Deep Learning Project: Final Report 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) are one of the most popular tools for video-based action recognition purposes and used in feature extraction part and in classification 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 and gesture recognition in RusLan dataset. The study provided the general CNN model for classifying different types of video datasets. With the general model the accuracy of about 77% and 74% was achieved for well-known UCF Sports and KTH datasets, respectively. The new RusLan dataset performed with 75% accuracy, which can be set as the benchmark for the future works. All results were obtained with common structure of the prepocessing: MobileNet feature extraction and general CNN model. Moreover, additional specific setup for UCF Sports achieved almost 95% of accuracy.**

***Index Terms*—Deep Learning, CNN, Canny Edge, Classification, Video Based Recognition, Features Extraction, KTH, UCF Sports, RusLan**

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. The goal of this paper is to examine the performance of neural networks in the classification of human actions with additional feature extraction tools. The research will be conducted with well-known action recognition datasets: KTH and UCF Sports. Additionally, the performance of our own dataset of Russian Sign Language from RusLan, Russian Sign Language corpus, will be analyzed. The general model architecture with identical preprocessing and feature extraction stage will be applied for each dataset and results will be compared with existed works.

II. Related Work

Convolutional Neural Networks have proven being 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. [6] 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. [12] 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 [9]. Authors presented human action recognition method based on transfer learning using a CNN with SVM-KNN architecture which deals with overfitting problem. As baseline architecture for CNN feature extractor was used AlexNet trained on ImageNet dataset, consisting of 5 convolutional and 3 fully connected layers. After that, hybrid classification model on SVM and KNN algorithms was applied for action recognition on video. 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 [2] 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.

Ravanbakhsh et al. [7] proposed method which captures human motion through a hierarchical structure. Using video files from KTH, UCF Sport and UCF-11 Human Action datasets, authors extract frames from each video and features are computed for each frame using CNN, which are later mapped into a short binary code space. Key-frames are selected using the changes in each bit of the binary codes in given time and fed into hierarchical decomposition. For dimension reduction PCA method is applied for all levels, which are saved as vectors for a video snippet. Finally, the histogram of temporal words for all the videos is built and SVM classifier is used to predict action labels.

Charalampous, K. and A. Gasteratos [3] described an unsupervised on-line deep learning algorithm for action recognition in video sequences. Authors used publicly available datasets, including KTH and UCF Sports. The proposed method applies clustering algorithm and forms representation vectors from learned space. Video files were divided into frames, from which the spatio-temporal features were extracted suing Viterbi Algorithm, which is used for assignment of indices according to similarity between representation vectors and transition probability. In addition, for described feature extraction from files were utilized essential ART-2 clustering algorithm and L1-norm minimization methods.

The study of Shamsipour, Ghazal and Pirasteh, Saied [10] aims to contribute a method of artificial intelligence and CNN for recognition of human interaction by video collected from drone. Authors used SVM method to classify the stationary features obtained from pre-trained CNN with ImageNet along with application of PCA to every window from frame, and tested obtained model on the UCF Sports Action data with sufficient accuracy score which has indicated improvement of more than 3.17% as compared to the previous methods.

Basha et al. [1] introduced method which aims to classify human actions in a video, where the videos are captured at a distance from the performer. Researchers trained and evaluated the proposed 3D CNN model on KTH and WEIZMANN datasets. In this method frames from video were inputted to the 3D CNN model where spatio-temporal features were extracted. CNN model generates feature vector, which transferred to LSTM model, consisting of 1 hidden layer, which gathers decisions of 4 neighboring frames to classify an image according to sport actions from videos.

Previous researchers successfully used CNN for the feature extractions of the video based action dataset. Studies conducted with KTH and UCF-Sports dataset mostly used classic Machine Learning classifiers, such as SVM, KNN, etc. This project aims to build classifiers based on CNN and compare results with other studies.

III. Datasets

We chose to work with three datasets suggested in the proposal: UCF sports action, KTH and RusLan.

A. UCF sports action

UCF sports action dataset was downloaded from the official website of the Center for Research in Computer Vision [11]. 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 salience 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).

B. KTH

KTH is an open access dataset, publicly available on the website [4]. The dataset was first introduced in 2004, currently it consists of 598 .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.

C. RusLan

Russian Sign language corpus [8] includes spontaneous speech (monologues and dialogues), texts recorded on the basis of stimulus materials (retellings of cartoons, stories from pictures), as well as materials partially obtained through questionnaires. More than 230 video texts from 43 Russian

sign language speakers were recorded and annotated into this dataset. The corpus includes texts from native speakers of the Russian Sign Language - men and women aged 18 to 63 with varying degrees of deafness: deaf, complicated hearing and CODA (Child of a Deaf Adult - individuals who are able to hear and are children of deaf parents, with sign language being their first language). Some of the texts are studio video recordings, but not all informants had the opportunity or willingness to be recorded in the studio, so the corpus also includes texts recorded in classrooms or at home. Each video file is named in the format "language and place of text recording-type of text-informant code-destination code-type of markup".

Our RusLan dataset consists of the extractions from the Russian Sign Language Corpus. 8 words (classes) were derived for validating the performance of the proposed general model, and trimmed out of the corpus. The 8 classes are: "Bird", "Cat", "Mom", "Dad", "Me", "Pipe", "Woman" and "Look". Each class is represented by 5 speakers showing each word 4 times, 20 instances overall. The whole newly derived dataset has 160 videos.

IV. Methodology

A. Data Preparation

Before starting to work with building a deep learning model, the data from the selected datasets went through a preparation stage. Here it was required to transform data according to the goals and requirements of neural network models.

From the previously studied training data, the main KTH dataset was not changed, while for the UCF Sports Action dataset, it was necessary to combine data from some sports classes, since these classes contained repeating sport types, but shooting was made from different angles. For example, 3 perspectives of golf video files (front, side and back) have been grouped into one general class Golf-Swing. In addition, since video files were missing in some class folders (ex. Diving-Side/014), but contained ready-made frames obtained from the video, they were used as an input for training data.

To form the classes of RusLan dataset, we have gone through the whole corpus and pointed out the most frequent words occurring in the text documents attached to each video speech. Some classes were denied because of the inconsistencies in the gestures of the speakers (for example, there are several variations to say "man" in Russian Sign Language), and because some words were not spoken by a speaker enough times. After the 8 final classes were determined, videos with chosen words were downloaded from the corpus and trimmed. Additionally, additional procedures of cutting frames were performed to make the background uniform and eliminate any visual noise (people walking at the back and operator's hand popping out at the edge of the frame).

B. Data Preprocessing

For experiments each video from datasets was stored as sequence of the particular number of frames. The number of frames was determined as minimum number of frames in all video within one dataset or was taken as maximum amount which RAM can processed as for KTH dataset. Moreover, in order to obtain some dynamic changes in the frames, the defined number of frames was skipped. The final list of frames was normalized and reshaped into 64*64 size.

Additionally, a list of videos from KTH dataset contains empty frames for active sports type such as walking, running and jogging. Empty frames are frames without human on the particular frame. Such frames were detected with Canny Edge detection method and were not included in the final data.

C. Feature Extraction

To determine the features of incoming frames from the video, it was required to find a solution for the model to be simultaneously fast, high-quality and economical in resources when processing a large number of images. The MobileNetV2 neural network met these criteria, the additional advantage of which is improved performance, compared to older models of neural networks which are also used to extract image features [5].

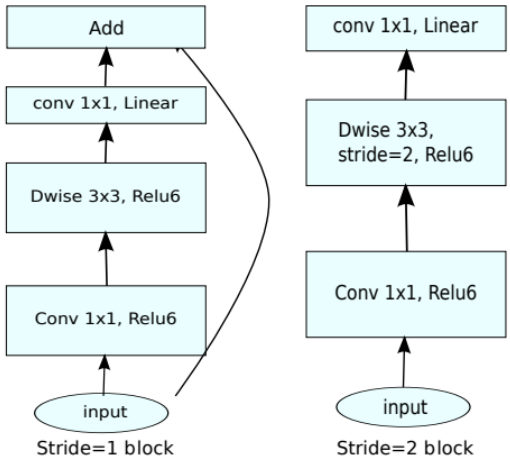

Fig. 1. MobileNetV2 architecture

On the Fig.1 shown that the MobileNetV2 block called the expansion convolution block consists of three layers. First comes the pointwise convolution with more channels, called the expansion layer. It is followed by depthwise convolution with ReLU6 activation. This layer, together with the previous one, essentially forms the already familiar building block of MobileNetV1. At the end there is a 1x1 convolution with a linear activation function that reduces the number of channels without losing useful information.

D. Model Architecture

To accomplish the task, several models of neural networks were considered, among which were various installations of Convolution 2D, RNN, LSTM and other approaches from previously conducted studies with selected datasets. From all tested models with basic settings, the highest accuracy was

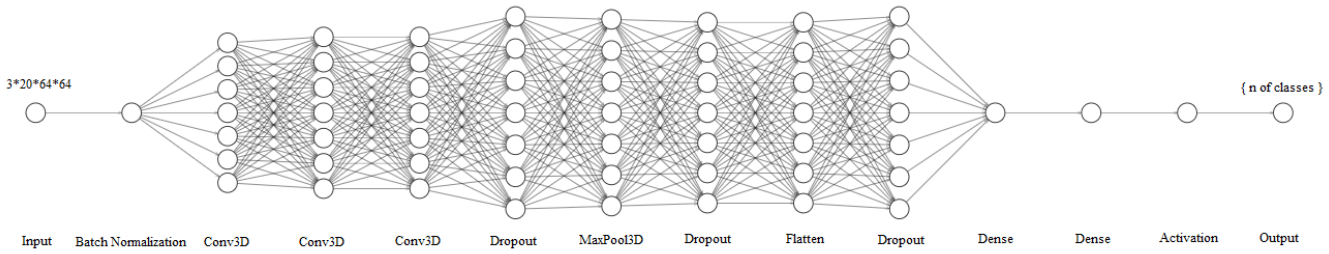

Fig. 2. Proposed CNN model

obtained using 3D CNN application. The baseline model for video action prediction purposes was constructed using 12-layers of Convolutional Neural Network reflected on Fig.2.

The architecture of the model starts with Batch Normalization at first layer for normalizing the input layer by adjusting and scaling the activations, followed by three 3D Convolution layers with "relu" activation function and with "channels first" data format. Next layers are Dropout with rate 0.1 and only one MaxPooling 3D with pool size = 2 due to small size of input data. Then the model adds another Dropout layer with 0.1 rate, one Flatten layer to flatten the input for next layer, and one more Dropout layer with rate = 0.1 to deal with overfitting issues. 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 a main metrics for the model.

V. Experimental Setup

A. KTH

For KTH dataset the final data contained 25 frames with skipping 2 frames in original data, which resulted in 3 seconds segment from the original video. After prepossessing and feature extraction the data was divided on the train and test with ratio 80% : 20%. For replication of the similar results the random state was set on 21. The best model obtained was trained with 20 epochs and with batch size equals to 32.

B. UCF Sports action

Due to the limited amount of the data for UCF datasets, two experimental setups were conducted. Initially 20 frames were stored as the final data and the number of frames skipped was defined for each class separately because of significant difference in the length of video for each class.

First setup uses the obtained final data, which was pre-processed and split to the train and test datasets with ratio 85% to 15%. For obtaining each class in the test dataset the random state of the splitting was set to 19. The best model was obtained by training the model with 30 epochs and by declaring batch size to 32.

Second setup used some additional data module by splitting existing final data to the 2 sub data. For each video frame stored, the even indexed frames and odd indexed frames were split and the final dataset in particular setup was increased twice in size. The train and test datasets size was declared as 80% to 20% on the 24 random state from the whole dataset. The model trained with 30 epoch and 32 batches outperformed other parameters for epochs, batches and splitting dataset parameter.

C. RusLan

The resulted data from the data preparation stage for RusLan dataset contained videos with minimum 13 frames up to 60 frames. The experimental setup set 10 frames for video with skipping frame number was internally calculated by looking on the number of frames per video. The obtained data was divided in 4:1 for train and test datasets, respectively, with 13 random state of splitting. Due to relatively small number of instances and number of frames comparing to other datasets, the batch size for training chosen as 15 and best model results was obtained on 30 epochs.

VI. Results

Four different setups was trained on the common general model described in Section IV-D. The best model setup for KTH dataset achieved 76.65% average accuracy on the test dataset. The Fig.3 represents the results of confusion matrix for KTH. Taking into account the factors the nature of actions in KTH are same, i.e boxing, handclapping, handwaving and running, jogging and walking, it was expected that model will misclassify the classes between each other. The most accurate predictions were on boxing, handclapping and walking classes with accuracies between 88%-89% and the least accurate classes: handwaving, running and jogging- correctly classified the test instances with 65%, 62% and 67% rates respectively. It can be also observed that boxing was misclassified as running and jogging due to the video instances when a subject performs boxing action while running in place.

Fig. 4 and Fig. 5 represents the result of the First and Second setups for UCF Sports action dataset respectively. First setup achieved an accuracy of 73.91% , which is close to the result for KTH dataset. As it can be seen from Fig. 4 6 classes provided 100% of accuracy, while Golf-Swing, Run-Side and Skateboarding-Front achieved only 40%, 33% and 67% of accuracy. The misclassification of data from these classes is

Fig. 3. Confusion Matrix (KTH dataset)

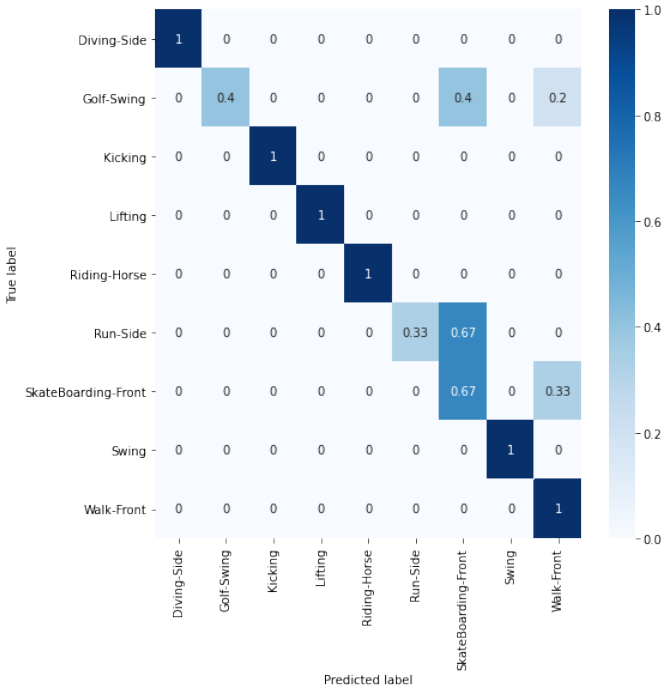

Fig. 4. Confusion Matrix (UCF dataset - First Setup)

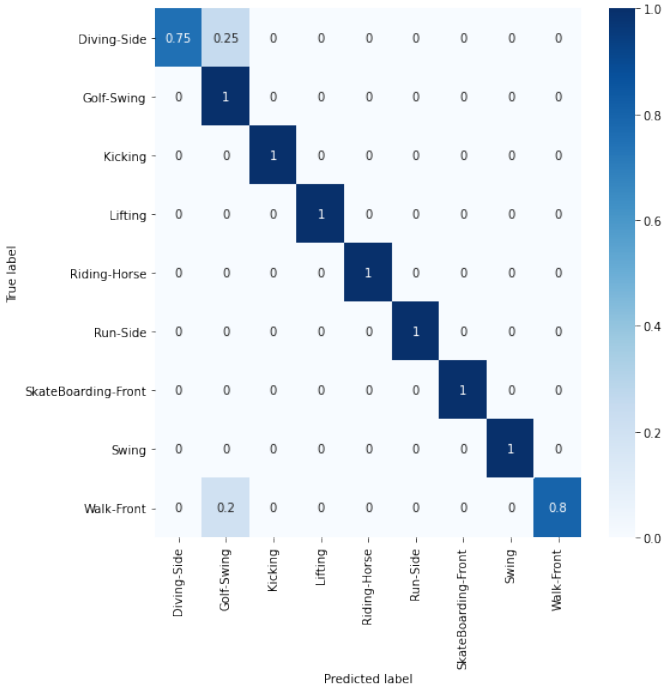

Fig. 5. Confusion Matrix (UCF dataset- Second Setup)

influenced by the presence of other people on the video and their poses in the frames, which the model also perceives as equivalent objects. In a Skateboarding class, people walk alongside the subject during the video, which can also be classified as Walk class. The same tendency is observed with the Golf and Skateboarding classes.

The Second Setup for the UCF Sports Dataset as it was expected outperformed the first Setup and achieved the 94.67%. In confusion matrix on the Fig.5, it can be observed that 7 classes except Diving-Side and Walking-Front achieved 100% of accuracy. However, it may be the result of adding additional video instances by splitting existing video to 2 pieces, which lead to the high accuracy on this dataset.

Fig. 6 illustrates the confusion matrix of the RusLan dataset which provides an average accuracy of 75% for testing set. The classes ”Cat”, ”Mom” and ”Woman” were predicted perfectly right. Words ”Dad” and ”Pipe” show an adequate accuracy of 80% as well. However, ”Bird”, ”Me” and ”Look” do not exceed 67%, and ”Look” falls down to 25%. The most probable reason for these confusions is the similarities in the sign language words. For example, the class ”Look” is often predicted as a ”Bird” because the gesture of showing bird’s flopping wings can be easily confused with a gesture of movement of pointing two fingers at one’s eyes and quickly pointing them out, as a sign of ”looking”. The word ”Me” must be confused with a word ”Look” for the same reason: a person quickly pointing at themselves looks almost the same as the sign ”Look”. Another problem might be the shortness and low quality of the videos. Low resolution and frame rate did not allow the neural network to capture more details of the signs, also the subjects sometimes were speaking very fast and the duration of one sign could be a split second, which affected the model’s performance. Additionally, we should keep in mind that every speaker has their own manner of the sign language,

so a word of one person could look different from what another person is showing.

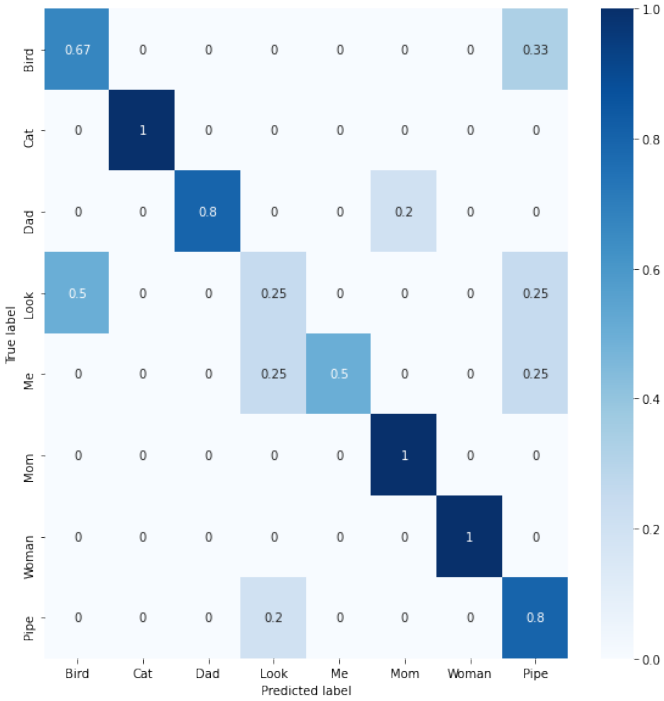

Fig. 6. Confusion Matrix (RusLan dataset)

The Table I describes the results obtained on three datasets with First Setup of UCF Sports dataset and provides the results obtained by other researchers. Overall, the general model performed less accurate than existed studies. However, other studies used as classifiers LSTM, SVM and KNN models and added some additional feature such as transfer learning and spatio-temporal feature extraction. Transfer learning is not recommended to use for small datasets and spatio-temporal feature extraction requiers some additional data preprocessing for each dataset explicitly. This study focused on the providing general CNN model classification for different type of video datasets.

Previous approaches	KTH	UCF Sports Action	RusLan
Proposed architecture	76.67%	73.91%	75%
Ravanbakhsh et al. [7]	74.5%	88.1%	-
Charalampous, K. and A. Gasteratos [3]	91.99%	88.55%	-
Shamsipour, Ghazal and Pirasteh, Saied [10]	-	93.67%	-
Basha et al. [1]	95.27%	-	-
Sargano et al. [9]	98.15%	91.47%	-

TABLE I
COMPARISON OF RESULTS WITH EXISTED STUDIES

In addition, there are some studies which worked with other Russian Sign Language datasets. The accuracy of most of the publications weren't mentioned and authors mentioned that their models didn't achieve good result. The minority of studies claimed that they achieved 70%-80% accuracy. However, their result is not comparable to this study, because the researchers were working with static data, i.e. image classification, while the subject of this project is video classification.

VII. Future Works

Since the RusLan dataset has not participated in any study previously, it is recommended to extend the data preparation stage by increasing number of samples of existing classes and increase number of classes. Moreover, it is suggested to increase the duration of each video instance and add some additional data of Russian Sign language video instances from open sources.

Taking into account the existence of studies for Latin Sign Language (LSA 64) and American Sign Language (ASL), it is proposed to extend the current study by replicating the models of the other sign language models and validate the performance of the existing model on the RusLan dataset.

In addition, the current study can be extended by adding more video datasets with different content. The future work will provide the common general model for different type of video data and can be improved for each type of data exclusively. The live or instant recognition can be used by adding to the existing model Hidden Markov Model by predicting the possible move of the subject.

VIII. Conclusion

Video based classification is one of the most popular topic in the Machine Learning field. Object, action and speech recognition is used in different spheres including medicine, engineering and others. This study proposed the general model for video based classification using Convolutional Neural Networks. This paper analyzed the performance of the well-known action datasets KTH and UCF Sports and compared it with existing studies with other researchers. The accuracy of proposed general model is lower than of existed papers; however, it is one of the minority papers that used CNN as classifiers for general model for both datasets.

Moreover, the specific setup of UCF Sports achieved almost 95% of accuracy and new Russian Sign Language dataset, which was derived from RusLan Corpus, was used for validation of the performance of the general model. The accuracy of 75% can be used as the benchmark for the future works. The possible future research was discussed for extending the current work. As a limitation, it should be mentioned the limited power capacity of the Google Colab and the time costs of each model's training and testing.

References

[1] Basha, Sh Shabbeer Pulabaigari, Viswanath Mukherjee, Snehasis. (2020). An Information-rich Sampling Technique over Spatio-Temporal CNN for Classification of Human Actions in Videos.

[2] 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.

[3] Charalampous, K. and A. Gasteratos, On-line deep learning method for action recognition. Pattern Analysis and Applications, 2016.19(2): p. 337-354.

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

[5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 4510-4520, doi: 10.1109/CVPR.2018.00474.

[6] 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.

[7] Ravanbakhsh, M., Mousavi, H., Rastegari, M., Murino, V., Davis, L. (2015). Action Recognition with Image Based CNN Features. ArXiv, abs/1512.03980.

[8] Russian Sign Language Corpus. Available at: http://rsl.nstu.ru/site/index.

[9] 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.

[10] Shamsipour, Ghazal and Pirasteh, Saied. (2019). Artificial Intelligence and Convolutional Neural Network for Recognition of Human Interaction by Video from Drone. 10.20944/preprints201908.0289.v1.

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

[12] 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.