Human Activity Recognition with Convolution Neural Network using TIAGo Robot

Irina Mocanu
Computer Science Department
University Politehnica of Bucharest
Bucharest, Romania
irina.mocanu@cs.pub.ro

Dana Axinte
Computer Science Department
University Politehnica of Bucharest
Bucharest, Romania
dana.axinte@cs.pub.ro

Oana Cramariuc, Bogdan Cramariuc
IT Center for Science and Technology
Research Department
Bucharest, Romania
{oanacramariuc, bogdancramariuc}@yahoo.com

***Abstract*—This paper presents a two layer convolutional neural network for performing activity recognition. We combine spatial and temporal information extracted from images acquired from RGB cameras. Spatial information are extracted from videos by splitting them into RGB channel frames and do a one frame at a time classification. Temporal information from videos are extracted by computing their optical flow. The results are combined in order to build a real time human activity recognition system. The network is tested using TIAGo robot for performing activity recognition. The accuracy of the system is 87,05%, that is comparable with the state of the art. Also, results are obtaining in real time.**

***Index Terms*—activity recognition, convolutional neural network, RGB images, optical flow, CaffeNet network, TIAGo robot**

I. Introduction

The need for monitoring of the daily activities has arisen for a variety of reasons. These include the fact that more and more people live on their own and want a certain degree of safety. Especially older people need to be permanently supervised because of possible accidents. Also, their behavior should often be monitored for checking activities such as administering medications, following medical treatments. Monitoring daily activities could also be used to achieve a style of healthy living. Feedback of the detection system could be reports of the number of activities that were considered healthy during a day, a week or a month or showing user's evolution. This information can help the user for improving their lifestyle.

The development of robotic software that allows reasoning and mimic of human-like behavior opens a world of possibilities where robots are potentially used in multiple ecosystems for human aid and assistance.

In this paper we propose a complex architecture for performing activity recognition through a robotic platform - TIAGo robot from Pal Robotics [1]. The proposed architecture identifies better information from the images, and also the temporal view of the actions. It is based on the model proposed in [2] and build a system composed of two Convolutional Neural Networks (CNNs) by taking into consideration both RGB frames, and also the movement between consecutive frames. The architecture is ported and tested on the TIAGo robot.

The rest of the paper is organised as follows: Section II describes related work. Section III gives the description of the proposed architecture. Evaluation of the proposed network are explained in Section IV. Conclusions and future work are given in Section V.

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CCCDI UEFISCDI and of the AAL Programme with co-funding from the European Unions Horizon 2020 research and innovation programme project "IONIS - Improving the quality of life of people with dementia and disabled persons", project number AAL-2016–074—IONIS-2 within PNCDI III and "Ecosistem de cercetare, inovare si dezvoltare de produse si servicii TIC pentru o societate conectata la Internet of Things" (Cod MySMIS = 105976) - subsidiary contract "Imbunatatirea intelegerii scenei prin recunoasterea unor acivitati umane de catre robotii asistivi", number 1268/22.01.2018.

II. Related Work

A lot of methods for activity recognition are already developed. Some of them are based on feature extraction or hand-crafted features, while others are using machine learning techniques such as CNNs to learn spatio-temporal features. The data representation is also a subject of differences as some methods are based only on RGB video processing In [3], the human activity recognition problem is approached using genetic algorithms and CNNs. A good classification is obtained by using the solutions of genetic algorithms to set the weights of a convolutional network. The input of the CNNs are action bank features for a RGB video [4]. Validation is made on UCF50 dataset [5], where an impressive recognition accuracy of 99.98% is obtained.

Paper [2] contributes in activity recognition by taking advantage of both spatial and the temporal features which exist in a RGB video and proposing an architecture consisting of two stream CNNs which would enhance this features. Any video is composed by two parts, one which consist of the spatial information given by the objects and the scenes present in the video and another of temporal information, containing the motion between frames. Each part is processed by one corresponding CNN and the SoftMax results are combined with late fusion. The input of the spatial stream CNN consist of frames of the video. The classification problem therefore becomes an image recognition problem, suggesting the use of different CNNs which have shown good results [2]. The input of the temporal stream CNN is obtained by stacking optical flow between consecutive frames [2]. The CNN architecture

remains the same as for the spatial CNN. Averaging and using a multi-class linear SVM are two methods of fusion of the CNNs results.

We extend the network from [2] in order to recognise other activities. The main difference consist in the way that information is highlighted from the convolutional layers and porting and testing the network on the TIAGo robot.

III. Description

In order to design the network architecture, we have tested different architecture for the convolutional neural network.

First we have chosen *CaffeNet*, a modification to *AlexNet* network brought by the Caffe framework [6]. We pre-training the network on the objects dataset ImageNet [7]. In 2012 many architectures tried to solve the visual recognition challenge using this dataset, with the algorithms requiring to deliver a list of five of the most certain objects identified in the images. Since then, this is a base point in training neural networks for the job of understanding images, specially when the new dataset has similar elements to it. We consider the pre-training step, already done, using the weights already computed. Two different approaches are considered:

1) the first one in which the the network is fully trained on the dataset.
2) only the final fully connected layers are trained.

The second case is useful for that depends on objects and the correlations would be easier to make. In the case of fine-tuning, the weights for the convolution layers are kept and only the ones corresponding to the three final fully-connected layers are modified. The accuracy greatly improves and a few conditions have to be observed. The actions we propose as input are very dependent on objects which the person interacts with. Only running the pre-trained network already identifies objects from *ImageNet* [7] classes such as piano, acoustic guitar.

The second tested architecture is based on *GoogleNet*, a network with 22 layers. Its main innovation is proposing an optimal component formed of convolutional layers and constructing a structure by spatially repeating this entity. Three convolutional layers and a max-pooling would form the Inception module, whose dimension is later reduced with two 1x1 convolutional layers and 3x3 max-pooling. Modules work in parallel and are later joined. These blocks are small networks, all having the role to identify specific information in images. We chose to have the weights already computed by training with *ImageNet* [7]. Even though the parameters is decreased, in our case, the training duration is still heavy. Same process of fine-tuning is done and good accuracy is obtained.

ResNet [8] an optimization of the deep neural networks architectures, even though the number of layers is increased, reaching the impressive number of 152 layers. The train and test error is decreasing very much. The difference between a normal network, where the pass of the signal is from one layer to another, in a linear way, a ResNet has skip connections that makes the transfusion of the input to another layer later. When an input is received in a layer, instead of computing the resulting volume, a find of what add to the input is done. After the addition, a batch normalization is done. Instead of a new representation directly, a delta on top of the input is computed. It is similar to doing a fine-tune to the results from previous blocks and do not to do the whole training in order to obtain information. This operations form a residual module. Therefore a network is now able to decide how far to send data and how much to increase its depth. For this model we also do fine-tuning of the last fully connected layers. Training the model from scratch or fine-tuning the whole layers would be very costly when the computational resources are small. The model is impressive, and despite that in our case the training time greatly grew, the accuracy increases far from the previously used models. With all these in mind, we chose to keep as the spatial CNN the ResNet architecture as although it requires more training time, it has best accuracy.

The architectures experienced above had the duty to extract and learn features from RGB images which only describe the surroundings of an action by identifying shapes and understanding objects within the first layers. Following this procedure, an action is solely tried to be recognized in a frame. This would not be a good solution as in most cases, the action is a combination of a series of moves and positions, which a single frame is unable to capture. Moreover, actions which do not depend on the objects to interact with are more difficult to observe and some activities have even overlapping frames as the starting, ending or even random captures from the video may be the same.

To take into consideration the *temporal feature* of an action, optical flow is computed for the dataset, based on the architecture presented in [2]. The optical flow images are set as input to a new CNN. Our first test is done with the CaffeNet, pre-trained on ImageNet. We choose not to train from scratch as we previously noticed that a pre-trained network is able to identify shapes rapidly in the first layers. Classification is also done by choosing the most frequent resulted class for each optical flow image of the video of an action. The accuracy improves and we notice that the network becomes able to distinguish actions which are not dependent on objects, just different motions and positions of hands or legs.

We build a system of two CNNs where one has a stream of RGB frames from videos as input and the other receives optical flow images, as given in Figure 1.

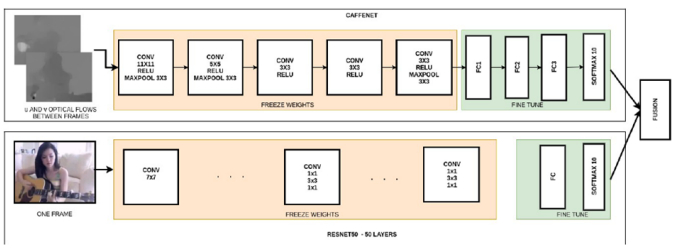

Fig. 1. The proposed architecture.

This approach has the purpose to retain as much informa-

tion as possible from the videos. This way both the spatial environment, where it can be noticed that an human action is often connected to objects or surroundings and the temporal aspect that outlines the fact that actions are composed of multiple poses that flow from one into another are taken into consideration. Firstly, the CNN we use to do the spacial recognition from videos of human actions is the ResNet architecture which is pre-trained on ImageNet dataset. Its input consist of one RGB frame at a time from a video. The net has frozen weights until the fully connected layers and the training is done only by doing fine tuning on these last layers. Secondly, the architecture used to build the temporal CNN is *CaffeNet*. The input now consist in images of optical flow displacement between each two consecutive frames of a video. The training is done the same as in the case of the spatial CNN, fine-tuning the last layers. The two CNNs are then connected by combining their Softmax layers and the classification is done by choosing the majority class from the frames of a video. This step is done with the classification step - that sets the final class of the video.

IV. Evaluation

A. CNN Evaluation

For evaluation we use UCF50 dataset [10] that is a RGB videos dataset collected from Youtube, selecting only the following activities: Drumming, PlayingGuitar, PlayingPiano, PullUps, PushUps. We select only these activities in order to be capable to extend the database with a set of images with ones captured in our laboratory and also to perform the network testing using the TIAGo robot.

We use FFmpeg framework [11], that is an open source multimedia framework for converting video and audio and is portable to multiple operating systems - for splitting the videos from the considered datasets into frames.

An open source OpenCV wrapper [12] is used to compute the optical flow of the videos. The wrapper is based on Lucas-Kanade method. Both the displacements of pixels on X and Y axes given by the motion between two frames are computed.

For implementing the CNN, we use Keras [13], that is a deep learning framework developed in Python and is an abstractization over another two libraries: Theano or Tensorflow. We decided to use this framework for the clear API provided, enabling us to fast write and modify a network and for the multiple pre-trained complex networks available. In our experiments we use the Tensorflow backend due to its vast documentation.

For each system, we start by initializing a model with the desired architecture from *ResNet50* and *GoogleNet*. Therefore, another small model is set on top with the role of a fully-conected layer. The small model is created of Flatten and Dense models. The weights of the previous layers are frozen, except for the last convolutional block of each architecture. The new model is compiled with Stochastic Gradient Descent as the learning algorithm, too. Learning rate is set to 0.001. Training is done with batches of 16 images, learning 4096 images in 256 iterations per epoch. 50 epoch are done to observe the convergence for each model, with a duration of 32 min per epoch when training on GPU, unoccupied with other tasks. Training time is greater than in the previous experiments due to the considerable complexity of the models. Although the number of iterations may seem small, it takes longer to process due to the depth of the models. This number is also reduced by the fact that fine-tuning is done only for the last layers, as training the whole networks would be time consuming. The greater accuracy is reached with the *ResNet* network. The final architectures of our action recognition system is made of a *ResNet50* CNN which is trained on RGB frames from videos and a *CaffeNet* CNN trained on optical flow images. We also implement the latter model in Keras and their fusion is done using a merge Concatenate layer beween the SoftMax layers of each convolutional network, giving so the final label of the input. All the parameters are kept the same as previous. The results for every frame and optical flow of a video are gathered and the most common class predicted is set as the final label of the video.

B. Porting the CNN to the TIAGo Robot

TIAGo robot (from Pal Robotics) is a robot designed for indoor environments. The robot has a height between 110 and 145 centimeters (depending on the extension of the torso) and weights approximately 70 kg. It presents features such as mobility and perception, being equipped with a robotic arm ending in a parallel gripper that offers object manipulation capabilities. These features makes TIAGo an optimal platform for conducting research experiments in robotics for human environments and machine learning areas.

The software capabilities of TIAGo include navigation applications for mapping and localization for indoor environments, human detection and obstacle avoidance. The interaction with a human factor is supplied by text-to-speech and speech recognition in English, remote control with tablet and telepresence. The RGB-D camera makes possible real-time object recognition and pose estimation but also human detection and recognition.

We tested the activity recognition module using the TIAGo robot. Thus we create a system composed of the TIAGo robot and a laptop, both running the Robotic Operating System (ROS). The laptop is mounted on top of TIAGo and connected to it through an Ethernet cable. The machines run a distributed instance of ROS, with the master node running on TIAGo. The first part of the pipeline is a ROS node on TIAGo captures the images outputted by the robots integrated cameras. The node subsequently publishes the captured images to a topic. A separate ROS node on the laptop subscribes to this topic to obtain the images, as given in Figure 2. The user performed activities in front of the robot. Thus, the robot doesn't need to track the user in order to capture the images.

For our final system, training accuracy increases rapidly as common shapes are quickly identified and correlated with the specific actions. We observe in Figure 3 that validation accuracy greatly improves, reaching an average of 85.06%.

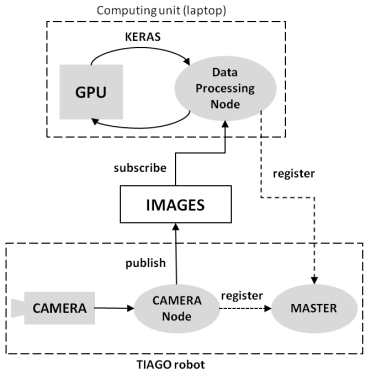

Fig. 2. Communication with the TIAGo Robot

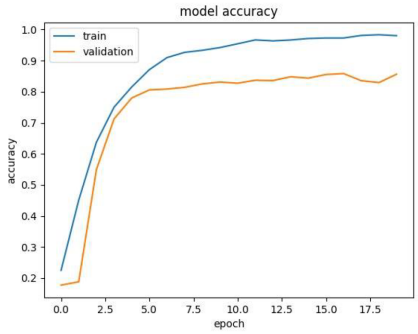

Fig. 3. The accuracy of our final system composed of a spatial ResNet50 CNN and a temporal CaffeNet CNN

We observe in Table I that validation accuracy also greatly improves, reaching an average of 87,05%, ahowing that the proposed model based on two CNNs comes close to the best recent results.

TABLE I
MY CAPTION

Method	Accuracy
Caffe Net fine tune	61,3%
Google-Net fine tune	74,7%
ResNet50 fine tune	78,14
ResNet50 + CaffeNet fine tune	**87,05%**
Two Stream CNNs [2]	88%

With this system the greatest accuracy is obtained, coming close to best recent results on UCF50 [5] for the considered activities. We observe what a great prediction our system performs in the cases of actions dependent on objects. Due to the previous train on *ImageNet* [7], the spatial network from our architecture is able to quickly determine shapes of known objects such as piano, guitar, drum. Thus, the network associates this activities with recognizing the environment surrounding them. The final systems does a classification per frame of 0.0012s and total of 0.06s per video, which we consider to be appropriate for solving the problem. However, a latency is introduced in our system by the requirement of processing the optical flow images of a video, step which brings a 0.05 additional duration to our recognition for our tests. This process can be rushed though the computation on GPU.

V. Conclusions and Future Work

Recognition of human daily activities from videos has become a main research field in terms of Computer Vision. Its application is introduced in real world tasks such as rehabilitation with physical medicine, security by video surveillance and sports videos analysis. We proposed a convolutional neural network based on the *ResNet50* CNN which does activity recognition from RGB images and a *CaffeNet* CNN which manages to classify optical flow images. The proposed network was tested on the TIAGo robot for performing activity recognition.

As future work we improve the fusion step between the two network architectures. An activity which is tried to be recognized by our system is labeled based on the classes of the predicted corresponding frames. However, not every frame of the video is relevant to the action itself, for instance the starting and ending frames of a video may not entirely describe an action, but different standing poses. Therefore, an heuristic for classifying an action has to be developed. Another improvement which can be bought to our architecture is performing fine-tuning for each layer of the networks used. Currently, fine-tune is done at the last fully connected layers, but a full training would probably increase the knowledge the networks gain and accuracy would improve.

References

[1] TIAGo Robot, http://tiago.pal-robotics.com/, last accessed April 2018.
[2] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos", In Advances in neural information processing systems, pp. 568-576, 2014.
[3] Earnest Paul Ijjina and Krishna Mohan Chalavadi, "Human action recognition using genetic algorithms and convolutional neural networks", Pattern Recognition, vol. 59, pp. 199–212, 2016.
[4] Sreemanananth Sadanand and Jason J Corso, "Action bank: A high-level representation of activity in video", IEEE Conference on Computer Vision and Pattern Recognition, pp. 1234–1241, 2012.
[5] Kishore K Reddy and Mubarak Shah, "Recognizing 50 human action categories of web videos", Machine Vision and Applications, vol. 24, issue 5, pp. 971–981, 2013.
[6] Caffe framework: http://caffe.berkeleyvision.org/, last accessed April 2018.
[7] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet Large Scale Visual Recognition Challenge", International Journal of Computer Vision, vol. 115, issue 3, pp. 211-252, 2015.
[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.
[9] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9, 2015.
[10] Kishore K Reddy and Mubarak Shah, "Recognizing 50 human action categories of web videos", Machine Vision and Applications, vol. 24, issue 5, pp. 971-981, 2013.
[11] FFMPEG: https://ffmpeg.org/, last accessed April 2018.
[12] OpenCV: https://opencv.org/, last accessed April 2018.
[13] Keras: https://keras.io/, last accessed April 2018.