Human Action Recognition for Pose-based Attention: Methods on the Framework of Image Processing and Deep Learning

Desislava Nikolova[1], Ivaylo Vladimirov[2] and Zornitsa Terneva[3]

Abstract – **This paper presents an overview of some approaches of Human action recognition (HAR) for pose-based attention. The paper focus is on algorithms that use video processing on a given dataset. A list of the best HAR datasets is given in order to show the variety of the available videos online. Local and Global feature extraction are reviewed. Also some of the most common Deep Learning methods are studied: Recurrent Neural Network (RNN), Convolutional Neural Network (CNN) and Generative Adversarial Network (GAN). All of the methods are directed to recognise the pose and the focus of the person in a recording.**

Keywords – **Human Action Recognition, Pose-based Attention, Image Processing, Feature Extraction, Deep Learning**

I. Introduction

The human recognition field presents increasingly harder cases to solve. In contrast with the past, one of the biggest challenges on the topic is the next level – human action recognition. The main idea is to have a very precise model, describing the exact action a given human is doing. This can be used in many fields: medicine, education, robotics, etc.

The Human Action Recognition (HAR) plays an important role in the interaction between people and interpersonal relationships. It provides information about a person's identity, personality and psychological state so it is difficult to be extracted and analyzed.

The ability to recognize the attention span of a person on a video or image is even more challenging. However this information is crucial in situations where the focus of the subject is more important than the actual activity he is performing. In schools, at work, even at the marketing agencies, the difference between a focused and unfocused person is determining the benefits of a certain activity.

In this paper a few of the best approaches in the field of HAR are presented. They are using RGB image preprocessing and Deep Learning for the actual recognition part. The main focus is only the methods which can help categorizing the attention of a person, based on the pose.

[1]Desislava Nikolova is with the Faculty of Telecommunications at Technical University of Sofia, 8 Kl. Ohridski Blvd, Sofia 1000, Bulgaria. E-mail: dnikolova@tu-sofia.bg

[2]Ivaylo Vladimirov is with the Faculty of Telecommunications at Technical University of Sofia, 8 Kl. Ohridski Blvd, Sofia 1000, Bulgaria, E-mail: ivladimirov@tu-sofia.bg

[3]Zornitsa Terneva is with the Faculty of Telecommunications at Technical University of Sofia, 8 Kl. Ohridski Blvd, Sofia 1000, Bulgaria. E-mail: zterneva@tu-sofia.bg

978-1-6654-2887-3/21/$31.00 ©2021 European Union

In the following sections the processing methodology is presented, including the three main points: preparation, feature extraction and classification.

II. Preparation

Different human actions could be analyzed through different data modalities, like RGB, skeletons, infrared, video, streaming, audio, radar, acceleration etc. Each and every one encodes different sources of useful and distinct information and has unique advantages based on the various application scenarios. [1]

In this paper we are going to focus on all of the features that can be extracted from an image (see Figure.1).

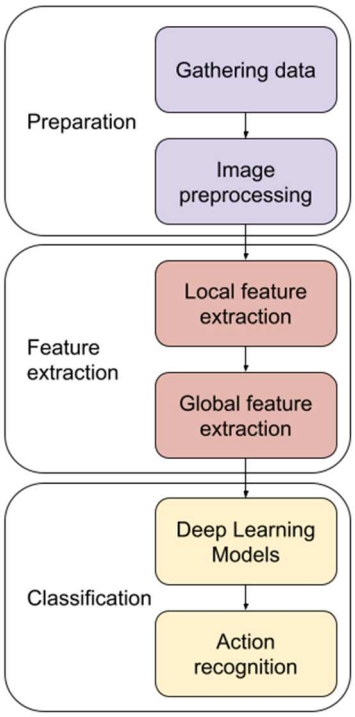

Fig. 1. Block scheme of HAR

A. Gathering data

There is a variety of datasets all over the Internet with a big variety of poses in the videos. The pose-based attention span is best analyzed in clips, because it gives the idea of how much movement is involved. An important feature in a given

dataset is the number of samples it provides. Some of the largest datasets are:

- YouTube8M [2]

YouTube8M was created in 2016 and the largest-scale video dataset. It contains 8 million YouTube videos – about 500 thousand hours of video content in total. The dataset consists of 3, 862 action classes and each video is annotated with one or multiple labels by a YouTube video annotation system. YouTube8M is split into training - 70%, validation - 20% and test - 10% sets. The validation set has additional human-verified segment annotations in order to provide one more layer security in the results.

- HACS [3]

HACS was first released in 2019 as a new large-scale dataset for recognizing and locating human activities in Web videos. There are two types of manual annotations. On 504 thousand files, HACS Clips has 1.55 million 2-second clip annotations. HACS Segments has 140 thousand full action segments (from start to finish) on 50 thousand videos. The videos are annotated with 200 human action classes.

- Sports1M [4]

Sports1M was first launched in 2014 as the world's first large-scale video activity dataset, with over 1 million YouTube videos annotated with 487 sports groups. Since the types are fine-grained, there are few differences between classes. For assessment, it uses a 10-fold cross-validation break.

- Moments in Time [5]

Moments in Time is a large-scale dataset optimized for event awareness that was introduced in 2018. It includes one million three-second video clips that are annotated with a 339-class dictionary. Unlike other datasets aimed at analyzing human behavior, the Moments in Time dataset includes individuals, animals, objects, and natural phenomena. By raising the number of videos to 1.02 million, pruning ambiguous groups, and increasing the number of labels per clip, the dataset was expanded to Multi-Moments in Time (M-MiT) in 2019.

- Kinetics Family [6,7,8]

The whole Kinetics Family is one of the most commonly used benchmarks. Kinetics400 is a series of roughly 240k instruction and validation videos trimmed to 10 seconds from 400 human activity groups that was launched in 2017. Kinetics Family continues to grow with Kinetics-600 which was released in 2018 with 480K videos. And the last Kinetics700 was released in 2019 with 650K videos.

- UCF101 [9]

UCF101 is a dataset of practical action videos compiled from YouTube with 101 action groups. This data collection is a supplement to the UCF50 data set, which contains 50 action groups.UCF101 has the most variety in terms of behavior, with 13320 videos from 101 action categories, and it is the most difficult data collection to date, with wide differences in camera motion, object appearance and posture, object size, angle, cluttered landscape, lighting conditions, and so on.

Since most action recognition data sets are unrealistic and staged by actors, UCF101 seeks to promote further action recognition studies by learning and testing new realistic action types.

As every dataset has its problems and challenges ahead, the next step is to process the videos in the most accurate way.

B. Image pre-processing

In this segment, image pre-processing options are recommended. They are driven by the feature descriptor method selected. As previously mentioned, raw image data directly from a given dataset can have a number of issues, and hence is unlikely to provide the best computer vision performance.

Image pre-processing is done by following the vision pipelines of the following fundamental families of feature description methods:

- Polygon Shape Descriptors (blob object area, perimeter, centroid)
- Local Binary Descriptors (LBP, ORB, etc.)
- Spectra Descriptors (SIFT, SURF, etc.)
- Basis Space Descriptors (FFT, wavelets, etc.)

Image pre-processing has to be carefully considered. A local binary descriptor using gray scale data, for example, would need different pre-processing than a color SIFT algorithm [10].

The processing part depends a lot on the dataset, but also depends on the features that are expected to be extracted.

III. FEATURE EXTRACTION

A. Vector generation of features

The HOG features descriptor can be used for both RGB and depth video sequences to reflect the appearance and motion information from RGBD video behavior. These feature vector values can be combined to produce the BoFs.

This descriptor is often used in both person identification and behavior recognition. It can be extended around each motion importance keypoint in video frames of RGBD videos images.

The HOG features descriptor can be well adapted to reflect local shape description from image channel and local motion information from modem host interface by computing the distributions of local acclivity and declivity.

B. Generation of key points

Applying the Speeded Up Robust Features (SURF) detector to remove more visually characteristic keypoints from the spatial domain, motion interest points can be located from RGB frame sequences.

The keypoints are then filtered using a temporal template method to detect motion and compute its direction. This restriction is provided by motion history images (MHI), which are created by calculating the difference between two adjacent frames.

The moving object with the most recent motion is defined by the points with higher intensities in MHI. The optical flow for certain keys retained after MHI filtering is then computed

using the Lucas-Kanade process. Pixel strength in an MHI is a feature of the temporal background of motion at that point.

The MHI is shown in equation (1):

$$H_\tau(x,y,t) = \begin{cases} \tau \\ \max(0, H_\tau(x,y,t-1) - 1) \end{cases} \quad (1)$$

The result is τ if D(x,y,t)=1.

Where:

- D(x, y, t) is a binary image of differences between frames.
- τ is the maximum duration of motion. It decides the temporal extent of the movement.

Following the computation of motion keypoints P(x, y, t) from RGB images, these motion points are aligned from RGB to the corresponding depth images Pd x, y, z, t, where x, y, t represent the coordinates and time of interest point P on RGB images and x, y, z, t correspond to the 3D coordinate and time of interest point on depth images [11].

C. Local feature extraction

Local representation systems derive characteristics from a geographic location. Local representation should not rely on people identification or human segmentation for action recognition. Partial occlusion, background clutter, perspective differences, individual appearances, and likely dimensions and rotations had little effect on local representation. [12].

D. Global feature extraction

Global features are extracted directly from original sources. The images are processed as a whole, so that all pixels are used to create the descriptor. One specific step detects the human and subtracts the background. This step generates a Region of Interest (ROIs) mainly used in background removal or human tracking. These ROIs include information for human appearance, geometric structures, motion, etc. They are often sensitive to variation in the viewpoint, noise and occlusion. [13]

IV. CLASSIFICATION

A. Deep Learning Models

With the increasing development and impressive performance in the recent years, the Deep Learning methods are in the center of the existing computer vision tasks, just like object detection, image recognition, etc. Three of the most used methods are Recurrent Neural Network (RNN) [14], Convolutional Neural Network (CNN) [15] and Generative Adversarial Networks (GANs) [16].

- RNN

Normally RNN-based methods are suitable for processing time series data because of its unique structure.

Some RNN-based methods like Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) have been adapted to skeleton based action recognition. They are used for improvement of the current learning techniques [17].

RNNs are commonly used to process sequential data, such as voice, text, images, and time-series, where the data at any given position is dependent on previously used data.

The model extracts the data from the current time Xi and the hidden state from the previous phase hi1 at each time-stamp and outputs a target value and a new hidden state (see Figure 2).

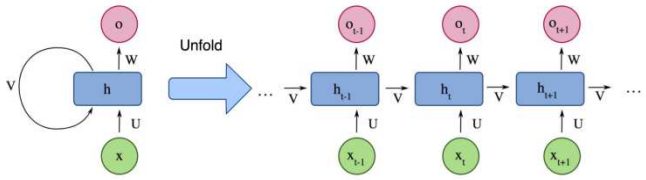

Fig. 2. Block Scheme of RNN

RNNs are troublesome with long sequences because they cannot catch long-term dependencies in some real-life implementations and often suffer from gradient vanishing or bursting problems.

Long Short Term Memory (LSTM) RNNs are designed to eliminate these problems. The LSTM architecture consists of three gates: input gate, output gate, and forget gate. They control the flow of information in and out of a memory cell which stores values over random time intervals [18].

- CNN

CNNs are among the most popular and commonly used deep learning architectures, especially for computer vision tasks (see Figure 3).

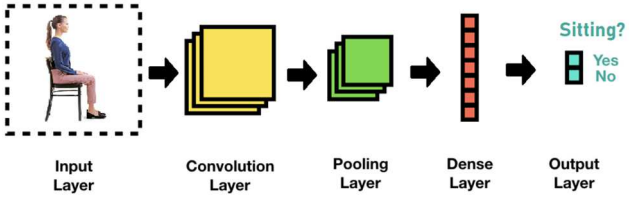

Fig. 3. Block Scheme of CNN

CNNs consist of three layers:

- Convolution layers

Those are the layers in which a kernel of weights is complicated in order to extract specific features.

- Pooling layers,

The pooling layers degrade spatial resolution by replacing a small region of a feature map with statistical data.

- Dense layers

In these layers a linear operation is performed, in which every input is connected to every output by a weight.

The units in layers are locally related, which means that each unit receives inputs from a small neighborhood of units in the previous layer. It is known as the receptive field. By stacking layers to construct multi-resolution pyramids, higher-level layers learn information from ever bigger receptive fields. CNNs offer a significant statistical advantage over fully-connected neural networks in that all of the receptive fields in a layer share weights, resulting in a considerably reduced number of parameters.

- Generative Adversarial Networks (GANs)

A generative adversarial network (GAN) is a newer class of machine learning system developed by Ian Goodfellow and his colleagues in 2014.

This method learns to produce new data with the same statistics as the training set given a training set. A GAN educated on photographs, for example, will produce new photographs that seem at least superficially genuine to human observers, with several practical characteristics. GANs were initially introduced as a kind of generative model for unsupervised learning, but they have also proved useful for semi-supervised learning, fully supervised learning, and reinforcement learning.

The core idea of a GAN is perfect for the human action recognition and only the final results can show whether it is the best option for the current problem.

B. Action recognition

The final results are expected to give a percent probability and reflect on the real action of the human. It can all depend on the model but with the proper preprocessing, the final answer should be close to 100% (see Figure.4).

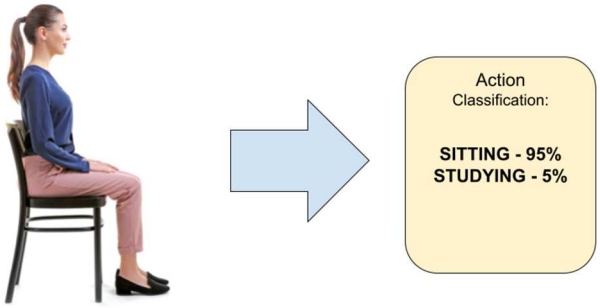

Fig. 4. Example of action recognition

In the given example the chances for the person to be focused are 100% as sitting and studying are both states of high attention span.

V. CONCLUSION

In general, one of the best ways to identify the attention of the user is through the pose. In order to better understand the pose and the activity of the person, the HAR is needed. In this paper are presented some of the best methods to tackle this problem.

The next step in this research will be to focus on the human face and detect even the expressions of the user. The Clustering approach, based on Genetic Algorithms [19], is one which should be examined because it is amongst the widely used techniques in the study of unsupervised data in activities of everyday life [20].

The future of the field of Human Action Recognition is not clear but can diverge in different directions. The pose-based attention will be a key factor for many areas of our lives.

ACKNOWLEDGEMENT

This work is supported by the National Program "Young Scientists and Postdoctoral Students", PMC №577, 2018-2021, Ministry of Education and Science, Bulgaria.

REFERENCES

[1] Z.Sun, J.Lui, Q.Ke, H.Rahmani, M.Bennamoun, G.Wang, „Human Action Recognition from Various Data Modalities: A Review", 2021;
[2] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B.Varadarajan. "YouTube-8M: A LargeScale Video Classification Benchmark", 2016;
[3] H. Zhao, Z. Yan, L. Torresani, "HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization", 2019;
[4] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, "Large-Scale Video Classification with Convolutional Neural Networks". In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014;
[5] M. Monfort, A. Andonian, B. Zhou, K. Makrishnan, S. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfruend, et al. "Moments in Time Dataset: One Million Videos for Event Understand"in IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2019;
[6] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The Kinetics Human Action Video Dataset", 2017;
[7] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman. "A short note about kinetics600", 2018;
[8] J. Carreira, E. Noland, C. Hillier, and A. Zisserman. "A short note on the kinetics-700 human action dataset.", 2019;
[9] K. Soomro, A. R. Zamir and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild", 2012;
[10] J. Verne, "Perspectives on Image Processing."
[11] R. Al-Akam and D. Paulus, "RGBD Human Action Recognition using Multi-Features Combination and K-Nearest Neighbors Classification", 2017
[12] R. Poppe, "A survey on vision-based human action recognition. Image Vis Comput.", 2010;
[13] I. Jegham, A. Ben Khalifa, I. Alouani, M. Ali Mahjoub, "Vision-based human action recognition: An overview and real world challenges", 2020;
[14] G. Lev, G. Sadeh, B. Klein, and L. Wolf, "RNN Fisher Vectors for Action Recognition and Image Annotation", 2016;
[15] G. Cheron, I. Laptev, and C. Schmid, "P-cnn: Pose-based cnn features ´ for action recognition," in IEEE International Conference on Computer Vision, 2016.
[16] J. Wang, Y. Chen, Y. Gu, Y. Xiao and H. Pan, "SensoryGANs: An Effective Generative Adversarial Framework for Sensor-based Human Activity Recognition," 2018;
[17] S. Pienaar, R. Malekian "Human Activity Recognition Using LSTM-RNN Deep Neural Network Architecture", 2019
[18] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey", 2020;
[19] A. Amelio and C. Pizzuti, "A New Evolutionary-Based Clustering Framework for Image Databases", 2014;
[20] P. Colpas, E. Vicario, E. De-La-Hoz-Franco etc., "Unsupervised Human Activity Recognition Using the Clustering Approach: A Review", 2020;