PROJECT REPORT

ON

3D CONVOLUTIONAL NEURAL NETWORKS FOR HUMAN ACTION RECOGNITION

(IEEE Citation for Paper: S. Ji, W. Xu, M. Yang and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221-231, Jan. 2013, doi: 10.1109/TPAMI.2012.59.)

ECE 535 (Jan 2023 - April 2023)

Submitted by:

Opara, Anita (V00996592)

Sekhon, Anuinder (V01022326)

Submitted to:

Prof Michael McGuire

INTRODUCTION

The chosen paper addresses the problem of human action recognition in surveillance videos using real-world data from airport surveillance as context. Human action recognition faces a lot of challenges as it is highly influenced by subjective factors like appearance, lighting ,pose , background and clothing. The existing system (at the time of paper) requires a lot of assumptions, which poses a big challenge in the real world as it is not practically feasible to know all the features in advance and ensure that those assumptions will be well suited in every environment in which that problem is being addressed. This paper proposes a deep learning model based on convolutional neural networks that is independent of the assumptions about the surrounding environment in which the video was taken as CNNs are invariant to such things. The paper proposes the use of 3D CNNs as they allow the features from both spatial and temporal dimensions.

In real world scenarios, it becomes necessary to analyze the objects in motion in surveillance videos for multiple reasons like security purposes, detecting panic situations in crowds like in airports or shopping centers, shopping behavioral analysis for personalized recommendations and also in the medical industry for the assistive care. This algorithm can be applied in such engineering tasks which directly and indirectly benefit the society.

The computing resources required for using the deep learning models for human action recognition will be intensive as this area finds major application in scenarios that require both real time decisions as well as research on the collected data. Generally, this will require high computation power GPUs,a cloud storage to process the dataset and hardware requirements for the specific softwares needed.

We believe that the research requires legal permissions of the geographical area to be implemented in the real world as it requires accessing the surveillance videos. Along with this, it requires taking into consideration the privacy of individuals as humans that appear in the video dataset have to be informed that their data is collected and the cultural beliefs of that place. The project was a part of an event surveillance system but didn’t exactly state for what the surveillance it will be used for. Assuming it was used for classifying panic situations used to automatically alert security agencies,misclassifications will cause waste of useful resources and personnel due to false alarm. Along with it, if the system makes any misclassification of the person during a sensitive situation like religious gathering, this can cause legal trouble for the authorities.

DATA COLLECTION AND PREPROCESSING

This paper uses video datasets from two different sources as mentioned below:

1. TRECVID dataset: The paper uses TRECVID development 2008 data set for the training and evaluation purpose. This video dataset has nine action classes and is generated from the real world London Aiport's surveillance video. The TRECVID, has clearly specified on their website that the data available is for research and prototyping purposes and clearly prohibits the use in commercial products or using the evaluation results for the production systems. [1]
2. KTH dataset: It contains six action classes performed by 25 subjects in four different outdoor and indoor environmental and situational conditions. The action classes are: walking, jogging, running, boxing, hand waving and hand clapping. This dataset has been used for the test and evaluation purposes too. [2]

From the TRECVID 2008 development dataset, the paper considered three action classes, namely, cellToEar, ObjectPut, Pointing. To explain, the cellToEar is an action when a cell phone reaches the head, ObjectPut based on having or dropping the object and Pointing is a gesture of pointing towards something or someone.

Since, here it is required to categorize the data into three class labels, the paper used the one-vs-all multi-class classification technique. This technique involves splitting the dataset to multiple binary classification problems and then training the binary classifier for each classification problem and hence training the model. Although this strategy makes it tough for handling the large datasets, we believe that the author chose this as it was a vast dataset and the author wanted to ensure that there were enough proper negative samples generated for the action that were not part of all those three action classes. The screenshot of the statistics for the dataset from the paper [3] is shown below:

Date\Class	CellToEar	ObjectPut	Pointing	Negative	Total
20071101	2692	1349	7845	20056	31942
20071106	1820	3075	8533	22095	35523
20071107	465	3621	8708	19604	32398
20071108	4162	3582	11561	35898	55203
20071112	4859	5728	18480	51428	80495
Total	13998	17355	55127	149081	235561

Figure 1. The statistics on the samples on TRECVID 2008 Development Dataset [3]

Based on our understanding, the author chose the TRECVID 2008, dataset as it is generated from a crowded place where along with diversity of people in terms of age and height, at the airport a person is required to go through multiple counters and hence varied gestures with no notable difference in the human gestures happening in one

video frame. Also, based on our further research, this dataset is highly imbalanced and cellToEar is the least frequent action in the dataset, followed by ObjectPut to Pointing in the increasing order of the frequency. Also,the data from camera 4 was not used because it didn't capture a lot of actions. Below shown screenshot shows the density of events in the TRECVID dataset.

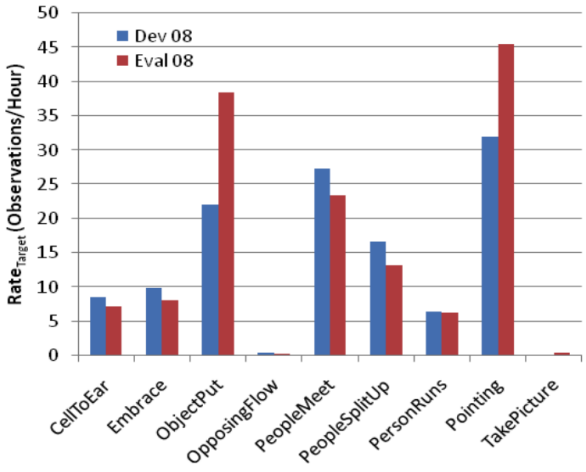

Figure 2. Density of events in TRECVID dataset [4]

The model was also evaluated on the KTH dataset, splitting the dataset into training and testing by taking 25 random subjects in the training set and the remaining 9 for testing.

After the comprehensive study of the paper and the resources available for the dataset on the internet[1][4], we believe that there is not enough information available to access the data preprocessing technique and recommend improvements on it.

FEATURE SELECTION

From the TRECVID dataset, a human detector was used to identify humans [6] and a motion tracker was used to determine if they were carrying out an action.The resulting output from the motion tracker determined the head tracking box which contained majorly the torso of a human carrying out an action.A bounding box was created from cropping the head tracking box to a frame of 60* 40 containing majorly the head of the human.Since the temporal dimension of the 3D CNN is 7 ,6 other frames were derived from the same position of this frame. Along with this, 3 frames were derived from the same position in the previous time-stamp and 3 frames were derived from the subsequent time stamps of the step size 2 in the same position.This forms the input cube for a 3DCNN representing a particular action. According to the paper [5] ,using 7 input frames which is about 0.3 seconds of video,the results are similar to passing the entire video sequence required for an action and thereby making 7 input frames as

optimal.This paper doesn't clearly specify if any special steps were done to ensure that the features were invariant to typical transformations of the relevant dataset.

Typically, the size of inputs to the 3D CNN system is small as convolution greatly increases the number of features to be extracted relative to the input frames which increases complexity .Also ,not all actions can be captured in small amount of frames.The paper highlights that for actions requiring a large number of picture frames the following can be done: a section of relevant picture frames of the action are used as input, features are extracted from the larger set of frames ,these features are combined with the existing state at the last layer of the CNN .This way the CNN has learnt more about the action.

Since, the paper doesn't mention any specific reason for which it chooses only those three action classes. So, we believe that the paper should have used the top three action classes with highest representation in the TRECVID dataset. Additionally, from the perspective of security, we believe that it is important to consider the action classes based on PeopleMeet and PeopleSplitUp which are the available actions with higher representation in the dataset.

MACHINE LEARNING/ PATTERN CLASSIFICATION ALGORITHM

The algorithm used in this paper is the 3D convolutional neural networks (3D- CNN) is able to take 3D data as an input , learn information from it ,create a feature vector containing information learnt from it.The feature vectors can be used to train a classifier or test an existing classifier.

3D CNN is a variant of the prior 2D CNN. 3D CNN consists of alternate 3D convolution and sub sampling layers which are used to create relatively large feature maps from 3D input such as videos and medical images.An advantage of 3D CNN is it is able to extract spatial and temporal information from feature maps generated as it also uses 3D kernels during convolution. After the alternate layers of the convolution and subsampling, the resultant feature vector can be passed down to a classifier. 3D CNN also supports regularization to prevent overfitting to noise and the use of auxiliary outputs in the intermediate layers as against the initial input to prevent the feature maps to be too large which is responsible for increasing the computational complexity.

We agree on the authors selection of the algorithm because of the type of task it is to be used for.Firstly, the input of the model is 7 contiguous picture frames as against just a picture frame in the 2D -CNN case.This extracts information as a collective and is able to identify correlation across the frames .This is more effective in action recognition as an action spans multiple images and cannot be easily identified from a 2D image. Secondly,the paper highlights the first layer of the 3D CNN model using a set of hard

wired kernels which contains prior knowledge on features.This kernel was used to generate 5 distinct channels from the input data.This leads to having better performance instead of using a set of random kernels.

In this paper, the 3D-CNN architecture comprises 6 alternate layers namely 3D convolutional layers, 2 sub sampling layers , 1 hardwired layer.The 3D-CNN takes an input of 7 contiguous picture frame of size 60 *40.The first layer uses the hardwired kernel and generates 33 features maps of 5 channels which are gray ,gradient-x ,gradient - y ,optflow-x and opt-flow-y .The gray channel holds the gray values from the image.The gradient- x and gradient- y channel holds values of the gradient in the x and y directions respectively.The optflow-x and opt-flow-y holds optical values in the x and y directions respectively. In the second layer,2 different 3D convolution kernels of size (7*7*3) we applied to the output of the first layer resulting in 2 set of feature maps of size (23*54*34).The next stage involved a a 2* 2 subsampling applied to each of the feature maps from the previous stage resulting in the same number of feature maps (2 * 23 = 46 feature maps) but with reduced dimension (27*17). The fourth stage is the second convolution layer.In this stage, 3 different 3D kernels of size 7*6*3 are applied to the 2 sets of existing feature maps resulting in 6 sets of feature maps of size (13*21*12) .This stage results in a total of 78 feature maps.Stage 5 is the second subsampling layer and a 3*3 subsampling is carried out reducing the dimension from 21*12 to 7*4. At this stage the temporal dimensions for all the 5 channels are small - 3 for gray , gradient-x and gradient- y channels and 2 for the optiflow-x and optiflow-y channels. The sixth and final stage is the third convolution layer and a 2D kernel of size 7*4 is applied to the output of the fifth layer resulting in 128 feature maps of size 1*1.Each of these feature maps are connected to the 78 feature maps from the previous stage.

The snapshot from the paper [3], which explains the above process is shown below:

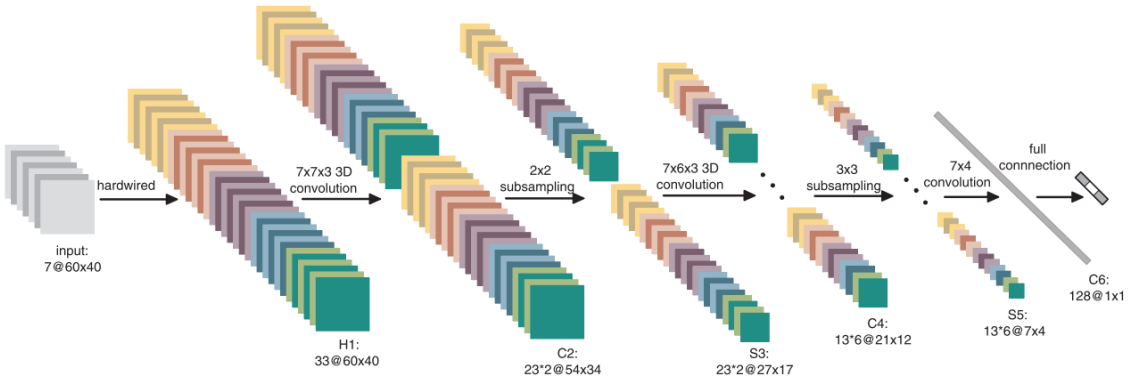

Figure 3: The 3D CNN architecture[3]

The model parameters from the second layer were randomly initialized and trained using the Stochastic Diagonal Levenberg- Marquardt method. It is an optimization algorithm over the space of the parameters and uses a hybrid of Gauss–Newton and steepest descent approach. As per our understanding, this method is useful in large imbalanced datasets as it takes into consideration additional information like damping factor while working on the optimization problem.A learning rate is generated for each parameter.

3D CNN action recognition systems are not typically stand alone as they are used as a module within a system.This paper used a dataset obtained from a surveillance event detection system.3D-CNN is computation intensive and is likely to be carried out using distributed systems.If the results of the algorithm are also required in real-time then an additional complexity has to be kept in mind.Also ,depending on the size and sensitivity of the data,cloud or on-prem solutions may suffice. Considering the No free lunch theorem, we can state that it is not possible to have one given architecture that will perform equally good for all the problems, even if the new problem is also based on human detection.

PERFORMANCE EVALUATION

The 3D CNN architecture was evaluated on 2 datasets.Majorly the TRECVID dataset and then the KTH dataset that we introduced in the beginning.On the TRECVID dataset ,various 3D CNN architecture combinations and 4 other methods were used for comparison.The KTH dataset was used to compare the proposed 3D-CNN architecture with the existing HMAX method.

On the TRECVID dataset, for testing the 3D CNN cross validation was used.Since the dataset was captured in 5 days, a five fold cross validation where every single day represented a fold.Hence, the testing and evaluation set were the same.While in the KTH dataset, 9 of the 25 samples was used for evaluation. Additionally, as stated in the paper, the performance measures that were used for the multiple values of the false positive rates, are as follow:

1. Precision
2. Recall
3. AUC (Area under the ROC Curve)

Only in the KTH dataset , the evaluation set was different from the test set as 16 of the 25 inputs were randomly selected for training while the remaining 9 were used for testing. Majority-voting is the criteria used to determine what a given input sequence was to be predicted. The action with the most frames is selected as the predicted action. .In the Kth dataset , the paper highlights 6 action classes.

Other methods and 3D-CNN combinations were used to get features from the TRECVID dataset .The features were evaluated through a classifier and their performances were compared to that of the 3D CNN architecture proposed in the paper.

Firstly ,we discuss the Spatial pyramid matching Method (SPM).SPM was used to extract specific types of features from the TRECVID dataset. Appearance information was extracted and stored as SPM-Gray,Shape and motion patterns were extracted and stored as SPM-MEHI while spatiotemporal information (TISR) were extracted and stored as SPM- TISR. Each of these were of size 8192D.A concatenation of SPM-GRAY and SPM-MEHI features were used for evaluation.One-vs-all linear SVM technique was used to train for 3 classifiers of the three action classes.The paper did not highlight what percentage of the dataset was used for testing and training with respect to the SPM approach.The evaluation section in the paper, mentioned 4 methods to be evaluated SPM-MEHI,SPM- GRAY,SPM-TISR and a combination of GRAY and MEHI SPM- GRAY + MEHI.

Additionally, to be evaluated was the 2D CNN and a regularized 3D RCNN which used auxiliary output from the SPM -GRAY and SPM -MEHI features.The SPM -GRAY and SPM -MEHI features were reduced using PCA to 150 D each.The paper compared 3D CNN performance to the performance of other similar methods using the same dataset.

While evaluating the performance of all the methods introduced for the 3 action classes, it was observed that the 3D -CNN performed best at recognizing CellToEar action,the combined SPM GRAY and SPM -MEHI performed best at recognizing pointing action, the regularized CNN - 3D RCNN performed best at the ObjectPut action and performed best on the averaged performance.

Although for the KTH dataset,the HMAX method slightly outperformed the 3D-CNN architecture, the paper highlighted the reason that this is due to hardcoded features computed from better resolution pictures in HMAX in comparison to the random features generated in the CNN algorithm.

The authors discuss exploring unsupervised 3D CNN training in the future as it requires a smaller number of labeled samples for training.The paper stated most of the misclassifications are not even recognized by humans, so at some point we can think of not considering them but at the same point we need to keep in mind that even slight misclassifications can have legal implications.

The paper also evaluates various combinations of the CNN architecture and compares their performance against each other using the TRECVID dataset.Here,a mixed CNN -M refers to a CNN where the channels are convolved separately but are connected to the same feature maps in the first layer of the CNN while a subscript CNN- S refers to a

CNN where the channels are convolved separately but are connected to the different feature maps in all the layers of the CNN.The number associated with each CNN signified the number and type of convolutions.For instance, a 3D CNN-M- 323 means that the CNN is mixed consisting of 3 convolution layers where the first and third layers are the 3D convolutions while the second one is a 2D convolution.

The performance of the following CNN architectures was evaluated:

- 3D-RCNN-S-332,
- 3D-CNN-S-332,
- 3D-CNN-M-332,
- 3D-CNN-M-322
- 3D-CNN-M-222.

The regularized 3D-RCNN-S-332 showed the best performance.

The paper tested and showed that although 3D-RCNN-S-332 gave the best outcome individually,better performance can be obtained by combining all the CNN algorithms hereby evaluated.This is done by combining incrementally the algorithms from the highest individual performance to the least performance.

Since the dataset is sufficiently large and also imbalanced, we believe that the cross-validation was not the best of the all available methods for the purpose of evaluation.

CONCLUSION

In considering continuing this project, we might reconsider the choice of the action classes of the author. In our opinion, we will choose to consider the classes which either have high representation in the dataset or will consider the ones which represent most frequent human actions in day to day life. Additionally, we will reconsider the validation scheme for our project. Since the training dataset was taken from a specific airport ,we believe that the best way to test true performance should be done by testing on an entirely new dataset and not using the method of cross validation. Since the data is only from a particular airport, we cannot ensure the performance of the model on the video datasets which are taken from places other than the airport.

We believe that using surveillance videos for the purpose of human action recognition is a sensitive topic which can create legal troubles even in the case of slight misclassification or violation of the rules regarding privacy. Along with this, according to us this needs to consider the validation techniques keeping in mind the environmental and situational parameters.

REFERENCES

[1] TREC Video Retrieval Evaluation Home Page URL:https://trecvid.nist.gov/ Date accessed: April 19, 2023

[2] Recognition of human actions URL:https://www.csc.kth.se/cvap/actions/ Date accessed:April 19, 2023

[3] S. Ji, W. Xu, M. Yang and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221-231, Jan. 2013, doi: 10.1109/TPAMI.2012.59.

[4] T. Rose, J. Fiscus, P. Over, J. Garofolo and M. Michel, "The TRECVid 2008 Event Detection evaluation," 2009 Workshop on Applications of Computer Vision (WACV), Snowbird, UT, USA, 2009, pp. 1-8, doi: 10.1109/WACV.2009.5403089.

[5] K. Schindler and L. van Gool, "Action snippets: How many frames does human action recognition require?," 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 2008, pp. 1-8, doi: 10.1109/CVPR.2008.4587730.

[6]M. Yang, F. Lv, W. Xu, and Y. Gong, “Detection Driven Adaptive Multi-Cue Integration for Multiple Human Tracking,” Proc. 12th IEEE Int’l Conf. Computer Vision, pp. 1554-1561, 2009.