# Early vs Late Fusion in Multimodal Convolutional Neural Networks

1st Konrad Gadzicki
*Cognitive Neuroinformatics*
*University of Bremen*
Bremen, Germany
gadzicki@uni-bremen.de

2nd Razieh Khamsehashari
*Cognitive Neuroinformatics*
*University of Bremen*
Bremen, Germany
rkhamseh@uni-bremen.de

3rd Christoph Zetzsche
*Cognitive Neuroinformatics*
*University of Bremen*
Bremen, Germany
zetzsche@informatik.uni-bremen.de

***Abstract*—Combining machine learning in neural networks with multimodal fusion strategies offers an interesting potential for classification tasks but the optimum fusion strategies for many applications have yet to be determined. Here we address this issue in the context of human activity recognition, making use of a state-of-the-art convolutional network architecture (Inception I3D) and a huge dataset (NTU RGB+D). As modalities we consider RGB video, optical flow, and skeleton data. We determine whether the fusion of different modalities can provide an advantage as compared to uni-modal approaches, and whether a more complex early fusion strategy can outperform the simpler late-fusion strategy by making use of statistical correlations between the different modalities. Our results show a clear performance improvement by multi-modal fusion and a substantial advantage of an early fusion strategy.**

***Index Terms*—Multi-layer neural network, Activity recognition, Sensor fusion**

## I. INTRODUCTION

The research of human activity recognition has gained attention over the years due to its utilization in various fields. The collaborative research center "EASE" (http://www.ease-crc.org) has the goal to develop robots capable of performing everyday activities. Examples from humans executing household activities like table setting, cooking etc. serve as an important information source for determining appropriate actions of the robot. The automated recognition of those activities is key for the access and analysis of data in a huge database of recorded human activities.

Deep learning had a tremendous impact on machine learning and pattern recognition, achieving results beyond the performance levels of classical approaches. Tasks like activity recognition profit substantially from deep learning, e.g. by utilizing the expressive power of Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). These approaches perform well on a large number of different sensory data relevant for activity recognition, e.g. image data, video, audio or skeleton data. Since activities are mostly recorded over time, the convolutional units of appropriate CNNs have to comprise the temporal dimension as well by applying some sort of spatio-temporal convolutions.

This work was supported by DFG (German Research Foundation) as part of Collaborative Research Center "EASE - Everyday activities Science and Engineering" (http://www.ease-crc.org)

Although CNNs featuring a single data source show already impressive performance for activity recognition, the fusion of several modalities could be a promising research direction for a further improvements of performance. Ambiguities of single data sources might be resolved and correlations between date sources could be exploited by integrating different modalities, thus improving the overall system performance.

Here we investigate the potential of multimodal fusion strategies in the context of spatio-temporal convolutional neural networks. We compare different multimodal architectures in relation to the unimodal variants without aiming for state-of-art performance on the dataset.

## II. RELATED WORK

### A. Data for Activity Recognition

Currently there is a large number of datasets for activity recognition available (for a review see [1]). They cover a wide range of modalities which are suitable for this task. RGB video data is often part of a dataset and recent approaches have focused on this modality, often together with additional modalities like depth maps. This combination of RGB+D is popular due to the frequent use of Microsoft Kinetic for recording of human activity. Since this device is also able to extract skeleton information from the RGB+D signal, skeleton data is often available as a third useful modality.

Every particular modality can provide different information, with its usefulness depending on the task at hand. Video data provide information about color, texture or shape of people and objects as well as about the whole scene in which an activity takes place. While this modality is very rich in information, the variation of illumination, color etc. in a real-world RGB channel makes it tricky to process.

Channels which are largely invariant to these variations might provide useful data which are easier to process. Depth maps, for instance, are rather invariant to illumination changes which might be very helpful for segmentation or for shape extraction. And there are further channels that can be derived from RGB or depth data. Optical flow is usually extracted from RGB and provides information about spatio-temporal changes in the scene. Skeleton data can be extracted from RGB or from depth, but can also be directly recorded with motion capturing. The information about skeleton points is especially

valuable for analyzing the human part of the scene in activity recognition.

The "NTU RGB+D" [2] which will be used in this paper, is a large-scale dataset providing RGB, depth and skeleton data. It offers large number of subjects and classes from different viewpoints.

### B. Activity Recognition Models

The analysis of human activities has drawn significant attention, with special interest in action recognition from RGB video. In the last decade the success of CNN-based approaches in image-based classification has led to the application of these methods to video data. Video data can be treated as a series of 2D image, each processed with a 2D-CNN. While this is sufficient to extract the spatial features, the temporal dynamics need to be captured as well. [3] introduced multi-frame optical flow as an input to a 2D-CNN together with RGB frames. This sort of network is basically a 2D image classification CNN which has the advantage of being pre-trainable on Imagenet [4].

In recent years spatio-temporal 3D-CNN were introduced for processing multiple frames directly [5]–[8]. Two stream CNN [9] add optical flow, derived from RGB video, as a second modality to the network. Multistage CNN [10] and structured segment network [11] add the ability to detect actions in untrimmed videos by generating proposals for time slices with actions. The fusion of multiple modalities in CNNs has been investigated by [25] and [26].

Apart from convolutional neural networks, recurrent neural networks offer another way to process video data. Here the temporal dynamics between individual frames are captured by the recurrent structure of the network [12]–[14].

Skeleton data, representing the positions of joints of a human body over time, offer rich information with regard to human activity recognition. CNN-based approaches can treat the $x, y, z$-position of joints as separate time series and process them with a time convolutional network [15]. Another way is to transform the joint information into a 2D structure and use 2D-CNN for processing. [16] interpret the joint positions as 2D information and color code the temporal dynamics, [17] use one image dimension for coding the spatial structure of joints and the other for the temporal dynamics and [18] project the 3D positions onto four different 2D planes and encode the joint distances in those planes in images.

As for the combination of modalities, the approaches which work well for a certain modality, are not necessarily suitable or working equally well in a multimodal system.

### C. Multi-Sensor Fusion

The fusion of multiple data sources is well established in literature. Bellot [19] identifies four gains which the fusion process might achieve:

- "gain in representation": the fused representation reaches a higher level of granularity or abstract level than the initial data sources.
- "gain in certainty": increase in the belief in the fused data.
- "gain in accuracy": the standard deviation on the data improves. Noise and errors are decreased.
- "gain in completeness": addition of new information make the view on the environment more complete.

With regard to the field of activity recognition, an overview of multi-sensor fusion can be found in [20] and [21]. Multi-sensor fusion is used when several sensors are placed in the environment [22], [23] or on the human body (wearable sensors) [24].

## III. MULTIMODAL FUSION

### A. Multimodal Fusion Strategies

Using multimodal approaches in a machine learning context is typically aimed at an improvement of the overall system performance with regard to recognition power or robustness. The idea is that individual data sources can provide different kinds of information which might resolve ambiguity, improve the overall quality of noisy data, or enable the exploitation of correlations.

One of the challenges of multimodal machine learning [27] lies in the methods for the fusion of the different modalities. The respective approaches can be broadly categorized as early and late fusion [28], depending on the position of the fusion within the processing chain. Hybrid fusion approaches try to combine the properties of the two basic methods [28].

Late fusion [29] is the simplest and most commonly used fusion method. It merges data after a separate full processing in different unimodal streams. The individual modalities can be processed by powerful targeted approaches, tailored to the specific properties of the particular modality. After a full chain of unimodal processing, typically after predicting labels in a recognition task, the results are merged, in the most simple case by summation or averaging. Late fusion has a major drawback which is the very limited potential for the exploitation of cross correlations between the different unimodal data.

Early fusion [29] is more powerful since it merges data sources in the beginning of the processing. Raw data can be fused directly without any pre-processing, but usually certain features are initially extracted. These basically unimodal features are then fused by concatenating the individual data into a joint representation. The unified representation has to make sure that the data is properly aligned, thus being suitable for further joint processing. If the data is properly aligned, cross-correlations between data items may be exploited, thereby providing an opportunity to increase the performance of the system. [25] argue that those fused low-level features might be irrelevant for the task, thus decreasing the fusion power.

Between late and early fusion as the extremes, it is also possible to use a halfway fusion [25] or middle fusion [26]. Here the fusion point somewhere in the middle of the network.

In this paper, we want to investigate whether the fusion of different modalities can provide an advantage as compared

to uni-modal approaches, and whether a more complex early fusion strategy can outperform the standard late-fusion strategy by making use of statistical correlations between the different modalities. We address this issue in the context of human activity recognition. To ensure a meaningful comparison we avoid special solutions but use one state-of-the-art convolutional network architecture (Inception I3D) for all different settings. Furthermore, we perform the tests on a sufficiently large and general dataset (NTU RGB+D). As modalities we consider RGB video, optical flow, and skeleton data.

### B. Multimodal Fusion for Convolutional Neural Networks

Convolutional neural networks have reached remarkable success in a variety of applications. Combining multimodal fusion and CNNs thus appears to be a promising direction for future research. In particular, the possible fusion methods described above can be applied to CNNs.

In the case of early fusion, one can combine raw data or early features. Since raw data from different data sources are rarely spatio-temporally aligned due to different resolutions or sampling frequencies, they require a certain amount of pre-processing before being concatenated for processing by a CNN. If one moves one step further in the network and starts fusion at an early features level, the requirement for spatio-temporal alignment remains, but there are several ways how to extract features. The most simple case is to use convolutional units for feature extraction and train them from scratch, or pre-train on a different dataset which offers the same modality. One could also bootstrap those units with weights learned on a similar data source. For instance one can train on Imagenet [4] and initialize a CNN with these pretrained weights. If the dimensionality changes to 3D as with video data, 2D weights can still be used for bootstrapping [9]. A last possibility is to use classical approaches for features extraction, e.g. a filter bank of parameterized filters.

For late fusion several unimodal networks are used as the basis for the fused architecture. The individual networks can be heterogeneous, fitting only the modality they are designed for. The actual fusion is then trivial requiring only the merging of the individual results of each network, i.e. the predicted labels in a recognition task. In order to achieve the fusion, the dense layers which usually form the output layer of a network need to be merged by summing or averaging.

The fusion of raw data resp. first features in the early fusion case and the fusion of final predictions in the late fusion case are the extreme variants. Apart from these two there are many more potential fusion points within a deep CNN. With increasing number of layers, the complexity of individual features typically rises. Fusing such correlated complex features for multiple modalities might result in increased performance.

## IV. METHODS

### A. Dataset

The reasons for using the "NTU RGB+D" dataset [2] are the size of the dataset (over 56k samples, ca. 40k for training and 16k for validation) and the different modalities (RGB video, depth video, IR video and skeleton data) it offers. There are 60 classes, 40 subjects performing the activities and three different view points. The data have been recorded with a Microsoft Kinect v2. Figure 1 shows sample frames from the dataset.

Fig. 1: Sample frames of the NTU RGB+D dataset [2].

### B. Modeling

For our investigation we have used early and late multimodal convolutional neural networks as well as the respective unimodal CNNs. Our basic architecture uses an established architecture from literature, "Inception-v1 I3D" [9]. This architecture uses "Inception" modules (see Figure 2) which introduce parallel pathways for processing of a given input, concatenating the results of each pathway as an output of the module. These "Inception" blocks are repeated several times with `MaxPooling` operations and down-sampling in between at certain points as shown in Figure 3.

We use RGB video, optical flow based on the RGB and skeleton data in our work. The CNNs have been implemented in TensorFlow [30].

The structure of the early and late fusion systems are shown in Fig. 4. The early fusion variant does not fuse the raw data, but the one of the first convolutional features. Architectures like "Inception I3D" are designed to have a set of convolutional layers before processing the data with a series of "inception modules" (see Figure 2) respectively. The first layers, leading to the first "inception" block, are called the stem of the network. We apply the term early fusion, if the fusion takes place within the stem.

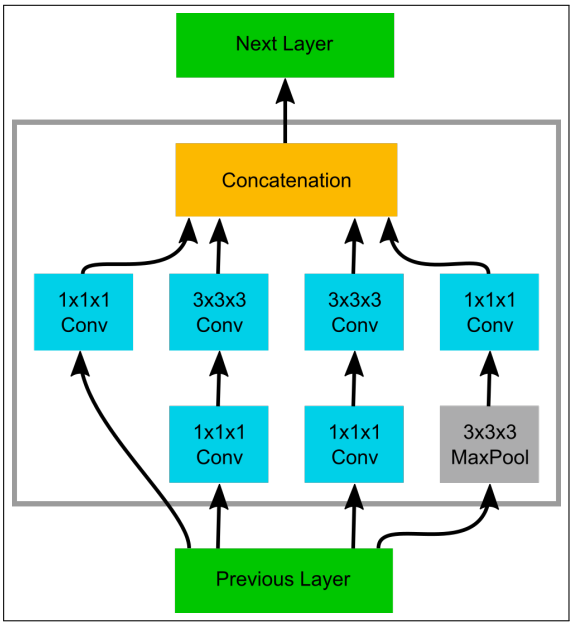


Fig. 2: Structure of an Inception-v1 block (adapted from [9])

The early and mid fusion are performed by concatenating the outputs of the partial networks for the given modalities. For instance, the early fusion of the first convolutional layers (stem layers 1[1]) concatenates the outputs for these layers which have been calculated for each modality, e.g. RGB and Optical Flow, separately. Afterwards the resulting tensor is fed into the remaining network without changing the architecture any further.

In the late fusion variant, the individual modalities are process on their own up to the dense layer at the top which are summed in the end (see Figure 4b).

The fusion of RGB video with Optical Flow is straight forward due to the same dimensionality of these modalities. The processing of skeleton data within the early fusion architecture would require a transformation of the input data, i.e. we transform the 1D skeleton data for a particular time step to 2D. Therefore, we tested the late fusion variant only with skeleton data which does not required any further transformation of the data. Here we can use an existing an existing CNN, build for 1D data, and sum its results with the outputs of the other branch. We use "Res-TCN" [15] for the late fusion variants. It is a residual network with temporal convolutions, thus the convolutions are 1D in nature.

The loss function is sparse softmax cross entropy. We use a Momentum optimizer with 0.9 momentum and a learning rate of 0.001

Based on the RGB videos, optical flow has been computed with "FlowNet2" [31]. The original RGB videos were downscaled to 256x256 and randomly cropped to 224x224 pixels around the center. The optical flow data has been processed in the same manner with the cropping positions being aligned

[1]For "Inception I3D": after 'conv3d_1a_7x7' in the implementation from `https://github.com/deepmind/kinetics-i3d`.

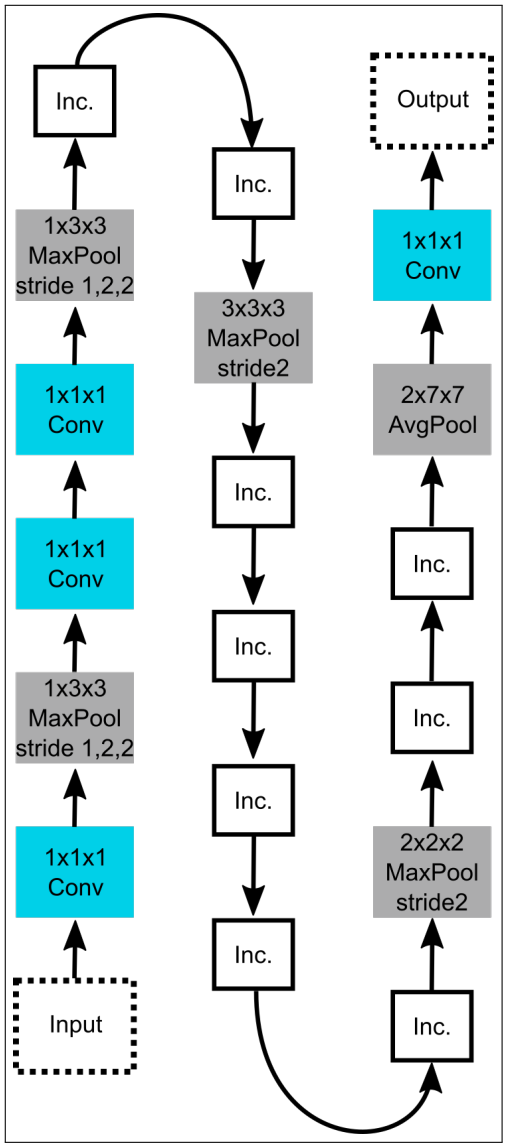


Fig. 3: Layout of the Inception-v1 I3D CNN (adapted from [9])

to the RGB video. The time slices were set to 10 seconds and looped if the data was shorter. The FPS was 25. The skeleton data consists of the $x$,$y$,$z$ coordinates for 25 joints per person and was computed with the Kinect v2. There were maximally two people tracked during the recording.

## V. RESULTS

We have used the "NTU RGB+D" dataset with the cross-subject split provided by [15] which provides 40320 samples for training. The unimodal and multimodal variants of the network are based on the "Inception I3D" architecture. The results involving skeleton data (unimodal and multimodal late fusion) were obtained with "Res-TCN" [15] for the skeleton part. Figure 5 shows an exemplary plot of the training of a multimodal network. There are no signs indicative of over-training.

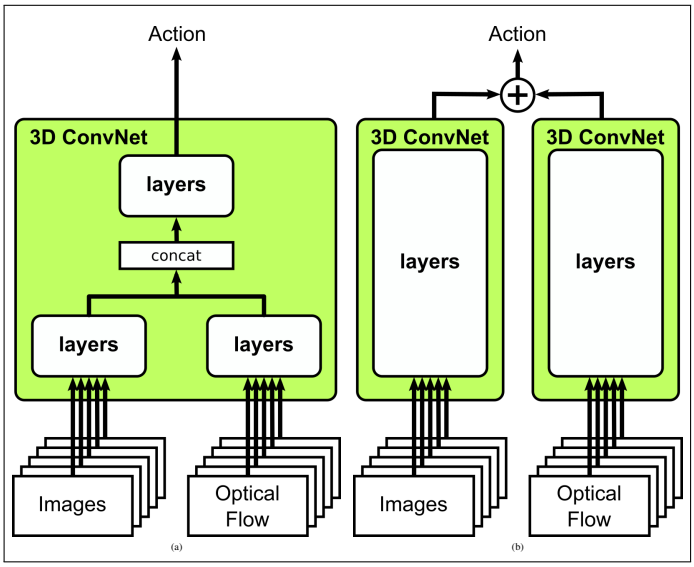


Fig. 4: CNN with (a) early fusion and (b) late fusion. The modalities shown here are RGB images and Optical Flow. For the early fusion the action label is directly output by the logits layer of the fused network. For late fusion the outputs of the logits layers of the individual CNNs for each modality are summed.

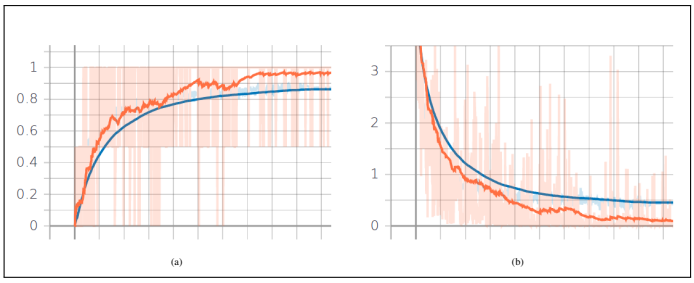


Fig. 5: Accuracy (a) and loss (b) for early fusion with RGB and optical flow with training set (orange) and validation set (blue).

Table I shows a summary of the results in terms of recognition performance. We measure the recognition accuracy by number of correct predictions divided by number of samples. For the unimodal versions of the CNNs classification performance ranges from 66.4% to 78.6%. Optical flow as a single modality provides a relatively poor performance of 66.4%. The other two modalities enable a significantly better recognition, with the best result of 78.6% being obtained on basis of the skeleton data.

The multimodal versions all show an improved performance in comparison to their unimodal counterparts. Somewhat surprising, the smallest improvement to a level of 82.3% is obtained by a late fusion of the RGB channel with the unimodally best performing skeleton channel. Nevertheless, the multimodal performance is superior to that of the individual channels alone (76.8% and 78.6%). The best multimodal performance of 86.7% is obtained with a network which makes use of an early fusion architecture which integrates RGB and optical flow.

TABLE I: Multimodal fusion results of "Kinetics I3D" [9] on NTU Dataset [2] for all results but skeleton data (marked with *) which used "Res-TCN" [15].

| | Fusion | Trained Layers | Modalities | Accuracy |
|---|---|---|---|---|
| Unimodal | RGB video | all layers | RGB | 76.8% |
| | Optical Flow | all layers | Optical Flow | 66.4% |
| | *Skeleton | all layers | Skeleton | 78.6% |
| Multimodal | Early Fusion | all layers | RGB, Op. Flow | 86.7% |
| | Late Fusion | all layers | RGB, Op. Flow | 82.9% |
| | *Late Fusion | last layer | RGB, Skeleton | 82.3% |

## VI. DISCUSSION

In this paper our investigation was focused on the question whether the fusion of information from several data sources is helpful for the task of human activity recognition by convolutional neural networks. Our results show that any sort of fusion will improve the performance. This is valid irrespective of whether the fusion is performed early or late, and irrespective of which modalities are combined.

On a detailed level, our investigations show a clear superiority of an early fusion strategy over a late combination (86.7% for early as opposed to 82.3% and 82.9% for late). This lends support to the hypothesis that a multimodal convolutional network architecture in which the information from different modalities can be combined and recombined across processing stages is able to exploit the multivariate correlational structure of the data sources.

It is interesting to note that in our setting the specific modalities used for the combination seem to have less relevance than the fact that a combination is used at all. Although the unimodal skeleton channel as such yields a much higher performance than unimodal optic flow (78.6% vs. 66.4%), a late fusion of this skeleton channel with the RGB channel cannot provide a better performance than the fusion of the seemingly inferior optic flow channel with the RGB channel (82.3% vs. 82.9%).

Future work can explore several directions for multimodal information fusion with convolutional networks. One direction is the full integration of skeleton data into an early fusion architecture. For this we have to bring the image raster data and the skeleton data into a format suitable for combination. Another direction is to investigate halfway fusion. [25] achieved best results by fusing in the middle of the network. On the other hand [26] reported worse performance for middle fusion under certain conditions.

A further direction is to consider hybrid approaches which try to combine the advantages of both early and late fusion methods. For example one could combine highly specialized processing architectures for unimodal data streams with more

general architectures for early fusion, e.g. for pairwise combination of modalities [32]. These pathways can finally be merged by late fusion, allowing to exploit potential cross-correlations residing in the different data streams, and at the same time use sophisticated models for each data stream.

In conclusion, our results yield further support for the general idea that fusion of and within convolutional network architectures could be a promising research direction for human activity recognition. The greatest potential, in our view, should be sought in an early integration of a variety of information sources.

## REFERENCES


[1] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "Rgb-d-based human motion recognition with deep learning: A survey," *CoRR*, vol. abs/1711.08362, 2017. [Online]. Available: http://arxiv.org/abs/1711.08362

[2] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3d human activity analysis," *CoRR*, vol. abs/1604.02808, 2016. [Online]. Available: http://arxiv.org/abs/1604.02808

[3] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 568–576. [Online]. Available: http://dl.acm.org/citation.cfm?id=2968826.2968890

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 4489–4497. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.510

[6] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 140–153.

[7] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, Jan 2013.

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 1725–1732. [Online]. Available: https://doi.org/10.1109/CVPR.2014.223

[9] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 4724–4733.

[10] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[11] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[12] Y. Shi, Y. Tian, Y. Wang, W. Zeng, and T. Huang, "Learning long-term dependencies for action recognition with a biologically-inspired deep network," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[13] D.-A. Huang, L. Fei-Fei, and J. C. Niebles, "Connectionist temporal modeling for weakly supervised action labeling," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 137–153.

[14] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. Carlos Niebles, "Sst: Single-stream temporal action proposals," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[15] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 1623–1631.

[16] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 3, pp. 807–811, March 2018.

[17] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, Nov 2015, pp. 579–583.

[18] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 624–628, May 2017.

[19] D. Bellot, A. Boyer, and F. Charpillet, "A new definition of qualified gain in a data fusion process: application to telemedicine," in *Proceedings of the Fifth International Conference on Information Fusion. FUSION 2002. (IEEE Cat.No.02EX5997)*, vol. 2, 2002, pp. 865–872 vol.2.

[20] A. A. Aguileta, R. F. Brena, O. Mayora, E. Molino-Minero-Re, and L. A. Trejo, "Multi-sensor fusion for activity recognition—a survey," *Sensors*, vol. 19, no. 17, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/17/3808

[21] B. Fu, N. Damer, F. Kirchbuchner, and A. Kuijper, "Sensing technology for human activity recognition: A comprehensive survey," *IEEE Access*, vol. 8, pp. 83 791–83 820, 2020.

[22] M. Tenorth, J. Bandouch, and M. Beetz, "The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition," in *IEEE International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS), in conjunction with ICCV2009*, 2009.

[23] R. S. Ransing and M. Rajput, "Smart home for elderly care, based on wireless sensor network," in *2015 International Conference on Nascent Technologies in the Engineering Field (ICNTE)*, 2015, pp. 1–5.

[24] C. Zhu and W. Sheng, "Human daily activity recognition in robot-assisted living using multi-sensor fusion," in *2009 IEEE International Conference on Robotics and Automation*, 2009, pp. 2154–2159.

[25] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," *CoRR*, vol. abs/1611.02644, 2016. [Online]. Available: http://arxiv.org/abs/1611.02644

[26] N. Damer, K. Dimitrov, A. Braun, and A. Kuijper, "On learning joint multi-biometric representations by deep fusion," in *Proceedings of the IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS 2019)*, Tampa, FL, USA, 10 2019.

[27] T. Baltrušaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, Feb 2019.

[28] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, Feb. 2015. [Online]. Available: https://doi.org/10.1145/2682899

[29] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, ser. MULTIMEDIA '05. New York, NY, USA: Association for Computing Machinery, 2005, p. 399–402. [Online]. Available: https://doi.org/10.1145/1101149.1101236

[30] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and et al., "Tensorflow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, ser. OSDI'16. USA: USENIX Association, 2016, p. 265–283.

[31] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2017/IMSKDB17

[32] Z.-Z. Lan, L. Bao, S.-I. Yu, W. Liu, and A. G. Hauptmann, "Multimedia classification and event detection using double fusion," *Multimedia Tools Appl.*, vol. 71, no. 1, p. 333–347, Jul. 2014. [Online]. Available: https://doi.org/10.1007/s11042-013-1391-2