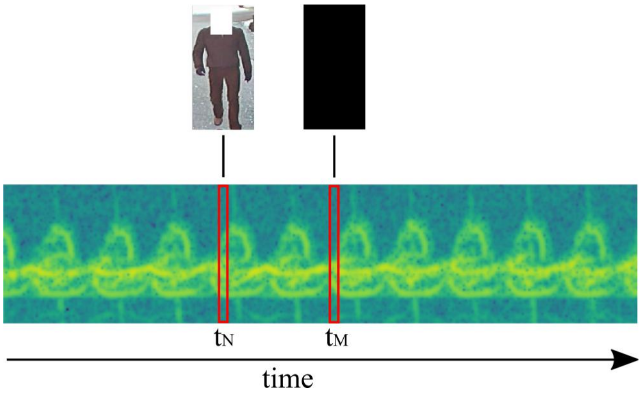

FIGURE 11 Data association of the radar and video frames. At the indicated time instances the most recent frame is associated with the STFT result. STFT, short-time Fourier transform

6.2 | Discussion

For some measurements, it takes several seconds before a moving person is detected by the SSD, due to the lighting conditions. During this time no reliable prediction can be made of the activity. Once a frame has been detected, however, the model is able to remember relevant past features and use them effectively for the classification of future frames. To what extent past features are used effectively may be optimised by training the model with longer or shorter time sequences. Although the current activities are mainly defined by identifying objects in the video frames. The CNN–LSTM model is in principle capable of exploiting the changes from frame to frame as well for the classification.

Being aware that someone was carrying an object or performing a certain activity can help in classifying the behaviour in the future. Especially the multimodal setup can benefit from this information. To further assess this notion, a data set is needed that includes transitions from one activity to another. With such a data set it can be investigated how heavily the LSTM model relies on the past information, this might be optimised by adjusting the length of the sequences during the training procedure.

For the radar-based CNN–LSTM classifier, a CNN stage was used to transform the input features. Alternatively, the CNN stage could be omitted and the result of the STFT can be directly fed to the LSTM stage. However, the feature extraction stage seems to slightly improve the multimodal classification performance, this can, however, also happen due to the increased size of the feature vector.

7 | CONCLUSION

Here, several multimodal learning methods have been demonstrated for the purpose of human activity classification based on radar and video data. First, an analysis was made of different fusion depths: data-level, feature-level and decision-level. Both feature-level and decision-fusion approaches show the possibility to improve classification accuracies. The disadvantage of the approach was that each frame was considered independently and 2D convolutional layers were used for the recognition of the activity in 1.28 s spectrogram excerpts.

The second approach demonstrated the use of a multimodal hybrid CNN–LSTM model to classify the human activity continuously. The result of a 0.1 s STFT and a single video frame was used for classification at each time instance. Unimodal implementations of hybrid CNN–LSTM models showed that both the radar and video sequences can be classified continuously with a high degree of accuracy. The multimodal classification accuracy improves by a few percent, although this improvement is for a large part contributed to the ability to classify missed detections in the video feed at the start of the measurement runs.

The CNN–LSTM model should be able to do transition between activities more rapidly than compared to the 2D convolutional approach as it can identify the necessary features that are related to changing the activity. However, to investigate this effect for human activity classification a data set is required that contains test subjects changing their activity or behaviour during the measurements. In addition, a more challenging data set including measurements in more realistic and varying (light) conditions should be obtained to test the robustness of the multimodal approach and show the added value.

ORCID

Faruk Uysal https://orcid.org/0000-0002-4518-5649

REFERENCES

1. Patel, N., et al.: Sensor modality fusion with CNNs for UGV autonomous driving in indoor environments. In: Proceedings of IEEE/RSJ

International Conference on Intelligent Robots and Systems (IROS), pp. 1531–1536, Vancouver (2017)
2. Neverova, N., et al.: ModDrop: adaptive multi-modal gesture recognition. IEEE Trans. Pattern Anal. Mach. Intell. 38(8), 1692–1706 (2016)
3. Abavisani, M., et al.: Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach (2019)
4. Nglam, J., et al.: Multimodal deep learning. In: Proceedings of International Conference on Machine Learning (ICML), Bellevue (2011)
5. Tatulli, E., Hueber, T.: Feature extraction using multimodal convolutional neural networks for visual speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP), pp. 2971–2975, New Orleans (2017)
6. Yasui, Y., et al.: Multimodal speech recognition using mouth images from depth camera. In: Proceedings of Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1233–1236, Kuala Lumpur (2017)
7. Vidanapathirana, M.: Real-time Human Detection in Computer Vision – Part 2. (2018). https://medium.com/@madhawavidanapathirana/real-time-human-detection-in-computer-vision-part-2-c7eda27115c6 Accessed 10 January 2019
8. Kim, Y., Moon, T.: Human detection and activity classification based on micro-Doppler signatures using deep convolutional neural networks. IEEE Geosci. Rem. Sens. Lett. 13(1), 8–12 (2016)
9. Trommel, R., et al.: Multi-target human gait classification using deep convolutional neural networks on micro-Doppler spectrograms. In: Proceedings of EuRAD, pp. 81–84, London (2016)
10. Amin, M.G., Erol, B.: Understanding deep neural networks performance for radar-based human motion recognition. In: Proceedings of IEEE Radar Conference, pp. 1461–1465, Oklahoma City (2018)
11. Li, X., et al.: A survey of deep learning-based human activity recognition in radar. Rem. Sens. 11(9) (2019)
12. Quan, D., et al.: Deep generative matching network for optical and SAR image registration. In: Proceedings of IEEE Geoscience and Remote Sensing Society (IGARSS), pp. 6215–6218, Valencia (2018)
13. Hoffman, S., et al.: Registration of high resolution SAR and optical satellite imagery using fully convolutional networks. In: Proceedings of IEEE Geoscience and Remote Sensing Society (IGARSS), pp. 5152–5155, Yokohama (2019)
14. Hu, J., et al.: FusioNet: a two-stream convolutional neural network for urban scene classification using PolSAR and hyperspectral data. In: Proceedings of Joint Urban Remote Sensing Event (JURSE), Dubai (2017)
15. Li, H., et al.: Magnetic and radar sensing for multimodal remote health monitoring. IEEE Sensor. J. 19(20), 8979–8989 (2019)
16. Molchanov, P., et al.: Multi-sensor system for driver's hand-gesture recognition. In: Proceedings of IEEE Automatic Face and Gesture Recognition (FG), Ljubljana (2015)
17. Liu, X., et al.: Multimodal-temporal fusion: blending multimodal remote sensing images to generate image series with high temporal resolution. In: Proceedings of IEEE Geoscience and Remote Sensing Society (IGARSS), pp. 10083–10086, Yokohama (2019)
18. Aydogdu, C.Y., et al.: Multi-modal cross learning for improved people counting using short range FMCW radar. In: Proceedings of IEEE International Radar Conference, Washington (2020)
19. de Jong, R.J.: Multimodal Deep Learning for The Classification of Human Activity. Master Thesis, Delft (2019)
20. de Jong, R.J., et al.: Radar and video multimodal learning for human activity classification. In: Proceedings of SEE International Radar Conference, Toulon (2019)
21. Loukas, C., et al.: Activity classification using raw range and I&Q radar data with long short term memory layers. In: Proceedings IEEE Digital Avionics Systems Conference (DASC), pp. 441–445, Athens (2018)
22. Sadreazami, H., et al.: On the use of ultra wideband radar and stacked LSTM-RNN for at home fall detection. In: Proceedings of IEEE Life Sciences Conference (LSC), pp. 255–258, Montreal (2018)
23. Li, X., et al.: LSTM based human activity classification on radar range profile. In: Proceedings IEEE International Conference on Computational Electromagnetics (ICCEM), Shanghai (2019)
24. Zhu, J., et al.: A hybrid CNN-LSTM network for the classification of human activities based on micro-Doppler radar. IEEE Access. 8, 2169–3536 (2020)
25. Li, H., et al.: Bi-LSTM network for multimodal continuous human activity recognition and fall detection. IEEE Sensor. J. 20(3), 1191–1201 (2020)
26. Otten, M.P.G., et al.: Light weight digital array SAR. In: Proceedings of IEEE ARRAY, pp. 177–182, Waltham (2010)
27. Heiligers, M.J.C., et al.: Deep learning for automatic target recognition with radar. In: Proceedings of EuRAD (special session), pp. 38–41, Madrid (2018)
28. Hyung, E.-J., et al.: Influence of load and carrying method on gait, specifically pelvic movement. J. Phys. Ther. Sci. 28, 2509–2062 (2016)
29. Liu, W., et al.: SSD: Single Shot Multibox Detector. arXiv(2016). https://arxiv.org/pdf/1512.02325.pdf Accessed 26 March 2020
30. Goodfellow, I., et al.: Deep Learning. (2016). http://www.deeplearningbook.org Accessed 18 September 2020
31. Chattopadhyay, A., et al.: Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe (2018)
32. Baltrušaitis, T., et al.: Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. 41(2), 423–443 (2019)

How to cite this article: de Jong, R.J., de Wit, J.J.M., Uysal, F.: Classification of human activity using radar and video multimodal learning. IET Radar Sonar Navig. 15(8), 902–914 (2021). https://doi.org/10.1049/rsn2.12064