CHAPTER 9

Visual Recognition of Abnormal Activities in Video Streams

Konstantinos Gkountakos, Konstantinos Ioannidis, Theodora Tsikrika, Stefanos Vrochidis, and Ioannis Kompatsiaris

9.1 Introduction

The massive streams of visual information captured by CCTV surveillance and body-worn cameras cannot be easily monitored by human operators, particularly in the field of law enforcement. To assist law enforcement officers in their daily tasks and to improve their operational and investigation capabilities, several tools have been developed in order to automatically process and analyse such video streams and subsequently alert the human operators when events of interest, such as any abnormal activities, take place. Abnormalities can be considered as non-normal states, unknown states, everything abnormal, deviant, or outliers. This work

K. Gkountakos (✉) · K. Ioannidis · T. Tsikrika · S. Vrochidis · I. Kompatsiaris
Centre for Research and Technology Hellas, Information Technologies Institute, Thessaloniki, Greece
e-mail: gountakos@iti.gr; kioannid@iti.gr; theodora.tsikrika@iti.gr; stefanos@iti.gr; ikom@iti.gr

B. Akhgar et al. (eds.), *Technology Development for Security Practitioners*, Security Informatics and Law Enforcement,
https://doi.org/10.1007/978-3-030-69460-9_9

focuses on such systems that aim to recognise actions of interest performed by humans or vehicles and categorise each action to one of existing predefined categories. Leveraging the significant advancements in deep learning neural networks, state-of-the-art action recognition methods are based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [10, 12]. Moreover, the architectures of such activity recognition systems typically consist of two parts: the feature extractor and the classifier. To this end, this work proposes an end-to-end activity recognition framework that extracts visual features from video streams and classifies them to predefined activities. The proposed framework is evaluated using the VIRAT [8] dataset and the activities considered in the TRECVID Activities in Extended Video (ActEV) evaluation series [3].

The main contributions of this work are the proposal of a complete end-to-end activity recognition framework based on deep learning neural networks, the investigation of early and late fusion techniques in the context of this framework, and the extensive evaluation experiments using the VIRAT dataset. Moreover, since some of the ActEV activities are fine-grained, we group similar activities together so as to consider coarser-grained activities that are likely to be of more interest to general activity-based recognition systems; we have thus performed evaluation experiments using both the finer- and the coarser-grained activities.

The remainder of the chapter is structured as follows. Section 9.2 discusses related work and relevant datasets, Sect. 9.3 presents the proposed framework, Sect. 9.4 describes the experimental setup and presents the evaluation results, and Sect. 9.5 concludes this work.

9.2 Related Work

State-of-the-art activity recognition methods are based on deep learning techniques. Simonyan et al. [9] proposed a 2D convolution-based architecture that takes into account the visual and stacked optical-flow features and generates a two-stream neural network that can learn simultaneously the motion and the appearance of the input video. Ji et al. [5] proposed a 3D convolution-based approach in order to extract spatio-temporal features, while Tran et al. [12] also trained a 3D convolutional neural network. Hara et al. [4] extended previous works that make use of 3D convolutional kernels with filter size equal to 3×3×3 by using varied kernel sizes and very deep convolutional neural networks. They also concluded that the Kinetics [6] dataset, consisting of more than 300,000 videos that

depict 400 human-related activities, can be widely employed for training and testing activity recognition systems, similarly to the wide use of the ImageNet [2] dataset for training object detection systems.

Apart from Kinetics, several other datasets have been built for the activity recognition problem. HMDB-51 [7] is one of such dataset that consists of more than 6766 videos, with a mean duration of approximately 3 seconds, categorised into 51 human activities extracted from movies. The ActivityNet [1] is another such dataset consisting of around 20, 000 videos categorised into 200 human activities. Finally, both the videos of the VIRAT [8] dataset and their annotations are provided by the National Institute of Standards and Technology (NIST – https://www.nist.gov/) in the context of the TRECVid Activities in Extended Video (ActEV – https://actev.nist.gov/) evaluation series.

9.3 Activity Recognition Framework

This work follows the supervised learning paradigm for human-related activity recognition that employs a deep neural network architecture, namely, the 3D ResNet neural network [4]. This 3D convolutional-based architecture achieves faster processing and can thus perform human activity recognition in (near) real time while using simultaneously (batch) frame processing. In particular, the architectures with 18, 50, and 101 layers as described in [4] have been deployed.

The 3D-ResNet-18 architecture consists of basic blocks, with each block consisting of two 3D convolutional layers followed by batch normalisation and ReLU (rectified linear unit) activation layers, as depicted on the left part of Fig. 9.1. The other two architectures (3D-ResNet-50 and 3D-ResNet-101) follow the bottleneck blocks approach (see right part of Fig. 9.1), where each bottleneck block consists of three 3D convolution layers followed by batch normalisation and ReLU activation layers, with the convolution kernels being 1×1×1 for the first and third convolution layers and 3×3×3 for the middle one.

Finally, it should be noted that the weights of the Kinetics dataset [6] were pre-loaded for all architectures. The Kinetics dataset was selected since it covers a large number of human activity classes (400 classes) and also contains videos that were not collected from sources in specific domains (e.g. movies, soccer games, etc.), but videos from diverse data sources uploaded on YouTube.

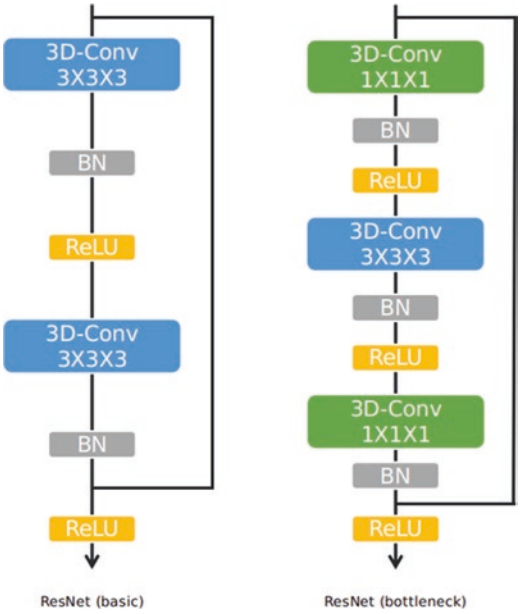

Fig. 9.1 3D-ResNet basic and bottleneck blocks [4]. "to" 3D-ResNet basic and Bottleneck blocks (as illustrated by [4])

9.4 Experiments

This section reports on the experimental evaluation of the proposed activity recognition framework by presenting first the datasets used in our experiments (Sect. 9.4.1), then the overall experimental setup (Sect. 9.4.2), and finally the evaluation results of our experiments (Sect. 9.4.3).

9.4.1 Dataset

In order to evaluate the proposed method, we selected the dataset provided by NIST under the ActEV evaluation series. This dataset was selected since it contains several human activities and vehicle actions that can be considered as abnormal in particular contexts. In particular, ActEV considers activities where one or more people generate movements or interact with objects (or groups of objects), such as other people (P) and vehicles (V). Specifically, ActEV defines and clearly annotates 18 human activities and vehicle actions listed in Table 9.1. The ActEV dataset consists of a total of 2446 annotated activities in its training and validation sets extracted

Table 9.1 ActEV activities official declaration

#	*Activity name*	*Objects acts*	*Description*
1	Closing	(P, V) or (P)	A person closing the door to a vehicle or facility
2	Closing trunk	(P, V)	A person closing a trunk
3	Entering	(P, V) or (P)	A person entering (going into or getting into) a vehicle or facility
4	Exiting	(P, V) or (P)	A person exiting a vehicle or facility
5	Loading	(P, V)	An object moving from person to vehicle
6	Open trunk	(P, V)	A person opening a trunk
7	Opening	(P, V) or (P)	A person opening the door to a vehicle or facility
8	Transport heavy carry	(P, V)	A person or multiple people carrying an oversized or heavy object
9	Unloading	(P, V)	An object moving from vehicle to person
10	Vehicle turning left	(V)	A vehicle turning left or right is determined from the POV of the driver of the vehicle
11	Vehicle turning right	(V)	A vehicle turning left or right is determined from the POV of the driver of the vehicle
12	Vehicle U-turn	(V)	A vehicle making a U-turn is defined as a turn of 180 and should give the appearance of a "U"
13	Pull	(P)	A person exerting a force to cause motion toward
14	Riding	(P)	A person riding a "bike"
15	Talking	(P)	A person talking to another person in a face-to-face arrangement between $n + 1$ people
16	Activity carrying	(P)	A person carrying an object up to half the size of the person
17	Specialised talking phone	(P)	A person talking on a cell phone where the phone is being held on the side of the head
18	Specialised texting phone	(P)	A person texting on a cell phone

from 118 videos of the VIRAT (release 1.0 and 2.0) dataset (http://virat-data.org/). The training set consists of 64 videos that contain 1338 recognised activities, while the validation set consists of 54 videos that contain 1128 recognised activities. The test set will not be considered as its annotations are not publicly available. The distribution of the activities both for the training and validation sets is depicted in Fig. 9.2. As it can be observed, ActEV is a challenging dataset, as it is highly unbalanced.

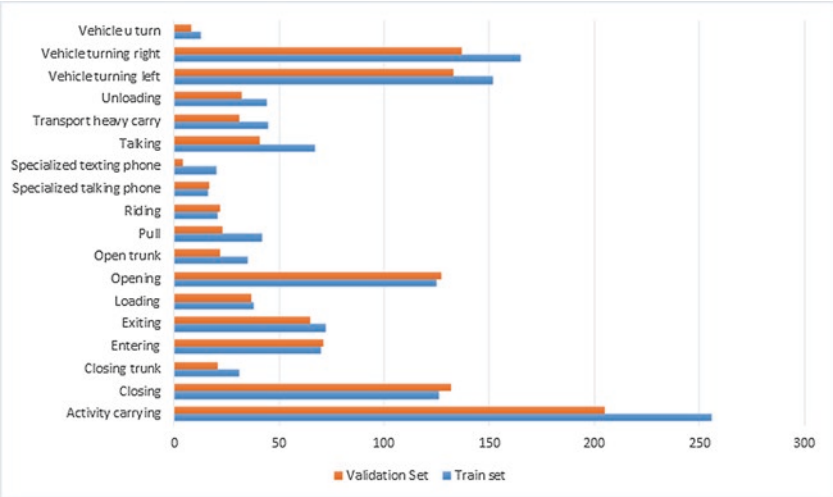

Fig. 9.2 ActEV dataset activities distribution

As some of the ActEV activities are rather fine-grained, we have also grouped similar activities together so as to consider coarser-grained activities that are likely to be of interest to more general activity-based recognition systems (e.g. recognition of vehicle-relevant activities). Table 9.2 lists these so-called super-activities, while Fig. 9.3 depicts the distribution of these super-activities for the training and validation sets, which is also highly unbalanced, similarly to before.

9.4.2 Experimental Setup

The aim of the evaluation experiments was to assess the effectiveness of the activity recognition system, and therefore they focused on processing and analysing only the parts of the video streams where some form of activity had been observed. To this end, first, the frames from all videos were extracted; to be more specific, one every four frames was extracted. Then, only the frames that depict an activity were considered and were stored in a valid format (.png).

The same training strategy was followed for each experiment. Specifically, the batch size was set to 32, the number of total epochs was set to 200, and stochastic gradient descent [11] was used as an optimiser

Table 9.2 ActEV activities grouped to "super-activities"

#	*Activity name*	*ActEV dataset activities*
1	Vehicle	1. Vehicle turning left 2. Vehicle turning right 3. Vehicle U-turn 4. Riding
2	Talking	Talking
3	Person exits	Exiting
4	Person enters	Entering
5	Person carrying activities	1. Loading 2. Transport heavy carry 3. Unloading 4. Activity carrying
6	Person interacts with phone	1. Specialised talking phone 2. Specialised texting phone
The following activities are not taken into account		
1		Closing
2		Closing trunk
3		Open trunk
4		Opening
5		Pull

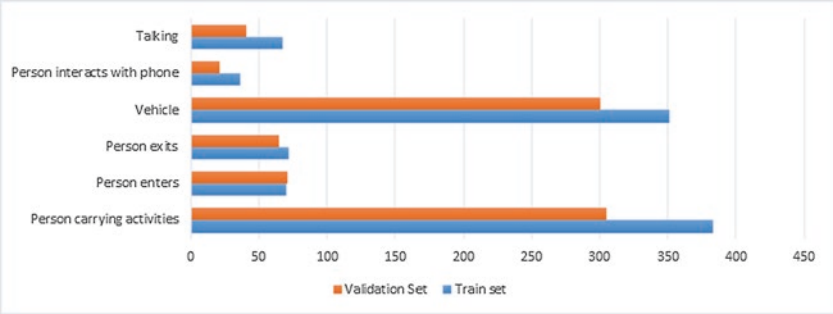

Fig. 9.3 ActEV dataset super-activities distribution

with an initial learning rate equal to 0.1. A "reduce on plateau" strategy was applied in order to create a learning rate schedule with max patience equal to 10 epochs. This strategy allows to reduce the learning rate by a factor once learning stagnates; if no improvement is seen for a "patience" number of epochs, the learning rate is reduced. Furthermore, five

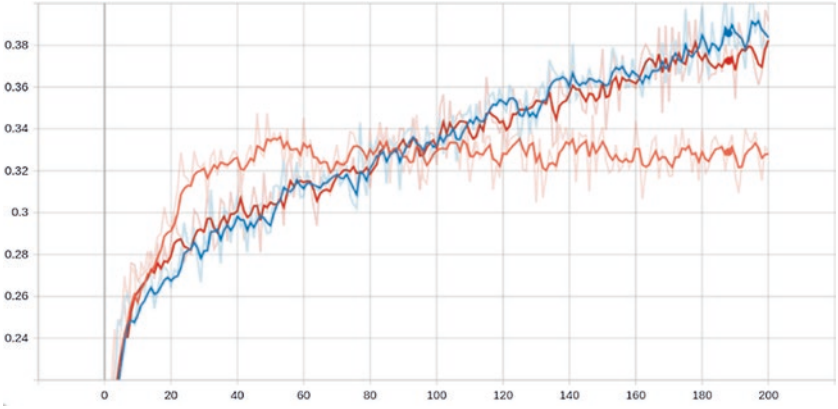

Fig. 9.4 Accuracy during training of ResNet-18(blue), ResNet-50(orange), and ResNet-101(red) with respect to the number of epochs

different scale factors were used for data augmentation [1.0, 0.84, 0.70, 0.59, 0.49], while a corner cropping strategy was also applied; this refers to the random selection of a cropped box from the four corners and the centre.

The training process was monitored for a complete evaluation by utilising the TensorBoard application downloaded from the TensorFlow[1] repository. Figure 9.4 presents the accuracy per epoch during training and denotes the 3D-ResNet architecture consisting of 18, 50, and 101 layers with blue, orange, and red, respectively. The correspondingly losses during training are depicted in Fig. 9.5.

The validation set of the ActEV dataset was used for evaluating the proposed activity recognition framework in order to investigate how the depth of a 3D-ResNet network architecture affects its effectiveness. To this end, we applied two different experimental settings, one that considers the 18 activities of the ActEV dataset and one that considers the 6 super-activities. Regarding the super-activities, we apply both late and early fusion. For the late fusion, the accuracy of each super-class comprises the summation of the subclasses' predictions during testing, whereas for early fusion, the super-activities are merged during training (i.e. a single training set is created for each super-activity by merging the training sets of its sub-activities).

[1] https://github.com/tensorflow/tensorboard

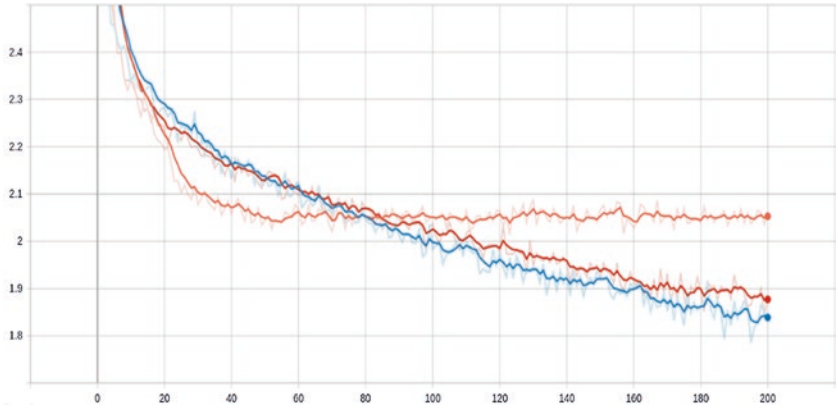

Fig. 9.5 Cross-entropy loss during training of ResNet-18(blue), ResNet-50(orange), and ResNet-101(red) with respect to the number of epochs

Precision@N is used as the basic evaluation criterion which allows us to show the accuracy of the framework for different numbers of retrieved activities where $N \in \{1, \ldots, 18\}$ in the case of ActEV activities and $N \in \{1, \ldots, 6\}$ in the case of super-activities. Precision@1 indicates the percentage of videos where the top prediction by our framework corresponds to the correct activity shown in the video. Hence, Precision@18 for the ActEV activities and Precision@6 for the super-activities should always be equal to 1, as the framework is bound to predict correctly if it simply provides all available activities. In addition, confusion matrices are also presented.

9.4.3 Results

This section presents the results for the different ResNet architectures both for the 18 activities and also for the 6 super-activities; in the latter case, the results listed below correspond to the late fusion, whereas the results for the early fusion are presented at the end of this section.

ResNet-50 results. Figure 9.6 presents the Precision@N using the ResNet-50 architecture. Precision@1 equals to 28% when all 18 activities are considered and 51% in the case of super-activities. As expected, coarser-grained activities can be more easily identified. Figures 9.7 and 9.8 present the confusion matrices of the prediction activities both for the 18 activities

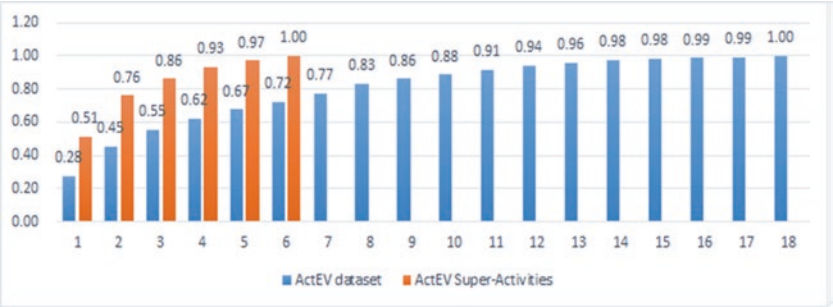

Fig. 9.6 Precision@N, ActEV, and super-activities trained using ResNet-503

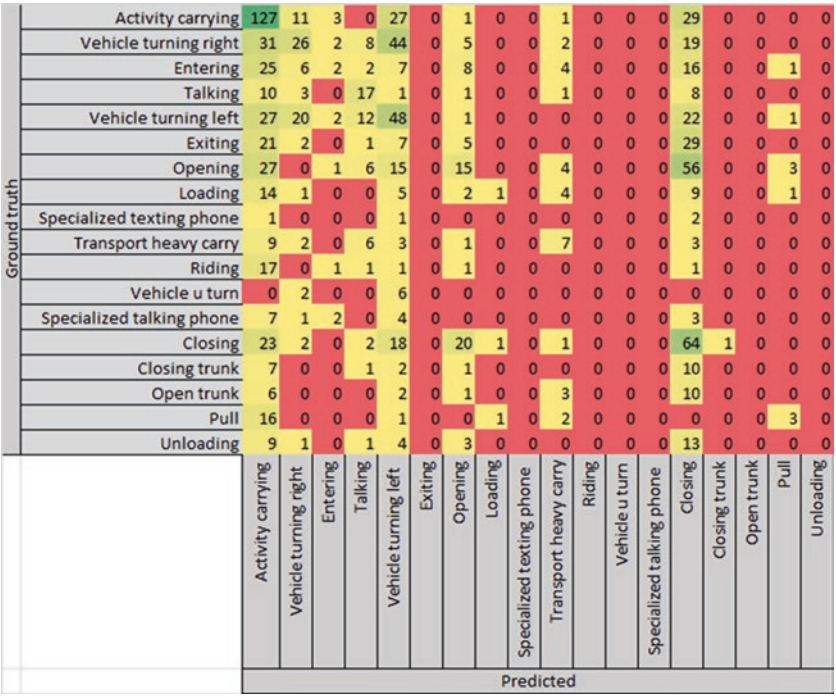

Fig. 9.7 Confusion matrix using ActEV dataset trained on ResNet-50

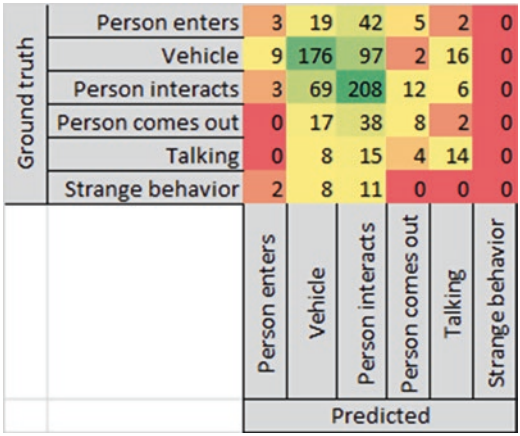

Fig. 9.8 Confusion matrix using super-activities dataset trained on ResNet-50

Fig. 9.9 Precision@N, ActEV, and super-activities trained using ResNet-18

and the 6 super-activities. A detailed examination indicates that the unbalanced characteristics of the ActEV dataset lead the model to a dominated learning state adapted to the activity with the highest occurrence ("activity carrying"). On the other hand, in the super-activities dataset, the number of false negatives and false positives has been reduced and disengaged from a dominating activity.

ResNet-18 Results. Figure 9.9 presents the Precision@N using ResNet-18 architecture. Precision@1 has decreased to 25%, compared to the 28% achieved by the ResNet-50 architecture for the 18 activities. Regarding the super-activities, Precision@1 has also decreased from 51% to 47%.

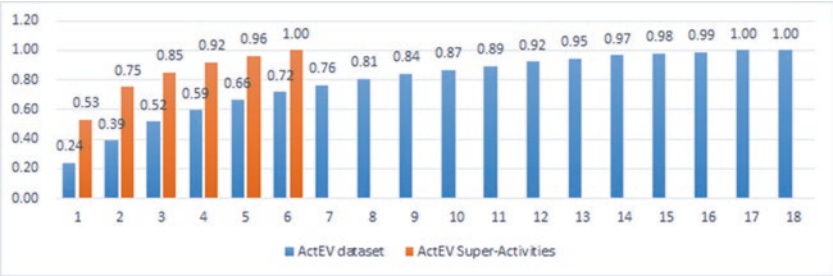

Fig. 9.10 Precision@N, ActEV, and super-activities trained using ResNet-101

ResNet-101 Results. Finally, the results of the experiments for the ResNet-101 neural network architecture are depicted in Fig. 9.10. As the results indicate, a higher capacity neural network can learn more accurately the classification problem. Specifically, the ResNet-101 architecture outperforms the previous ones when considering the super-activities, but the results for the 18 activities dataset are even lower than the ResNet-50 architecture. A detailed examination indicates that many of these 18 activities are closer (in terms of visual content) to each other, and thus, a higher capacity neural network which tries to differentiate between them aggressively results in lower Precision@1, even though the Precision@5 remains similar to the ResNet-50 results.

Early Versus. Late Fusion. In addition to the late fusion experiments presented above, we also carried out early fusion experiments for the case of super-activities.

To compare the effectiveness of the two approaches, we select the ResNet-101 architecture as it achieves the best performance in the case of super-activities. Figure 9.11 depicts the Precision@N both for early and late fusion. Specifically, early fusion increases the system performance for all N except for Precision@1. Furthermore, Fig. 9.12 compares the confusion matrices for early and late fusion and indicates that although the Precision@1 is lower when applying early fusion, the value of the error of misclassified activities is smaller and Precision@N for $N > 1$ is higher.

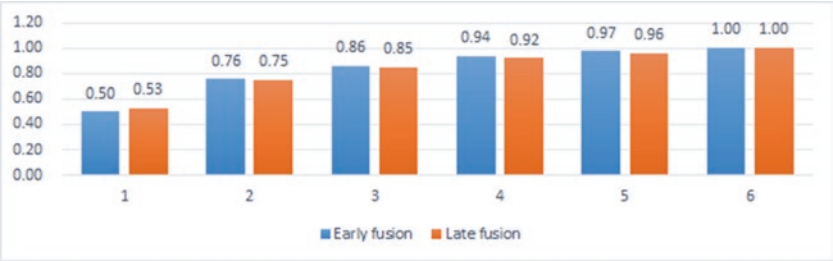

Fig. 9.11 Precision@N both for early and late fusion using ResNet-101

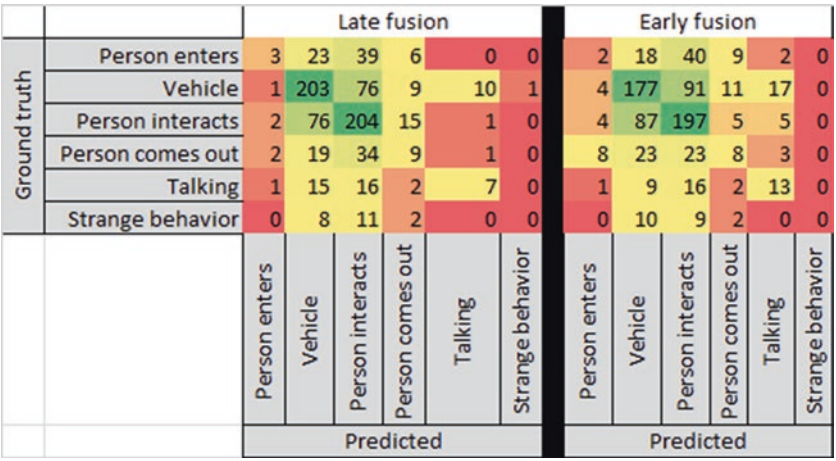

Fig. 9.12 Confusion matrices both for early and late fusion using ResNet-101

9.5 Conclusions

This work presented a framework for recognising activities in video streams. Specifically, the framework makes use of 3D convolutional filters in order to learn the spatio-temporal representation of activities. The framework was evaluated using the challenging ActEV dataset and also a second dataset that was created using the same data and which merges the ActEV activities into super-activities in order to evaluate the proposed framework in a more general activity-based recognition domain. The

experimental results indicate that our framework can capture coarse level representations as it performs satisfactorily in the super-activities dataset. Finally, the early fusion approach proved to be advantageous in contrast to the late fusion when more than one activity were retrieved.

Acknowledgements

This research has received funding from the European Union's H2020 research and innovation programme as part of the CONNEXIONs (H2020-786731) project.

Bibliography

1. Caba Heilbron, F., Escorcia, V., Ghanem, B., & Carlos Niebles, J. (2015). Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 961–970). IEEE.
2. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 248–255). IEEE.
3. Awad, G., Butt, A. A., Curtis, K., Lee, Y., Fiscus, J., Godil, A., ... & Quenot, G. (2020). Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval. arXiv preprint arXiv:2009.09984.
4. Hara, K., Kataoka, H., & Satoh, Y. (2018). Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6546–6555). IEEE.
5. Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 35*(1), 221–231.
6. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950.*
7. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: a large video database for human motion recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 2556–2563). IEEE.

8. Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C. C., Lee, J. T., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., et al. (2011). A large-scale benchmark dataset for event recognition in surveillance video. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3153–3160). IEEE.
9. Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Proceedings of the international conference of advances in Neural Information Processing Systems (NIPS),* (pp. 568–576), NeurIPS foundation.
10. Singh, D., Merdivan, E., Psychoula, I., Kropf, J., Hanke, S., Geist, M., & Holzinger, A. (2017). Human activity recognition using recurrent neural networks. In *Proceedings of the international Cross-Domain conference for Machine Learning and Knowledge Extraction (CD-MAKE)* (pp. 267–274). Springer.
11. Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 1139–1147).
12. Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 4489–4497). IEEE.