# Deep Learning Project: Interim Deliverables Video Based Action Recognition Using Deep Learning Models

Aitore Issadykova
*Data Science*
*Nazarbayev University*
Nur-Sultan, Kazakhstan
aitore.issadykova@nu.edu.kz

Assem Kussainova
*Data Science*
*Nazarbayev University*
Nur-Sultan, Kazakhstan
assem.kussainova@nu.edu.kz

Zhaniya Koishybayeva
*Data Science*
*Nazarbayev University*
Nur-Sultan, Kazakhstan
zhaniya.koishybayeva@nu.edu.kz

***Abstract*—Convolutional Neural Networks (CNN) is one of the most popular tool for video-based action recognition purposes and used in feature extraction part and in classifiers part. Due to runtime costs of CNN, researchers prefers to use it either in data pre-processing part or in the classifier part. This project examined the performance of CNN in the video based action recognition for different sport activities from KTH and UCF Sports datasets. Models architectures include the Histogram of Gradients (HOG) as feature extractor and CNN as classifier. The preliminary experimental results with specified parameters are expected to be improved for the final stage of the project.**

***Index Terms*—Deep Learning, CNN, HOG, Canny Edge, Classification, Video Based Recognition, Features Extraction**

## I. Introduction

Image processing was developing exponentially during the last few decades because of the fast performance improvements in deep learning architectures. Convolutional Neural Networks (CNN) are considered to be the most popular tool for extracting the features from images and for performing classification. One of the main application of the CNN classification is the video based object recognition, which is a foundation of many developing technologies in robotics, computer science and engineering fields. The most popular video based classification applications are : advanced face and voice extraction and classification from real time video is used for security purposes and data access granting, and artificial intelligence systems' real time learning is used for object detection around the system, which is used in self-driving cars . This paper will examine the performance of neural networks in the classification of human actions with additional feature extraction tools. The research are conducted with most popular action recognition datasets : KTH and UCF Sports. Models are based on simple feature extractions like HOG and Canny Edge and neural network classifiers will be examined on efficiency which is runtime of the model and accuracy of the model. The future improvements and further works will be discussed for both models.

## II. Related Work

Convolutional Neural Networks have proven highly successful at static image recognition problems. By using feature pooling operations, CNNs are capable of automatically learning complex features required for visual object recognition tasks achieving superior performance to hand-crafted features. Encouraged by these positive results several approaches have been proposed recently to apply CNNs to video and action classification tasks.

Related to this field research has been done by Ng et al. [3] where authors use an approach of processing video files as frames using recurrent neural network with max-pooling that uses Long Short-Term Memory classificator which is connected to the output of the CNN. Authors used two CNN models for processing each individual frame of the input video files per second from Sports-1M and UCF-101 dataset: AlexNet and GoogleNet. In addition, for better capturing of motion information without loss an optical flow is added to the model. As a result, best achieved accuracies of action recognition for Sports-1M and UCF-101 dataset were 73.1% and 88.6% respectively.

Later Ullah et al. [6] improved previous approach performance in action recognition software by using features of the CNN and processing it through bidirectional LSTM algorithm with increasing the number of layers in the neural network models. Authors derived features from the UCF-101 dataset video frames with six frame jump, which are after sent to DB-LSTM in portions within time interval, where two layers are stacked on both forward and backward pass of the LSTM. In comparison with earlier work, current model is capable of learning long term complex sequences in videos and the proposed method improved the recognition rate on UCF-101 dataset from 88.6% to 91.21%.

Two most popular action datasets that are used for video-based classification purposes are KTH and UCF Sports. There are a lot of studies that used two datasets and built several distinctive models with high accuracies. One of such studies was conducted by Sargano and others [4]. Authors used the well

known ImageNet CNN model for feature extraction and built a hybrid SVM-KNN classifier. The proposed model achieved 98.15% and 91.47% accuracies for KTH and UCF Sports datasets, respectively, with Leave-One-Out cross validation.

The research conducted by Beikmohammadi and others [1] used improved 14 layers CNN- MobileNet, which was trained on ImageNet databases, for feature extraction. Since the classifiers such as SVM, LSTM, KNN, CNN and others are claimed to be costly for this research, authors used simple logistic regression for the classification purposes. The models used only 7 selected frames from the video for training and testing and achieved 98.81% and 96.47% accuracies for KTH and UCF-Sports respectively.

The past researches successfully used CNN for the feature extractions of the video based action dataset. Studies conducted with KTH and UCF-Sports dataset mostly used not Neural Networks classifiers, i.e. SVM, KNN, etc. This project aims to build classifiers based on CNN and compare results with other studies.

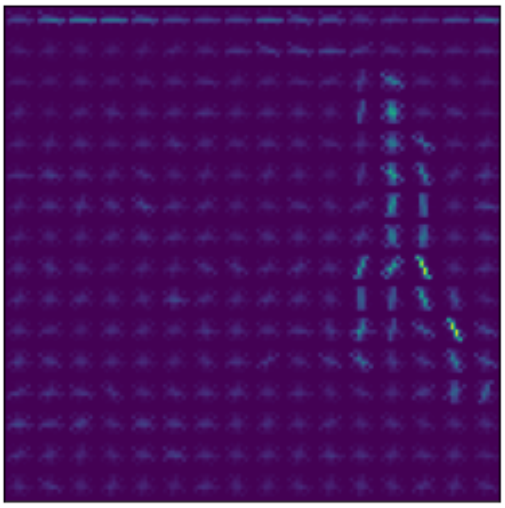

Fig. 1. HOG feature extraction for KTH dataset

## III. Datasets

For this project we chose to work with two datasets suggested in the proposal: UCF sports action and KTH.

### A. UCF sports action

UCF sports action dataset was downloaded from the official website of the Center for Research in Computer Vision [5]. The dataset represents video and image files of sports activities that are broadcasted through different media channels including ESPN and BBC. This collection of data includes 150 video sequences (almost 15 minutes), where each video file has a resolution of 720 x 480. In addition to the main viewpoints and perspectives of the objects in the video, additional sides and angles of the shooting were provided for some of the sports activities. The dataset has been used in research and projects, which are directed to the fields of action recognition, action localization, and saliency detection. The dataset includes 10 sports actions: Diving (14 videos), Golf Swing (18 videos), Kicking (20 videos), Lifting (6 videos), Riding Horse (12 videos), Running (13 videos), SkateBoarding (12 videos), Swing-Bench (20 videos), Swing-Side (13 videos), Walking (22 videos).

Before calculating HOG features it was necessary to preprocess the images to obtain square-blocks of histograms which are used for a fast HOG calculation. The preprocessing helped feature extraction algorithms to better describe the image features. At this stage the images are cropped according to given ground truth and resized to the new size of 64*128.

### B. KTH

KTH is an open access dataset, publicly available on the website [2]. The dataset was first introduced in 2004, currently it consists of 600 .avi videos of all combinations of 25 subjects performing six actions (walking, jogging, running, boxing, hand waving and hand clapping) in four different scenarios ( outdoors, outdoors with scale variation, outdoors with different clothes and indoors). The database consists of 2391 sequences shot on a static camera with 25 fps and 160x120 pixels resolution.

As a preprocessing part of the dataset, the deletion of empty frames, edge detections, resizing of frames and normalizing data were performed. Since the videos contain the “empty” frame, which does not contain a human, the Gaussian blurring and Canny Edge Detection were used for deleting empty frames and for edge detection for future classifier model. Each frame was resized to size 128*128 and then it was normalized. The categorical type of classes was converted to a binary array for training purposes.

## IV. Methodology

In the preliminary stage we created two simple classification models for KTH and UFC Sport datasets separately. Both models will be improved also separately and then results will be compared with each other. Initial models used simple edge detection/ feature extraction using HOG and Canny edge approaches. Moreover, it was decided to separate models due to limits of keras 3D CNN functions which can be runned only with GPU and not TPU.

### A. KTH based model

For the KTH dataset we used only 70 videos of each category and limited the number of frames to be kept to 12. These numbers were chosen due to computation limits for the dataset and computational power of GPU accelerator of Google Colab. Fig. 1 represents the frame from the running video after Canny Edge and HOG feature extraction.

*1) Architecture of classifier:* For the KTH dataset-based model was constructed an 9-layers CNN model. The architecture of the model is following: first is Convolution3D layer with “relu” activation function and with “channels first” data format. Next layers are MaxPooling3D, Dropout with

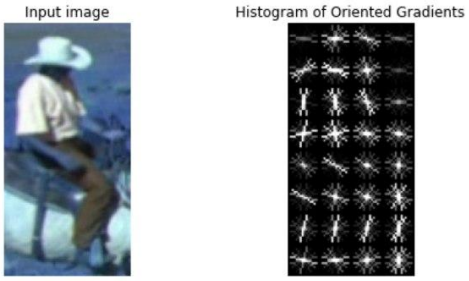


Fig. 2. Input image and HOG for UCF Sports

rate 0.5, Flatten and Dropout with rate 0.5 Layers. Then the model adds a Dense layer with kernel initializer "normal" and activation function" relu" and again a Dropout layer with 0.5 rate. The last two layers are the Dense layer with kernel initializer "normal" and "softmax" Activation layer. The whole model is compiled with categorical cross-entropy loss because of multiclass classification problem, RMSprop optimizer was chosen due to relatively small size of batch and accuracy was chosen as metrics.

*2) Final Model:* The initial setup for this model used Canny edge detection procedure and HOG descriptor and then sent the data to the classifier directly. The final model of the video-based action recognition the additional feature detection will be added. Simple descriptors such as HOG and SIFT will be evaluated on performance by comparing with ImageNet. The more optimal in terms of running time feature extraction model will be added to the final architecture. Moreover, the parameters of classifiers will be iteratively changed towards the final model in order to get the optimal solution.

### B. UCF Sports based model

For the UCF sports action dataset we used all the available data since the Google Colaboratory services provide sufficient TPU for working with large amounts of data. Before applying data to the model, the Histogram of Oriented Gradients was implemented for the simple extraction of features from video frames. HOG works according to the following principle. In a previously transformed image, it is needed to look at each pixel to determine how dark the current pixel is compared to the pixels directly adjacent to it. An arrow is then drawn showing the direction in which the image becomes darker. If this process is repeated for each individual pixel in the image, then eventually each pixel will be replaced by an arrow. These arrows are called the gradient and they show the flow from light to dark throughout the image. Thus, we get the basic structure of the image through the streams of light and dark. Feature vector length for this implementation is 780*3=11340. Fig. 2 illustrates the example of application of HOG feature extraction method on Riding-Horse image.

Next, the Convolutional Neural Network was built for action recognition purposes. The initial parameters for the models for first testing were set based on general recommendations for the CNN models. For our model the initial parameters are following:

- Filters: 64
- Kernel size: 5
- Strides: 2
- Activation: Relu

*1) Architecture of model:* For this dataset was implemented a Convolutional Neural Network with 10 layers, which includes 4 Convolution2D layers, 1 Dropout layer, 1 Flatten layer, 3 Dense layers and 1 Activation layer. All convolution layers are activated using Relu function, have kernel size equal to 5 and strides = 2. Next, data goes to the Dropout layer with rate 0.5, and Dense layers with Relu activation functions are added to the network. After data passes these layers, final activation is made using the 'softmax' function, from which the model is finally compiled. This step is done with 'mse' loss and 'sgd' optimizer, which has the learning rate = 0.01, and for metrics was chosen 'accuracy' parameter. Fitting of data to the model goes through 5 epochs.

*2) Further improvements:* Initially, features of data are extracted using HOG and transferred to the network for action prediction. In the next step, more complex feature extraction methods mentioned in the project proposal will be implemented for results comparison, and other architectures of models will be tested and compared by accuracy score. Better parameters will be identified for current model architecture in order to increase the accuracy of action recognition. All the results will be summarized and described in the final report.

### C. Evaluation Metrics

For the preliminary part we used Leave-One-Out (LOO) cross-validation scheme for UCF Sports Action dataset and train/test split scheme for KTH dataset. Both of them will be used in final models for presenting our final project report. For the final model accuracy was chosen as the main metric and other major factor which is runtime efficiency will be included in the final result analysis and model selection.

## V. Experimentations and Results

### A. KTH based model

The data was divided for train and validation datasets using train_test approach with 8:2 division ratio. Due to the small amount of training data, batch size was used as 15 and in order to not overtrain the neural networks number of epochs was chosen as 20 (For general CNN it is recommended to use less than 21 epochs). As for CNN model parameters, the number of filters, pool size and level of convolution were set as 32, 3 and 5 respectively.

The initial setup with described above parameters provides 81.85% accuracy for train data and 72.62% accuracy for test data (Fig. 3).

As it can be seen from the confusion matrix on Fig. 4, the running, walking, and jogging can be misclassified to each other class. However, the overall accuracy of the dataset is good and can be increased with additional feature extraction method.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.87 | 0.81 | 0.84 | 16 |
| 1 | 0.71 | 0.71 | 0.71 | 14 |
| 2 | 0.76 | 0.87 | 0.81 | 15 |
| 3 | 0.54 | 0.58 | 0.56 | 12 |
| 4 | 0.83 | 0.59 | 0.69 | 17 |
| 5 | 0.62 | 0.80 | 0.70 | 10 |
| accuracy | | | 0.73 | 84 |
| macro avg | 0.72 | 0.73 | 0.72 | 84 |
| weighted avg | 0.74 | 0.73 | 0.73 | 84 |

Fig. 3. KTH metrics results


Fig. 4. Confusion matrix KTH

As part of model testing, the previously unseen running video was tested by model.The classifier classified the running action correctly despite the fact that accuracy of the running class is low. It is expected to decrease misclassification of the classes by tuning parameters and updating architecture of the whole model.

### *B. UCF Sports based model*

The results of the current model for the whole training data achieved the accuracy of 92.7%, and then it was applied to the test data. In the testing step, each sport action class was separately checked and accuracy for each of the sport types was obtained. Following is the list of each sport action class and corresponding to it accuracy score.

From the results on Fig. 5 it can be concluded that even if the accuracy of the model is quite high for the whole data, when it comes to recognition for each class, the accuracy score is significantly dropped. Therefore, further improvements should be made for obtaining better results.

## VI. Working plan

The working plan is the same as was provided for the initial project report. The remaining timeline can be reviewed in the graph on Fig. 6.

```
Accuracy of label Kicking-Front is
0.0
Accuracy of label Swing-SideAngle is
0.807176923077
Accuracy of label Golf-Swing-Back is
0.31666
Accuracy of label Run-Side is
0.3006
Accuracy of label Golf-Swing-Front is
0.327075
Accuracy of label Swing-Bench is
0.736
Accuracy of label Kicking-Side is
0.03913
Accuracy of label Diving-Side is
0.724671428571
Accuracy of label Walk-Front is
0.439272727273
Accuracy of label Riding-Horse is
0.402766666667
Accuracy of label SkateBoarding-Front is
0.554766666667
Accuracy of label Golf-Swing-Side is
0.27
```

Fig. 5.

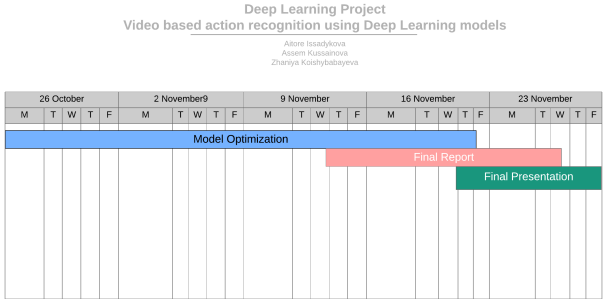


Fig. 6. Project schedule illustrated by Gantt chart

## VII. Conclusion

This paper is focused on classification of human action from video by CNN as classifier. For project purposes two models were build with using HOG as feature extractors and two different architectures for CNN as classifier layer. Both models performance were examined on the KTH and UCF Sports datasets separately. It is expected that future works will present the each model performance for both datasets and will be improved by tuning parameters, applying more advanced feature extraction and possible CNN architecture changing. The final stage of the project will be present an optimal model for KTH and UCF Sports dataset which will use CNN as classifier layer.

## REFERENCES


[1] Beikmohammadi, A., Faez, K., Mahmoodian, M. H., Hamian, M. H. (2019, December). Mixture of Deep-Based Representation and Shallow Classifiers to Recognize Human Activities. In 2019 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS) (pp. 1-6). IEEE.

[2] KTH dataset. Available at: https://www.csc.kth.se/cvap/actions/.

[3] Ng, Joe Hausknecht, Matthew Vijayanarasimhan, Sudheendra Vinyals, Oriol Monga, Rajat Toderici, George. (2015). Beyond short snippets: Deep networks for video classification. 4694-4702. 10.1109/CVPR.2015.7299101.

[4] Sargano, A. B., Wang, X., Angelov, P., Habib, Z. (2017, May). Human action recognition using transfer learning with deep representations. In 2017 International joint conference on neural networks (IJCNN) (pp. 463-469). IEEE.

[5] UCF sports action dataset. Available at: https://www.crcv.ucf.edu/data/UCF_Sports_Action.php.

[6] Ullah, Amin Ahmad, Jamil Muhammad, Khan Sajjad, Muhammad Baik, Sung. (2017). Action Recognition in Video Sequences using Deep Bi-directional LSTM with CNN Features. IEEE Access. PP. 1-1. 10.1109/ACCESS.2017.2778011.