Human Action Recognition based on 3D Convolution Neural Networks from RGBD Videos

Rawya Al-Akam
Active Vision Group, AGAS
Institute for Computational Visualistics
University of Koblenz-Landau
Universitatsstr. 1
56070 Koblenz, Germany
rawya@uni-koblenz.de

Dietrich Paulus
Active Vision Group, AGAS
Institute for Computational Visualistics,
University of Koblenz-Landau
Universitatsstr. 1
56070 Koblenz, Germany
paulus@uni-koblenz.de

Darius Gharabaghi
Institute for Computational Visualistics,
University of Koblenz-Landau
Universitatsstr. 1
56070 Koblenz, Germany
darius.gh@gmx.de

ABSTRACT

Human action recognition with color and depth sensors has received increasing attention in image processing and computer vision. This paper target is to develop a novel deep model for recognizing human action from the fusion of RGB-D videos based on a Convolutional Neural Network. This work is proposed a novel 3D Convolutional Neural Network architecture that implicitly captures motion information between adjacent frames, which are represented in two main steps: As a **First**, the optical flow is used to extract motion information from spatio-temporal domains of the different RGB-D video actions. This information is used to compute the features vector values from deep 3D CNN model. **Secondly**, train and evaluate a 3D CNN from three channels of the input video sequences (i.e. RGB, depth and combining information from both channels (RGB-D)) to obtain a feature representation for a 3D CNN model. For evaluating the accuracy results, a Convolutional Neural Network based on different data channels are trained and additionally the possibilities of feature extraction from 3D Convolutional Neural Network and the features are examined by support vector machine to improve and recognize human actions. From this methods, we demonstrate that the test results from RGB-D channels better than the results from each channel trained separately by baseline Convolutional Neural Network and outperform the state of the art on the same public datasets.

Keywords

Action Recognition, RGBD videos, Optical Flow, 3D Convolutional Neural Network, Support Vector Machine.

1 INTRODUCTION

The human action recognition from videos is challenging field in real-world actions and has advanced rapidly over the last few years. Due to the large intra-class variations, high dimension of video data, varying motion speed, partial occlusion and clutter background, precise action recognition is still a big challenging task. And the efficient solutions to this challenging and difficult problem can facilitate several useful applications such as visual surveillance, human-robot cooperation, and medical monitoring systems [JBCS13]. A recent development of range sensors had an incontrovertible influence on research and applications of machine and computer vision field. Sensor devices provide depth information of the scene view and objects, that helps in solving problems which are looked hard for RGB images or videos [HSXS13].

Classical action recognition tasks mainly depend on hand-crafted features which can be divided into local and global approaches. Local feature extraction methods which consist of two steps: detection and description, such as the spatio-temporal interest point detection (STIP) [Lap05], improved dense trajectories (IDT) [WS13] and histogram of optical flow (HOF) [LSR08] are widely used as a local feature for human action recognition task. Local feature extraction approaches are much more efficient and robust in real scenes applications. While the global feature extraction approaches represent the video sequence as a whole which is capturing the general appearance and mo-

tion information from each frame in video sequences. In spite of the global approaches are very sensitive to occlusion, cluttering and shift but it is still using and commonly existing for human action recognition tasks [SLY16]. Regrettably, the hand-crafted features-based encoding methods such as fisher vector [PD07] and bag-of-words (BoW) [VVV16, VBM11], which are represented as universal visual that does not consider much about temporal information for video-based action recognition [YCXL17].

In recent years the use of neural networks and deep learning algorithms has shown significant progress in several fundamental problems in computer vision, including action and activity recognition and also the Convolutional Neural Network (CNN) is utilized to solve different image processing tasks such as object and action recognition, and also the classification done from not only static images but also from dynamic sequence of images. And besides that, the recent development of depth camera sensors that enabled us to capture effective 3D structures of the scenes and objects [HSXS13]. This helps the vision to move from 2D towards the 3D vision, like 3D scene understanding, 3D object recognition, and 3D action recognition.

The focus of this work is improving the human action recognition from RGB and depth information by using the big public datasets. The contribution of this paper is represented in two folds: **First**) we used a 3D Convolutional Neural Network (3D CNN) model for recognizing action based on optical flow information. The CNN is used for learning high-level descriptors from low-level motion features (optical flow) by using different input video channels, such as RGB and depth information which are represented by OF-RGB-CNN and OF-Depth-CNN models; **Second**) we are examined the possibilities of feature extraction from 3D CNN and classified by multi-class support vector machine (SVM). This approach is improved that the combination of these two previous models with SVM which outperform the results of each model separately.

The rest of this paper is organized as follows. Section 2 explores the previous work related to this area. Section 3 introduces the proposed approach of the system model. Section Section 4 provides our experimental and results, and section Section 5 concludes the paper.

2 LITERATURE REVIEW

Human action recognition is one of the most interesting topics of computer vision, and it has many use cases within the academy, surveillance, robotics, games, and entertainment multimedia; because of this, the quantity and variety of works is impressive and impossible to cover in a single work. This section is focused on reviewing the works which consider being relevant to this particular problem and to our approach. There are different works which applied deep neural networks to multi-modal learning. Yosinski et. al. [YCBL14] explained a method of how the transferability of features from each layer of a neural network, and exposes their generality or specificity. Further, they evaluated and examined while a layer is general or specific to the training data, and how got good features from these layers transfer to be better than other tasks. For this purpose, several CNN is trained and the weights from different layers are transferred to other networks. Then these networks are trained again, respectively fine-tuned, using different transfer learning strategies such as freezing the weights of certain layers. Simonyan and Zisserman [SZ14] proposed a two-stream CNN architecture using multiple optical flow images computed from RGB video frames for action recognition. A temporal CNN is trained on optical flow volumes, and storing the horizontal and vertical displacement vectors from consecutive action frames, which is finally combined with a spatial CNN trained on RGB frames, to include individual scene and object features. Razavian et. al. [RASC14] improved the possibilities of feature extraction from CNN using the OverFeat network. They extracted feature vectors, respectively the network activations from a fully connected layer, and finally, they have used a support vector machine for the classification task. In addition, they selected datasets and tasks which were different from the OverFeat networks original task, e.g. classification of birds and flowers or image instance retrieval for buildings. Despite these differences, their method has achieved superior performance compared to state of the art methods and proven that pre-trained deep CNN is suitable for generic feature extraction. Athiwaratkun and Kang [Ath15] also improved the possibilities of utilizing features extracted from a pre-trained network and by depending on these features values, they evaluate the quality and performance of these vectors gained form different network layers. These feature vectors are used to train SVM [AEV17] and Random Forest classifiers which even outperform their baseline CNN.

In recent year because of the development of depth sensor, there are different works by using RGB and depth data as input to the CNN. Xinhang et. al. [SHJ17] improved the scene recognition by transmitting pre-trained RGB-CNN models and fine-tuning from RGB to the target of the RGB-D dataset. For RGB-D scene recognition, they combined RGB and depth features by projecting them in a common space and further leaning a multilayer classifier, which is jointly optimized in an end-to-end network. In the other research, A 3-Channel CNN based on rotated 3D points generated from depth maps is introduced by Wang et. al. [PW14]. they are used the weighted hierarchical depth motion maps to store temporal motion information from different views: top, side and front views, for generat-

ing synthesized data, respectively rotation and temporal scaling. The three views are formed to be an input to distinct CNN. Another groups [Wan14] have demonstrated the model of 3D activity recognition from RGB-D data with reconfigurable Convolutional Neural Networks which is handled realistic challenges in 3D data, and it is enabled to perform recognition from it acts directly on the raw inputs (grayscale-depth data) to conduct recognition rather than relying on hand-crafted features. Our deep structured 3D model can be viewed as an extension of these existing approaches, in which make the network could be reconfigurable during learning and inference.

3 PROPOSED APPROACH

The proposed methodology comprises of the major states as shown in Figure 1. The 3D convolution operation is applied to extract spatio and temporal features from video data for action recognition. These 3D feature extractors operate in both the spatio-temporal domains, thus capturing motion information from video streams, as illustrated in next steps:

- The first CNN model has utilized an optical flow representation from RGB videos based on multi-frame dense optical flow (OF-RGB-CNN). This optical flow is computed to hold temporal motion information from temporal domains that fed directly to the CNN (RGB-CNN) to compute the feature vector values which are trained and evaluate directly inside CNN and also these feature vectors are testing with multi-class SVM.
- The second CNN model has used the depth data from corresponding actions of the same RGB video. In contrast to the OF-RGB-CNN approach, the depth-CNN is trained and evaluated in a similar manner.
- Finally, this 3D CNN architecture generates multiple channels of information from adjacent video frames of OF-RGB and OF-depth dimensions and performs sub-sampling and convolution separately in each channel. The final feature representation is obtained by combining information from all channels of both CNN models (OF-RGB-Depth). To explore the possibilities of feature extraction and evaluation with another classifier. The multi-class support vector machine (SVM) classifier is used for this testing evaluation. For this purpose, each CNN serves as a fixed feature extractor. The evaluation of this classification method is then done separately for each CNN model and when combined these two previous models with SVM classifier.

The general structure steps of our human action recognition system are explained in the next subsections in details.

3.1 Preprocessing

3.1.1 RGB Video Preprocessing

The original RGB datasets come in sets of "avi" videos format and have a resolution of 1920 x 1080 pixels. As a first, the video is converted to a sequence of frames and each frame is cut to quadratic size since the subjects appear mostly around the image center. Then each frame is resized to 360 x 360 pixels and for lower computation complexity each frame is converted to a grayscale image.

3.1.2 Depth Video Preprocessing

The original masked depth datasets are coming as the sets of individual frames in "png" format and have the resolution of 512 x 424 pixels. The individual image values are given in millimeters. The masked depth data is already preprocessed and extract foreground data from it. However, the masked depth data still involves challenges. Strong noise can be found in the ground area in all samples and can not easily be removed because of occlusion with feet and legs. This noise could be caused by lighting conditions or camera parameters. Each sample comes in a folder associated with sample number, action ID, camera setup and so on. To keep track of the sample order a shell script is used to sort the samples by their action ID. The frames are first cut to resolution 400 x 400 to further reduce unnecessary image space, also the quadratic shape can be beneficial for the matrix. The image values are then converted from millimeters to the range [0,255], additionally, histogram spreading is applied for better visualization. To reduce memory consumption and training time the images are finally resized to 64 x 64 pixels. The example of this process is shown in Figure 2.

3.2 Feature Extraction

3.2.1 Dense Optical Flow

Optical flow displacement data is generated to capture temporal motion information as in [SZ14, RF16] from RGB and depth data to trained a CNN. Each video sequence is divided into the pairs of consecutive grayscale frames. Then the dense optical flow is computed between each pair of consecutive frames t and $(t+1)$. For the optical flow computation, the Farneb"ack optical flow [Far03] method is applied. The output is two channels image storing the horizontal dx and vertical dy displacement of each pixel location (u,v). Maximum and Minimum values over all frames are used for image normalization. Figure 3 shows a sample optical flow volume and corresponding RGB frames. In this work, the two channel images are further resized to 64x64 pixels and then split into a vertical and a horizontal component. Then, these components are stored in a sequence of image vectors and finally used as input to the CNN.

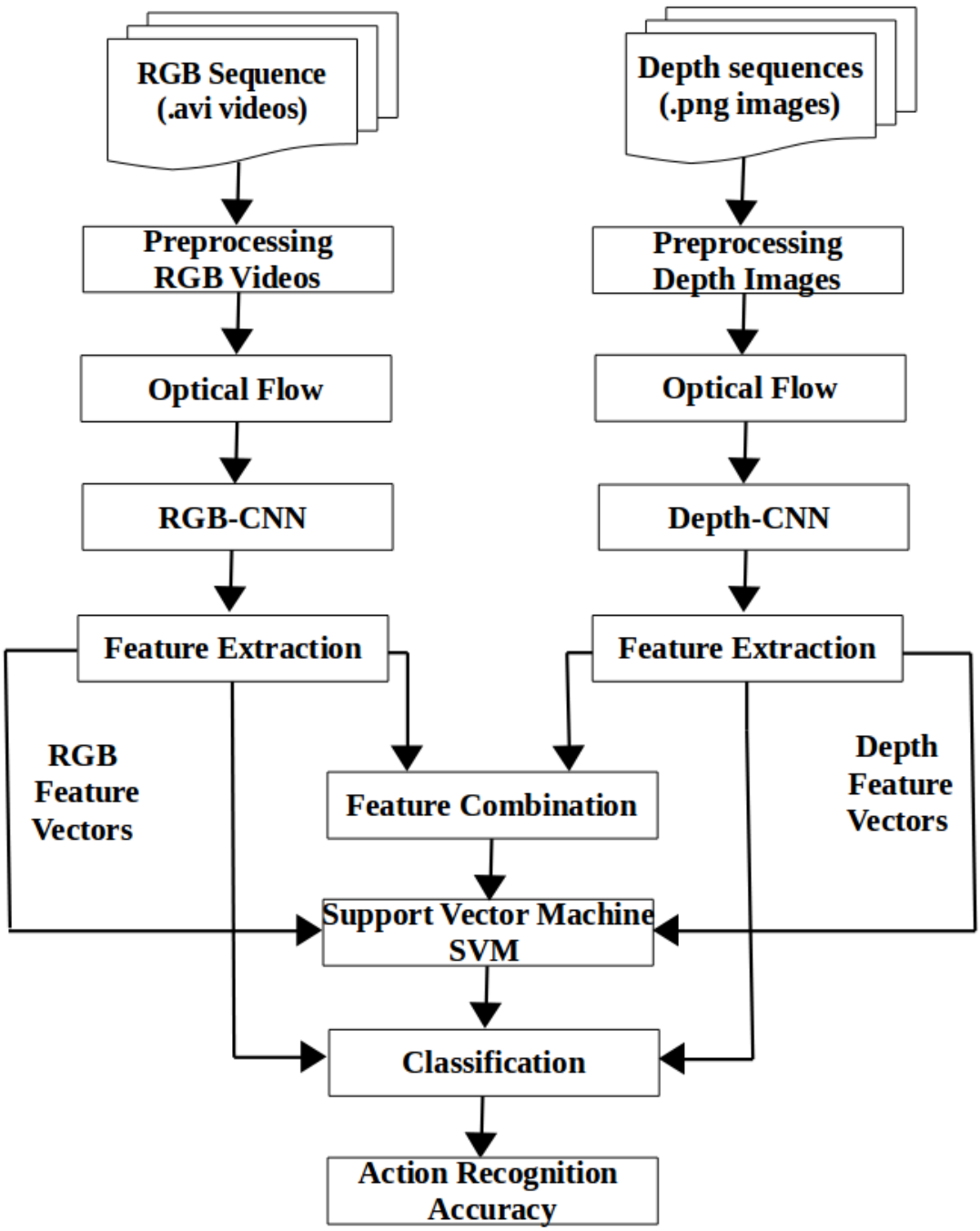

Figure 1: General convolutional Neural networks structure for human action recognition form RGB and Depth video action

3.2.2 Convolutional Neural Networks

The Convolutional Neural Networks (CNN) are representing a hierarchical architecture that can be trained to perform various detection, classification and recognition tasks. A standard CNN consists of two essential components: a feature extractor and a classifier. The feature extractor is used to filter input images into feature maps which are represented a set of features from the images. These features are represented a low-dimensional vector and include corners, lines, edges, etc., which are relatively invariant to position shifting or distortions [CS17]. Then the output from the feature extractor is fed into the classifier, which is usually based on traditional artificial neural networks. In this work, the 3D CNN task involves not only tracking the temporal movement information but also extraction of spatial features. Feature Extraction from 3D CNN describes the process of utilizing the network weights and architecture to fit a new problem. For this purpose, data similarity and data size has to be taken into account.

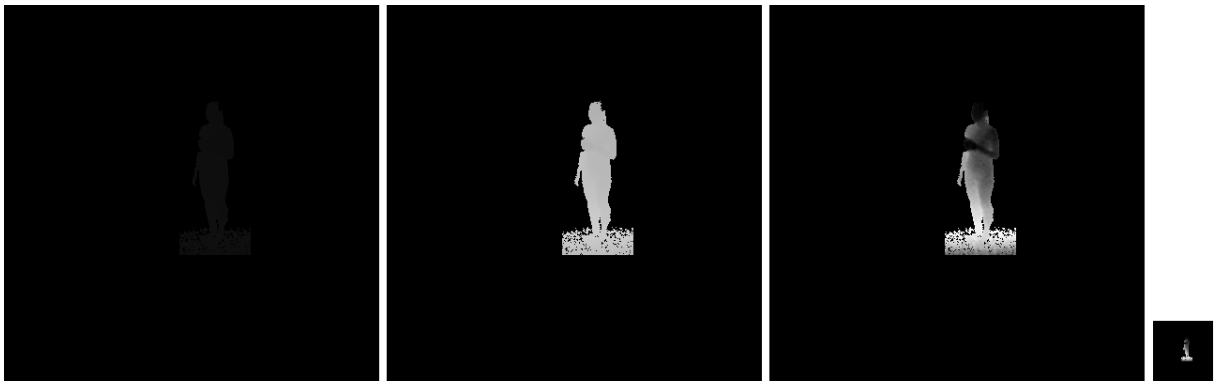

Figure 2: Sample frames of action drinking, from left to right: original depth map, depth map rescaled to range [0,255], rescaled depth map after histogram spreading, final depth map resized to 64x64 pixels.

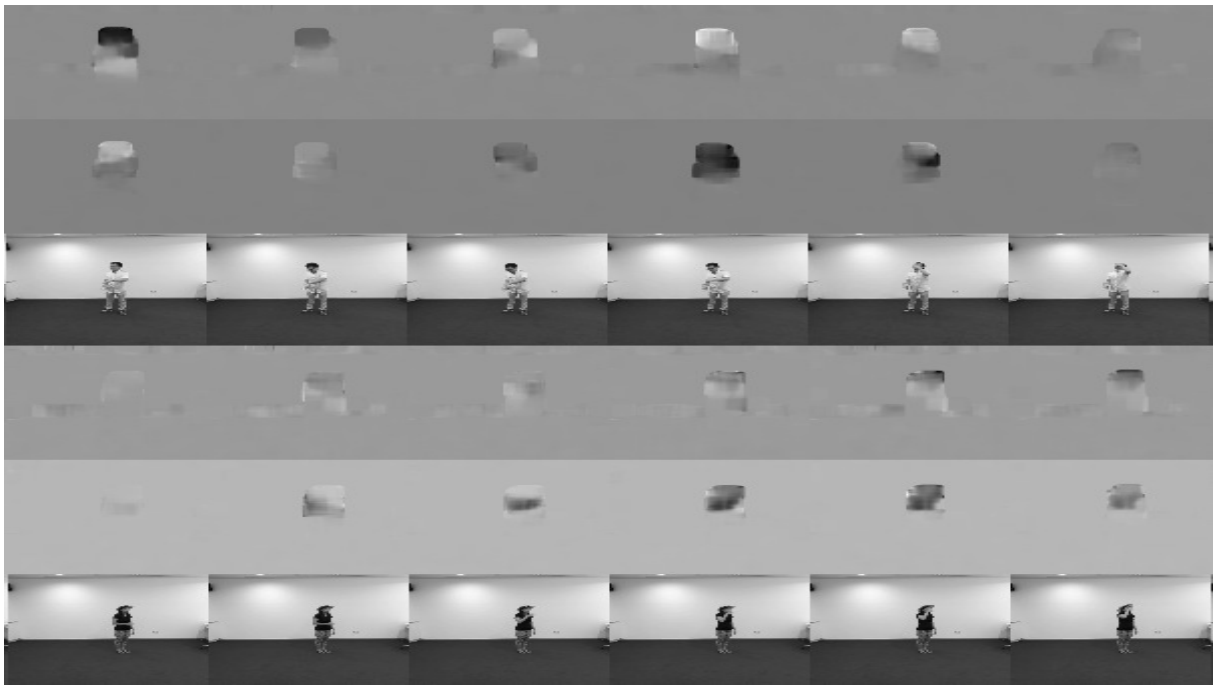

Figure 3: From top: optical flow displacement dx, dy, original gray-scale frame, these frames are extracted in the middle of optical flow volume and corresponding RGB frames, action eating (top) and drinking (bottom)

This CNN system can be trained directly for the task of motion prediction in terms of optical flow for the task of human action recognition. The OF-CNN model which represented in Figure 4, that consists of 4 convolution layers each followed by a max-pooling layer and relu activation. Two fully-connected layers and softmax activation generated the prediction output. In this task, the input takes the optical flow vectors of a single optical flow image volume and treats the individual frames as image channels, and the size of convolution filters is adjusted from 5 to 3. This CNN system can be trained in two different fold, either evaluated directly OF-CNN or by depending on the feature extracted from OF-CNN which are evaluated by using multi-class SVM classifier.

4 EXPERIMENT AND RESULTS

To improve our method and because of CNN require large datasets for training and testing purposes, the Nanyang Technological University's Red Blue Green and Depth information (NTU RGB+D) datasets [SLNW16] is used in order to provide reliable results and the accuracy of the system. The NTU RGB+D is one of the largest scale benchmark dataset for 3D action recognition. It provided 56880 RGB+D video samples of 60 distinct actions. The 60 action classes in NTU RGB+D dataset are presented as: *"drinking, eating, brushing teeth, brushing hair, dropping, picking up, throwing, sitting down, standing up, clapping, reading, writing, tearing up paper, wearing jacket, taking off jacket, wearing a shoe, taking off a shoe, wearing on glasses, taking off glasses, putting on*

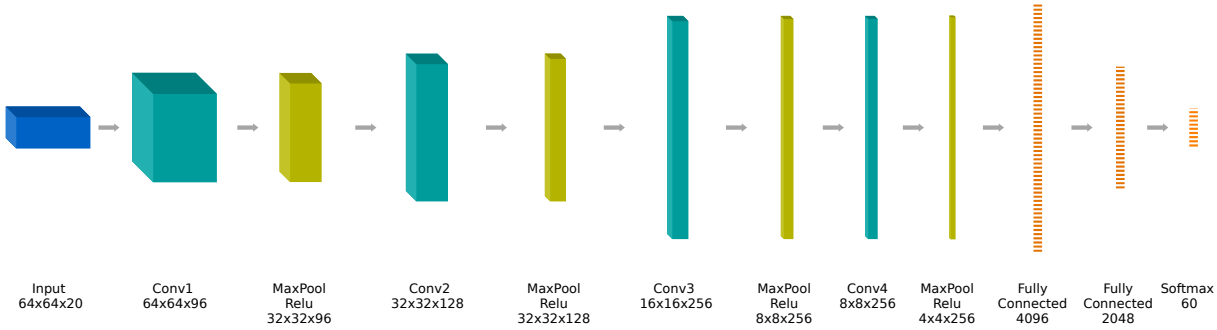

Figure 4: OF-CNN architecture, transformation of input volume, convolutional, pooling and relu layer and softmax output

a hat/cap, taking off a hat/cap, cheering up, hand waving, kicking something, reaching into self pocket, hopping, jumping up, making/answering a phone call, playing with phone, typing, pointing to something, taking selfie, checking time on watch, rubbing two hands together, bowing, shaking head, wiping face, saluting, putting palms together, crossing hands in front, sneezing/coughing, staggering, falling down, headache, touching chest, touching back, touching neck, vomiting, fanning self, punching/slapping other person, kicking other person, pushing other person, patting other's back, pointing to the other person, hugging, giving something to other person, touching other person's pocket, handshaking, walking towards each other, and walking apart from each other". All tested results are done on an Intel Pentium G4600 at 3.6GHz and 8GB RAM, for training a single class of NTU60 model takes 1650s on average, respectively 44 hours (without data loading). See Figure 5 which is showed different action from RGB channel. In this work, only RGB data (136 GB) and masked depth maps (83 GB) are considered. Two different train-test splits for the NTU dataset are proposed. A cross subject split divides the dataset into two groups each containing 20 distinct subjects with 40,320 training and 16,560 test samples, respectively 71% and 29%. The cross view split utilizes the different camera views for each action. The training set contains 37,920 samples (66,6%) with front and side views of the action and the test set holds 18,960 samples (33,3%) with a 45-degree view. Due to time constraints the OF-RGB-depth-CNN, the cross subject split is used in our experimentation results. Our **case study** of this system is illustrated in the next three steps:

- Optical flow is computed from RGB video data, by reducing a single video sequence to 10 optical flow images (OF-volume), this resulting in 20 channel as an input from each video sequence from vertical and horizontal components (d_x, d_y), the image width and height is reduced to 64 pixels. These optical flow volumes are expected to hold temporal motion information suitable for the action recognition task, and they are fed to CNN model for feature extraction and training.

- Optical flow is computed by using depth data from corresponding actions of the same RGB dataset is trained for comparison. In contrast to the OF-RGB-CNN approach, a depth data volume utilizes 10 frames with equal distance which are extracted from the full sequence of depth images. Further, the OF-depth-CNN is trained and evaluated in a similar manner of previous RGB-CNN.

- Both CNN models (OF-RGB-Depth-CNN) are used to explore the possibilities of feature extraction in combination with support vector machine classification. For this purpose, each CNN serves as a fixed feature extractor. The evaluation of this classification method is then done separately for each CNN as well as for a fused model combining feature vectors of both CNN. To explore another common method which can yield further accuracy improvement, the SVM is trained combining feature vectors extracted from the OF-RGB-CNN and OF-depth-CNN models. The larger feature vectors are expected to deliver improved performance. For this purpose the feature vectors are joined by concatenation, to form a single feature vector of dimension 3072 from each video frames. Feature vectors are not further processed, e.g. normalized and rescaled. The multi-class SVM with RBF Kernel is used and setup for training and test procedure.

The comparison results are presented in Table 1, which is compared the previous case study results based on three different models from different input channels

Figure 5: NTU dataset images example [SNGW18]

(RGB, depth, RGBD). From this results, we demonstrate that a CNN with low prediction accuracy can give feature values that yield better classification results with SVM.

Table 2, shows the comparison results with the other state-of-the-art methods using the same datasets.

Model	Modality	Accuracy
OF-RGB-CNN	RGB	40,6%
OF-RGB-CNN-SVM	RGB	**44**%
OF-Depth-CNN	Depth	46,9%
OF-Depth-CNN-SVM	Depth	**50,2**%
OF-RGB-Depth-CNN-SVM	RGB+Depth	**65**%

Table 1: accuracy comparison of all CNN and SVM approaches

Method	Modality	Accuracy
HOG^2 [SNGW18]	Depth	32.24%
Super Normal Vector [SNGW18]	Depth	31.82%
HON4D [SNGW18]	Depth	30.56%
LSTM Encoder-Decoder [Luo17]	RGB	56%
OF-RGBD-CNN- SVM (our)	**RGB+Depth**	**65**%

Table 2: Comparison with the state-of-the-art methods on NTU-RGBD cross subject split dataset

5 CONCLUSION

This paper has presented a deep 3D convolutional neural network based model for classifying and recognizing human actions based on RGB-D data. These models extract features from both spatio and temporal dimensions by performing 3D CNN. The experimental results on NTU RGB+D datasets demonstrate that fusion of different modalities can give better performance than using each modality individually, which mean that the incorporation of RGB and depth modalities to compute 3D CNN feature vectors and supervised learning for the evaluation that yields better prediction accuracy compared to the original CNN. In this work, a support vector machine classifier is used and the accuracy results values outperform the results from baseline CNN in the individual modalities. Training a 3D CNN which provides reliable results and requires not only large datasets but also time, hardware and knowledge. Especially the impact of dataset size is crucial to CNN applications. For Future, We will explore the unsupervised training of 3D CNN models.

ACKNOWLEDGEMENTS

This work was partially supported by Ministry of Higher Education and Scientific Research (MHESR), Iraq, and University of Koblenz-Landau, Germany.

6 REFERENCES

[AEV17] Ana Paula G S De Almeida, Bruno Luiggi M Espinoza, and Flavio De Bar-

ros Vidal. Human action recognition in videos: A comparative evaluation of the classical and velocity adaptation space-time interest points techniques. In *Conference on Computer Graphics, Visualization and Computer Vision WSCG*, 2017.

[Ath15] Keegan Athiwaratkun, Ben Kang. Feature Representation in Convolutional Neural Networks. *arXiv1507.02313 [cs]*, pages 6–11, 2015.

[CS17] Jing Chang and Jin Sha. An efficient implementation of 2d convolution in cnn. *IEICE Electronics Express*, 14(1):1–8, 01 2017.

[Far03] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis*, SCIA'03, pages 363–370, Berlin, Heidelberg, 2003. Springer-Verlag.

[HSXS13] Jungong Han, Ling Shao, Dong Xu, and Jamie Shotton. Enhanced computer vision with Microsoft Kinect sensor: A review. *IEEE Transactions on Cybernetics*, 43(5):1318–1334, 2013.

[JBCS13] Yu Gang Jiang, Subhabrata Bhattacharya, Shih Fu Chang, and Mubarak Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, 2(2):73–101, 2013.

[Lap05] Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[LSR08] Ivan Laptev, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. *Computer Vision and Pattern Recognition, CVPR 2008*, pages 0–7, 2008.

[Luo17] Boya Huang De-An Alahi Alexandre Fei-Fei Li Luo, Zelun Peng. Unsupervised Learning of Long-Term Motion Dynamics for Videos. *CVPR 2017, Computer Vision and Pattern Recognition (cs.CV)*, 2017.

[PD07] F Perronnin and C Dance. Fisher Kenrels on Visual Vocabularies for Image Categorizaton. *IEEE Computer Vision and Pattern Recognition, CVPR '07.*, 2007.

[PW14] Zhimin Gao-Jing Zhang Chang Tang-Philip O. Ogunbona Pichao Wang, Wanqing Li. Action Recognition From Depth Maps Using Deep Convolutional Neural Networks. *IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS*, 46(4):498–509, 2014.

[RASC14] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 512–519, 2014.

[RF16] M. Radolko and F. Farhadifard. Using trajectories derived by dense optical flows as a spatial component in background subtraction. In *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, WSCG*, 2016.

[SHJ17] Xinhang Song, Luis Herranz, and Shuqiang Jiang. Depth CNNs for RGB-D Scene Recognition: Learning From Scratch Better Than Transferring From RGB-CNNs. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, 2017.

[SLNW16] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[SLY16] Ling Shao, Li Liu, and Mengyang Yu. Kernelized Multiview Projection for Robust Action Recognition. *International Journal of Computer Vision*, 118(2):115–129, 2016.

[SNGW18] Amir Shahroudy, Tian-Tsong Ng, Yihong Gong, and Gang Wang. Deep Multimodal Feature Analysis for Action Recognition in RGB+D Videos. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 40(5), 2018.

[SZ14] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *CoRR*, abs/1406.2199, 2014.

[VBM11] Crystian Sadiel Venegas-Barrera and Javier Manjarrez. Visual Categorization with Bags of Keypoints. *Rev. Mex. Biodivers.*, 82(1):179–191, 2011.

[VVV16] Kushal Vyas, Yash Vora, and Raj Vastani. Using Bag of Visual Words and Spatial Pyramid Matching for Object Classification Along with Applications for RIS. *Procedia Computer Science*, 89:457–464, 2016.

[Wan14] Xiaolong Lin Liang Wang-Meng

Zuo Wangmeng Wang, Keze. 3d human activity recognition with reconfigurable convolutional neural networks. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 97–106, New York, NY, USA, 2014. ACM.

[WS13] Heng Wang and Cordelia Schmid. Action Recognition with Improved Trajectories. *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3551–3558, 2013.

[YCBL14] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems 27 (Proceedings of NIPS)*, 27:1–9, 2014.

[YCXL17] Sheng Yu, Yun Cheng, Li Xie, and Shao-Zi Li. Fully convolutional networks for action recognition. *IET Computer Vision*, 11(8):744–749, 2017.