Radar and video multimodal learning for human activity classification

de Jong, Richard J.; Uysal, Faruk; Heiligers, Matijs J.C.; de Wit, Jacco

DOI
10.1109/RADAR41533.2019.171283

Publication date
2020

Document Version
Accepted author manuscript

Published in
2019 International Radar Conference (RADAR)

Radar and Video Multimodal Learning for Human Activity Classification

Richard J. de Jong
Microwave Sensing, Signals & Systems
Delft University of Technology
Delft, The Netherlands
richard.dejong@tno.nl

Matijs. J.C. Heiligers
Department of Radar Technology
TNO
The Hague, The Netherlands
matijs.heiligers@tno.nl

Jacco J.M. de Wit
Department of Radar Technology
TNO
The Hague, The Netherlands
jacco.dewit@tno.nl

Faruk Uysal
Microwave Sensing, Signals & Systems
Delft University of Technology
Delft, The Netherlands
f.uysal@tudelft.nl

***Abstract*—Camera systems are widely used for surveillance in the security and defense domains. The main advantages of camera systems are their high resolution, their ease of use, and the fact that optical imagery is easy to interpret for human operators. However, particularly when considering application in the defense domain, cameras have some disadvantages. In poor lighting conditions, dust or smoke the image quality degrades and, additionally, cameras cannot provide range information. These issues may be alleviated by exploiting the strongpoints of radar. Radar performance is largely preserved during nighttime, in varying weather conditions and in dust and smoke. Furthermore, radar provides range information of detected objects. Since their qualities appear to be complementary, can radar and camera systems learn from each other? In the current study, the potential of radar/video multimodal learning is assessed for the classification of human activity.**

Keywords—multimodal learning, radar, video, human activity classification, micro-Doppler

I. INTRODUCTION

Classification of human activity is an important asset in the defense and security domains. The activity or behavior a human exhibits may (partly) reveal someone's intent. Someone strolling on a parking lot may be just on his way to his car or someone may be scanning the cars for possible valuable items to steal. Physical behavior of a person, e.g., walking speed and walking pattern, may reveal the actual intent.

In the civil security domain, cameras are widely used for surveillance; CCTV systems can be found in city centers, in malls, on parking lots, in train stations, 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.

Cameras do have some disadvantages, in particular when considering defense applications. The quality of (daylight) camera imagery degrades in poor lighting conditions, smoke and dust. Furthermore, an image sensor does not provide information about the range to a subject.

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

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

In literature multimodal learning has been used in different applications. For example the fusion of video data with laser range measurements for autonomous navigation [1]. Multimodal learning is also used to achieve robust speech recognition using the video data only in absence of the audio signal by fusing the video data of a speaking person and the related audio signal [2]-[4]. Extending the previous approaches, in this paper, we particularly investigate the use of multimodal learning for radar and optical sensors for human activity classification.

In the current study it was investigated whether the classification of human activity can be improved when feeding corresponding video and radar data to a neural network based classifier, as compared to performing classification using either the video data or the radar data. If indeed there is some performance improvement using a multimodal neural network, 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 (for instance during the nighttime 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 neural network for classification.

The concept of multimodal learning is explained in more detail in the following section. Subsequently, in Section III the radar and video measurements and the resulting data sets are described. These data sets were used to assess the potential of radar/video multimodal learning. The design of the multimodal network architecture and the obtained results are discussed in Section IV. Finally, the conclusion is given in Section V.

II. Multimodal Learning

Modality refers to the way the environment or events are perceived. Different types of sensors, e.g., acoustic, optical, and RF sensors, perceive the environment in different ways and are therefore referred to as different modalities. Multimodal convolutional neural networks (CNNs) are networks that are able to jointly interpret data from different modalities [5]. Within a network architecture, the integration of the data coming from various modalities may be done at different levels. The different levels of integration considered in this work are presented in Fig. 1.

The architecture on the left depicts decision-level fusion. The radar and video data are essentially considered independently resulting in two classification results. Only a final stage is added in which the individual classification results are aggregated in some way to obtain the overall classification result. By using this decision-fusion architecture, the individual CNNs for the video and radar data are trained and validated separately and thus 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 obtain features, but all features are then fed to a single (fully-connected) neural network to obtain the overall classification. By using this architecture, the individual radar and video data CNNs may learn from each other in the back propagation stage, if the weights are adapted on the basis of the overall classification result.

Finally, the architecture on the right depicts data-level fusion. By using the data-fusion architecture, the video and radar data are simultaneously fed to a single CNN. Potentially, data-fusion allows deep exploitation of the correlation (or complementarity) between the different modalities. A drawback of this architecture is that the video and radar images need to be of the same size (expressed in image pixels). Another drawback of the architecture is that it requires the convolutional stages to have the same parameters which is not necessarily the optimal feature extraction process for the individual modalities.

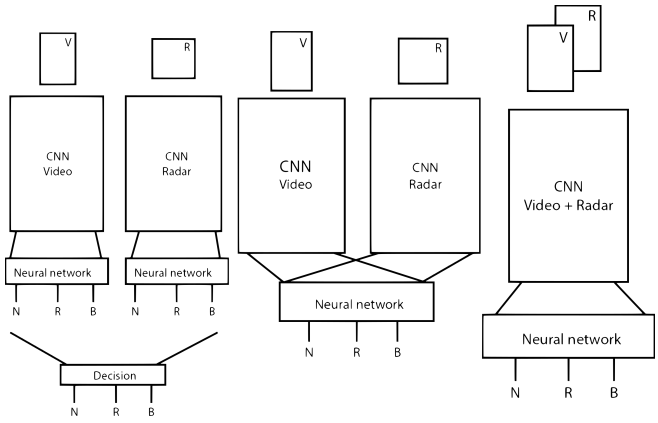

Fig. 1. Schematic block diagrams of three multimodal fusion approaches. Here "CNN" indicates the convolutional and pooling layers of the CNN used for feature extraction, whereas "Neural network" refers to the fully connected layer(s) (including a softmax layer) used for the final classification. The three output classes are denoted by "N," "R," and "B."

To assess the added value of video/radar multimodal learning, it should be analyzed what information is actually exploited for feature extraction by the individual CNNs. Do the CNNs exploit complementary information or do they exploit similar information from the video and radar data? In the latter case, the added value of the multimodal approach may be limited. For this assessment the gradient-weighted class activation mapping (Grad-CAM++ [6]) visualization method is applied. With the aid of visualization methods, such as Grad-CAM++, a saliency or heat map can be generated, which contains a coarse localization of the parts of the feature maps which contribute to the class score. A saliency map mitigates the black-box nature of CNNs and highlights the image pixels that are actually used for feature extraction and classification, e.g., [7], [8].

III. Measurement Setup and Data Sets

To assess the potential of multimodal learning for human activity classification, measurements were conducted with a compact radar system and a high-definition camera, see Fig. 2. The test subjects walked toward the measurement setup starting at around 40 m range, with a typical walking speed of about 1.5 m/s, resulting in measurement runs of about 20 s. The measurements were conducted on different days both in the morning and the afternoon. The weather conditions were similar during the different measurements, but the lighting conditions varied depending on the time of day.

Measurements were made of (*a*) people just strolling, i.e., walking without luggage or objects in their hands (class N), (*b*) people carrying a relatively heavy backpack (class B), and (*c*) people holding a weapon-like object with both hands (class R). These three cases (N, R and B) are assumed representative for different types of human activity, as persons carrying heavy items may be regarded suspect in particular situations (such as a person carrying a crowbar on a parking lot). It should be noted that the test subjects inclined to swing their arms when not carrying the object.

A total of thirty-five test subjects were available. Each test subject performed the activities twice; as a result 210 measurements are available for training and validation.

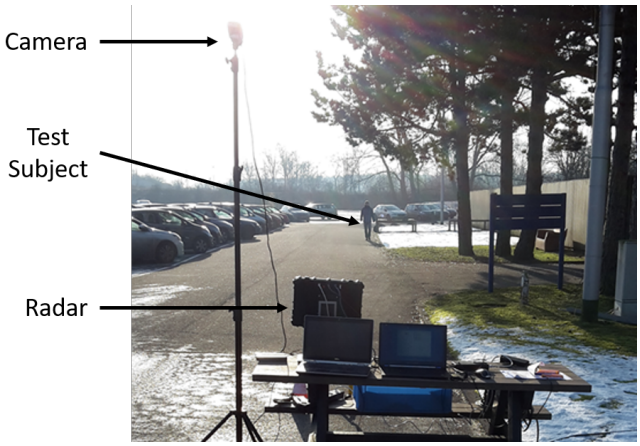

Fig. 2. Human activity measurement setup with a high-definition camera and a compact radar system.

A. Radar

For the radar measurements the AMBER frequency modulated continuous wave (FMCW) radar was used. AMBER operates in X-band and for these measurements the range resolution was 1.5 m. Since the focus is on the micro-

Doppler signatures of the test subjects, the range resolution was relatively coarse to ensure that a test subject is contained within a single range resolution cell (including swinging arms and legs). For the current study, it is assumed that the gait and body motion change if a person carries a heavy item. If this difference in gait and/or body motion can be recognized in the person's micro-Doppler signature, it can be determined whether the person carries a heavy item or hefty backpack.

To highlight the micro-Doppler signature, spectrograms were generated of the measured radar data. An example spectrogram excerpt of each activity is presented in Fig. 3. These examples include approximately one human gait cycle. From an earlier study it was already concluded that the swinging arms or the lack thereof (in the case the person is carrying an object in both hands) is the most distinguishing feature [8]. This can be seen in the examples, in the case the person is strolling or carrying the backpack, the spectrogram exhibits an 'arc' related to the (lower) arms' motion (indicated by the black arrows). When both hands of the test subject are engaged, this arc is absent (indicated by the black circle).

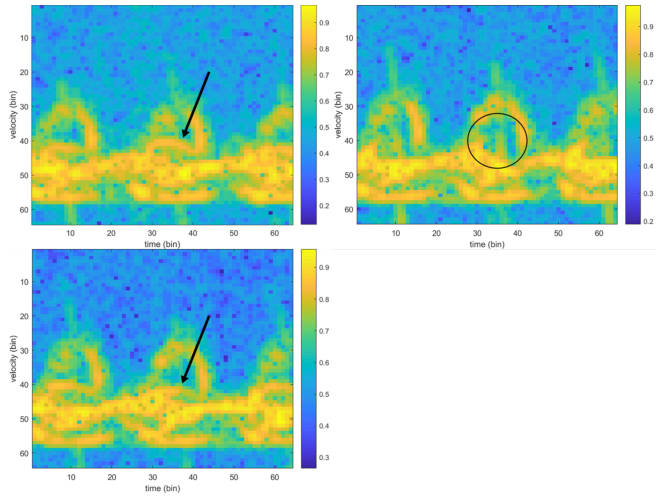

Fig. 3. Examples of measured spectrograms for the test subject strolling (N) (top left), the test subject carrying an object in both hands (R) (top right) and the test subject carrying the backpack (B) (bottom).

B. Video

The video recordings were made using a high-definition camera with a frame rate of about 13.5 frames per second. A single shot detector (SSD) [9] was applied to the individual frames in the video recordings to detect the test subjects. The area within a bounding box, i.e., the image pixels related to the person, is then extracted from the frame. Due to the test subjects walking from an initial range of 40 m toward the radar, the bounding boxes differ in size. As the test subjects approach the measurement setup, the bounding box size increases. However, for the CNN all input images need to be of the same size. Consequently, all extracted subimages were resized to a width of 64 pixels and a height of 128 pixels.

The SSD has a high detection rate (of the order of 90%) and a low false alarm rate. Nevertheless, in some frames the test subject was missed or an object was labeled as a person. These frames were deleted from the data set manually.

C. Training and Validation Set

As mentioned, spectrograms were generated from the radar measurements. From a spectrogram, an excerpt was selected of 1.28 s long, ensuring that at least a single human gait cycle is included. This spectrogram excerpt was paired with the video frame corresponding to the start time of the spectrogram excerpt. The maximum synchronization error between the start time of the spectrogram excerpt and the actual time of the corresponding video frame is 0.01 s. In total over 40,000 of such video/radar data pairs were available for training and validation. The data pairs of 28 randomly selected test subjects were used for training and the data 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.

IV. Network Architecture and Results

The application of CNNs for classification of persons or objects in pictures and video is already well-established, e.g., [10]. Recently, CNNs have also been successfully applied to classify human activity based on radar micro-Doppler signatures [7], [11], [12]. The idea of fusing video and radar data using CNNs in a multimodal setup is however novel.

Keras [13] with TensorFlow was used to implement the CNNs and the multimodal topologies. First two models were optimized for the two modalities separately, the single modality models are discussed in Section A and the related results in Section B. Saliency maps for the unimodal implementations are discussed in Section C and D. In Section E the multimodal fusion architectures and results are discussed.

A. Convolutional neural network architecture

The model parameters were devised based on a grid search. The radar CNN consists of four convolutional stages (alternately a convolutional layer and max pooling layer) with 5x5 kernels. The amount of kernels at each stage is 20-30-40-50 respectively. The (sub)optimal CNN for the video data also had four convolutional stages with kernel size of 3x3. However it uses a double convolutional layer at each stage. With the amount of kernels at each stage 16-32-64-128 respectively. Both the radar and video classifier use two fully connected layers with 500 neurons each and a softmax function is used to perform the final classification.

B. Single modality results

The classification performance of the single modality implementations are shown in Table I. Two scenarios are taken in consideration, classifying just the N and R class (the first row) and classifying all three classes. The classifier for the video outperforms the radar classifier for both scenarios.

As was previously shown in [8], the radar classifier has difficulties to distinguish the N and B class. The overall classification performance in this case is 62.6%. The confusion matrices in Fig. 4 emphasize this, the radar classifier is however still able to distinguish the R class with a high degree of accuracy, which is also the main contribution to the overall classification performance. The confusion matrix for the video data shows a high classification accuracy for the R class as well, which is most likely related to the clear visibility of the object in the video frames (see Fig. 5).

TABLE I. UNIMODAL CLASSIFICATION PERFORMANCE (RESULTS ARE IN %). COLUMN 'VIDEO' AND 'RADAR' CONTAIN THE RESULTS FOR THE VIDEO AND RADAR MODALITY RESPECTIVELY.

Set of classes	Modality	
	Video	*Radar*
{N,R}	95.2	88.9
{N,R,B}	87.0	62.6

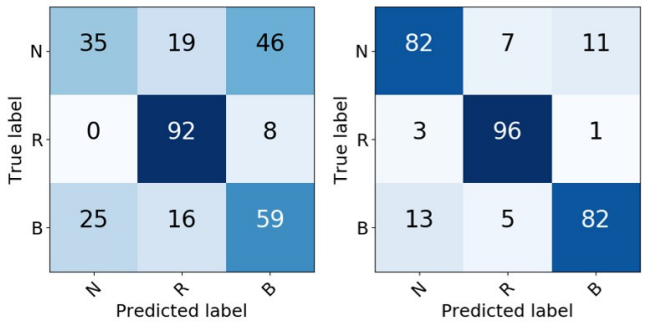

Fig. 4. Confusion matrix for the radar classifier (left) and video classifier (right). 'N' refers to a person strolling, 'R' to a person carrying an object with both hands and 'B' to a person carrying a backpack. Classification results are in %.

Due to the cold weather during the recordings some people walked stiffly with their arms besides their body, due to the reduced arm motion this behavior can result in an absence of the previously described arc in the spectrograms, this behavior slightly explains the confusion between the nothing and the rifle class. Furthermore occasionally a person grabs the strap of the backpack whilst walking, therefore reducing the arm motion.

C. *Saliency maps video*

The result of the Grad-Cam++ saliency maps has been superimposed on video frames obtained from the SSD in Fig. 5. The images are classified correctly with a certainty over 0.99. An effort is made to give a fair representation of the saliency maps observed on the data set. The first row of the saliency maps illustrates the general findings, the second row more questionable saliency maps. Red regions indicate the most relevant regions for a correct classification.

In case of the person strolling the pixels around the lower arms and hands have high saliency. This suggests that the arms hanging loose next to the torso are an important feature to classify a person strolling. In case of the person carrying the object, the pixels around the hands and the object have high saliency, indicating that the presence of the object in front of the torso is the key feature. For the person carrying the backpack the pixels around the strap of the backpack have high saliency, thus rightly contributing to the classification. The free hand also has high saliency, but is on its own not discriminative for the class activity. The same type of information is used to separate the person carrying an object or a person just strolling: are the arms loose next to the torso or engaged in some way. For these two classes the multimodal approach may have added value. The person carrying a backpack however shows similar features as the person just strolling, the backpack strap however is not a feature directly visible in the radar spectrograms (it might indirectly be observed due to a changed gait). If it is not possible to extract any information on the particular activity the multimodal approach may only have limited added value.

For the person carrying a backpack the main feature of interest in the video data is the presence or absence of the straps of the backpack. For a significantly heavy backpack the micro-Doppler signatures (see Fig. 3) are also expected to change. However, the impact of the backpack (of 10 kg) in this setup does not seem to be significant enough as there is no clear noticeable difference in the micro-Doppler components. Which is emphasized by the inability of the radar classifier to distinguish the Nothing and Backpack class.

The second row of images in Fig. 5 are also classified correctly, in these images the feet are seemingly a relevant feature. It is unclear how the feet can contribute to the classification and whether this is a desired property, although some frames have been cropped (by the SSD) such that the feet are no longer visible, this can cause the feet to contain some information about the position of the person and the objects in question. Furthermore the saliency map on top of the person carrying an object shows that the background has high saliency. This does however not contain any information about the activity and is therefore an undesired property.

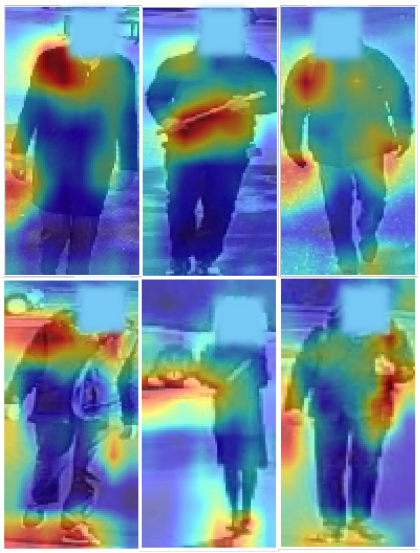

Fig. 5. Saliency maps with a correct classification (certainty over 0.99) superimposed on the corresponding video frame, for a person just strolling (left column), a person carrying an object with both hands (middle column) and a person carrying a backpack (right column).

D. *Saliency maps radar*

In Fig. 6, measured spectrograms and corresponding saliency maps are shown of a person just strolling with his arms swinging and a person carrying an object with both hands. These results were obtained for a CNN trained to distinguish between these two classes only for the current implementation (right two columns) and a CNN with a larger last convolutional layer (left two columns). As it was also known from the earlier study, [8], that it is difficult to classify the person carrying a backpack on the basis of radar spectrograms (including the backpack class resulted in the overall saliency maps to look like noise), this class was omitted from this assessment. In case of the strolling person, the area where the arc of the moving arm is (cf. Fig. 3) has high saliency, which is most clearly visible in the top left figure in Fig. 6. In case of the person carrying the object, the

response to the torso has high saliency. This assessment confirms the notion that the arm motion or lack thereof is the key feature to distinguish a person just strolling from a person carrying an object 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, the resulting saliency maps are not always clear as is illustrated by the two columns at the right.

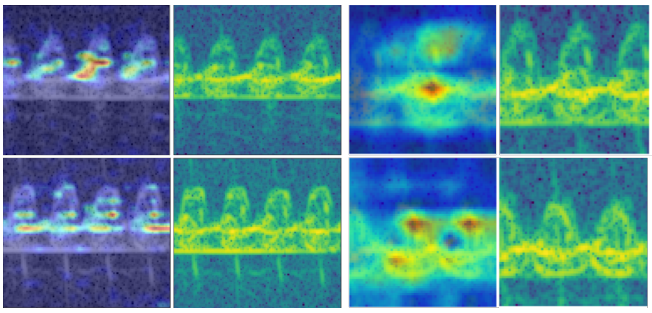

Fig. 6. Saliency maps superimposed on the spectrograms. Right two columns show the results for the current implementation and the left two columns for an architecture with larger feature maps at the last convolutional layer. Images show a person just strolling (N)(top row) and a person carrying an object with both hands (R)(bottom row).

E. Multimodal classification performance

The results of the multimodal fusion strategies are presented and discussed in the order of the fusion depth in the following sections.

1) Data-level fusion

The data-level fusion architecture was implemented by adding the spectrogram excerpt as an additional channel to the video CNN. The results of this procedure are presented in Table II. There is no significant difference between the implementation of this early fusion strategy and the classifier for just the video input (see Table I). The radar input either just introduces noise into the feature extraction process or the model seemingly learns to ignore the radar input.

TABLE II. DATA-LEVEL FUSION CLASSIFICATION PERFORMANCE. RESULTS ARE IN %.

Set of classes	Classification Accuracy
{N,R}	95.5
{N,R,B}	87.0

2) Feature level fusion

The feature level fusion architecture was implemented by concatenating the single modality implementations after feature extraction and training and validating this model from scratch. In Table III the overall classification accuracy of the feature-level fusion is presented when the model is trained with distorted data (removing either the radar or video frame during training with a 1/3 chance) or when both modalities are present (without distortion). The columns 'Vid' (Video) and 'Rad' (Radar) indicate whether the video or radar, or both, are used in the validation stage.

When dropping modalities ('with distortion' column) the relevance of the individual modalities is similar to the unimodal trained models, this shows the possibility to make such neural networks more robust to missing modalities. In case the models are trained with both modalities present ('without distortion' column) the overall architecture seems to prefer the data from the video as removing it only reduces the classification by 0.1%. More interestingly the performance of the model trained with both modalities present for the three class classification (set {N,R,B}) shows an improvement in classification accuracy when the radar input is removed, it improves from 86.9% to 87.2%.

In Fig. 7 the confusion matrices for the feature-level fusion model validated with a single modality or both modalities present are presented. As the radar classifier part is not able to distinguish the N and B class the overall classification accuracy reduces. This suggests that the radar model trained for all three classes learns bad features from the data and therefore starts introducing noise into the model leading to a worse classification accuracy. Most likely the radar model trained on all three classes starts to recognize the peculiarities of the training data set such as noise and specific human gaits.

Overall the data from the video recordings was found to be dominant in the classification process, which is likely 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 expected to introduce a bias towards the video data. It is expected that the classification performance can be further improved by further optimizing the multimodal architecture.

TABLE III. FEATURE-LEVEL FUSION CLASSIFICATION PERFORMANCE VALIDATED WITH JUST VIDEO 'VID', RADAR 'RAD' OR BOTH (VID+RAD). RESULTS ARE IN %.

Set of classes	Validation data					
	Without distortion			With distortion		
	Vid+Rad	*Vid*	*Rad*	*Vid+Rad*	*Vid*	*Rad*
{N,R}	96.3	96.2	74.0	95.3	92.0	86.7
{N,R,B}	86.9	87.2	46.2	87.2	86.8	62.2

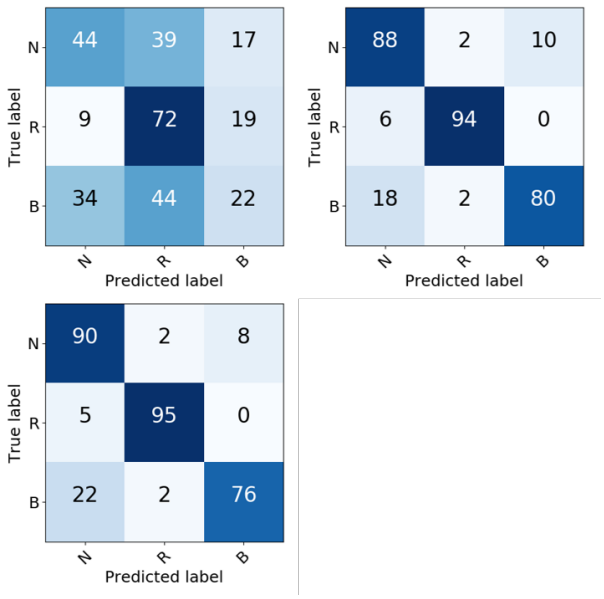

Fig. 7. Classification results of feature-level fusion without distortion. Top left: validated with just radar data. Top right: validated with just video data. Bottom left: validation with both modalities. The classes are: 'N' for person strolling, 'R' for a person carrying an object with both hands and 'B' for a person carrying a backpack. The numbers are percentages.

3) Decision level fusion

As aggregation strategy for the decision-fusion method in Fig. 1 the average of the individual classifiers is used. The models from Table I were used for the single modalities. The results of this strategy as well as the unimodal results are presented in Table IV for comparison. For just the N and R class the overall classification improves by 1.2% up to 96.4%. A small improvement is observed for the three class classification problem, as the radar classifier is not able to distinguish the N and B class it does not contribute to the classification for these classes. It does however contribute to the classification for the R class which is depicted in the confusion matrix in Fig. 8. The R class is identified in 99% of the cases, albeit there is some confusion. There is however no significant difference in the classification performance of the feature and decision-level fusion methods. A better solution might be to use e.g. a decision tree to first resolve the N and R class (allowing for more optimal feature extraction) and then use just the video data to resolve between the N and B class.

TABLE IV. CLASSIFICATION ACCURACY OF DECISION LEVEL FUSION. THE RESULTS FROM THE UNIMODAL CLASSIFIERS ARE ALSO STATED. RESULTS ARE IN %.

Set of classes	Unimodal		Decision Fusion
	Video	*Radar*	*Video + Radar*
{N,R}	95.2	88.9	96.4
{N,R,B}	87.0	62.6	87.3

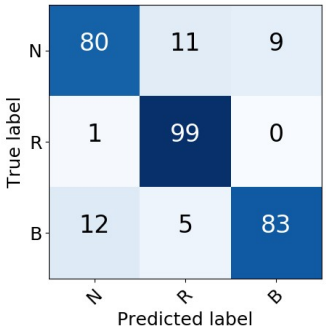

Fig. 8. Confusion matrix results of the multimodal architecture with decision-level fusion. The classes are: 'N' for person strolling, 'R' for a person carrying an object with both hands, and 'B' for a person carrying a backpack. The numbers are percentages.

V. CONCLUSION

Unimodal implementations showed that both single video frames and radar spectrogram excerpts can be used to discriminate a person just strolling and carrying an object with both hands up to 95.2% and 88.9% respectively. On basis of just spectrogram excerpts it was not possible to differentiate a person carrying a backpack clearly.

Several possibilities related to the fusion depth of radar and video multimodal 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 or distorted modalities by adapting the training phase by taking into account the missing or distorted modalities. Feature level fusion is expected to improve the classification performance when the activities are better resolved using the correlation of the single modalities. This was not found to be the case for a person carrying a backpack.

Furthermore Grad-CAM++ saliency maps were used to identify the relevant features for the different modalities. In case of the radar spectrograms the presence or absence of the micro-Doppler component related to the lower arms motion was identified as key feature. For the single video frames the recognition of the objects (metal pole and backpack straps) was identified as discriminating feature. In case of a person just strolling the presence of the hands hanging loose besides the body was identified as key feature. Although the Grad-CAM++ method gave some insight in the pixel regions relevant for the classification it does not give an explanation why certain regions are important for classification.

The data set contains measurements recorded during daytime. Preferably, data should also be recorded at night time. Although a single shot detector is used which might fail to recognize humans in darker lighting conditions training a model jointly for detection and classification might improve the overall detection and classification performance. Which might be a topic for future research.

Furthermore the current implementation uses only single video frames that are associated with 1.28 seconds of radar data. During this time however multiple frames from the video stream are available, ideally all the data collected up to a certain time is used for detection and classification simultaneously, and this will be a topic for future research.

REFERENCES

[1] N. Patel, A. Choromanska, P. Krishnamurthy, and F. Khorrami, "Sensor modality fusion with CNNs for UGV automous drving in indoor environments," in *Proc. IEEE/RSJ IROS*, pp. 1531-1536, 2017.

[2] J. Nglam, A. Khosia, M. Kim et al., "Multimodal deep learning," in *Proc. Int. Conf. Machine Learning ICML*, 2011.

[3] E. Tatulli and T. Hueber, "Feature extraction using multimodal convolutional neural networks for visual speech recognition," in *Proc. ICASSP*, 2017, pp. 2971-2975.

[4] Y. Yasui, N. Inoue, K. Iwano, K. Shinoda, "Multimodal speech recognition using mouth images from depth camera," in *Proc. APSIPA Annual Summit and Conf.*, 2017.

[5] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: a survey and taxonomy," IEEE Trans. Pattern Anslysis Machine Intelligence, vol. 41, no. 2, pp. 423-443, February 2019.

[6] A. Chattopadhay, A. Sarkar, P. Howlader and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[7] M.G. Amin and B. Erol, "Understanding deep neural networks performance for radar-based human motion recognition," in *Proc. IEEE Radar Conf.*, pp. 1461-1465, 2018.

[8] M.J.C. Heiligers, A.G. Huizing and J.J.M. de Wit, "Deep learning for automatic target recognition with radar," in *Proc. EuRAD (special session)*, pp. 38-41, 2018.

[9] W. Liu, D. Anguelov, D. Erhan et al., "SSD: single shot multibox detector," 2016. [Online].

[10] M. Vidanapathirana, "Real-time human detection in computer vision – part 2," 2018. [Online]. Available: https://medium.com/@madhawavidanapathirana/real-time-human-detection-in-computer-vision-part-2-c7eda27115c6 (accessed Jan. 10, 2019).

[11] Y. Kim and T. Moon, "Human detection and activity classification based on micro-Doppler signatures using deep convolutional neural networks," IEEE Geoscience and Remote Sensing Letters, 2016.

[12] R. Trommel, R. Harmanny, L. Cifola, and J. Driessen, "Multi-target human gait classification using deep convolutional neural networks on micro-doppler spectrograms," in *Proc.* EuRAD, 2016, pp. 81–84

[13] F. Chollet and others, "Keras", 2015. [Online]: https://keras.io