IOSR Journal of Business and Management (IOSR-JBM)
e-ISSN: 2278-487X, p-ISSN: 2319-7668 PP 09-17
www.iosrjournals.org

Embedded Machine Learning for Actual Image Reconstruction for Partial Criminal Sketch

Amit Kumar, Abhishek Jain
Department of Computer Science & Engineering, Quantum University, Roorkee, Uttarakhand India
Department of Computer Science & Engineering, Uttaranchal University, Dehradun Uttarakhand

ABSTRACT
We are dwelling in an electronic period where most trades are contact-less, virtual diversion stages are conventional and a piece of our everyday schedule is recorded either in a merciful of course furtive manner. Whether we are free in an online social event, everyday cordial media feed, a buddy related plan, a live gaming or video move, numerous bytes of our information are sent through an association to a server. The exceptional improvement of limit is in like manner engaging extraordinary many blended media content to be placed away locally on electronic devices but testing progressed assessments that are hampered by the assortment of such contraptions that were taken care of in a criminological examination community expecting to be dealt with by an expert as fast as could be expected. The size and proportion of information that requires examination is growing, provoking a wild high level criminological build-up. Phone clients can convey special substance like sound, pictures and accounts, and due to the web, can impart data by and large in no time flat.
To address the modernized quantifiable excess, the making mature enough evaluation models to help progressed logical assessments has been proposed. Though a couple of models perform well for the entire age range, specifically age ranges, for instance, the underage social event, the models perform totally lacking. Affecting elements on underage age appraisal have been surveyed and it has been settled that particular parts have strong, delicate or fragile associations with the machine-expected execution. These thoughts are key on the span of datasets and will yield further developed results on future arranged models.
The greatest underage dataset with age and direction marks has been accumulated and a few models have been investigated various roads in regards to various picture pre-taking care of methodology, cerebrum network plans, etc. Hyper-limit smoothing out was introduced and the best score for facial age appraisal was gotten. The scores were surveyed with a picked test dataset that contains faces that can be spotted by striking face finders like Viola Jones. A unique facial embedding approach was proposed and a scattering evaluation metric was introduced instead of a lone worth. The exhibition achieved beats the top tier facial age locaters for subjects more youthful than 25.
Keywords: *Spatial Temporal Graph Convolutional Network (ST-GCN), TRN etc.*

I. Introduction

This dissertation's main focus is cybercrime investigations and the use of deep learning (DL) to help investigators find missing people, victims, and suspects as well as to stop the backlog of unprocessed digital devices that are seized from crime scenes—a process known as "digital forensic backlog."

The amount of gadgets found at crime scenes has increased due to the impact of digital data on our daily lives. Finding digital information on a disk's surface has always been a laborious process for forensic investigators. Law enforcement agencies (LEAs) face challenges in obtaining information that holds probative value because processing such devices takes time and requires professionals, which can be costly. After it has been gathered, the evidence is kept in a digital forensic lab until it is examined. However, the backlog is growing as a result of the previously noted resource shortage. Processing the substantial volume of data that is often seized and processed requires a laborious, highly skilled digital forensic analysis.

Police officers and witnesses have used human soft biometric features, like age and gender, to describe unidentified victims. In some circumstances, the victim's age may influence how a crime is classified. In most jurisdictions, the creation, dissemination, and ownership of child sexual exploitation material (CSEM) are regarded as illicit actions. The availability of CSEM on the deep web has contributed to its rise, which in turn has led to a significant increase in cybercrime. Digital forensic investigators come into contact with such material during CSEM investigations, and some implicated personnel suffer as a result of such interaction.

II. Background Theory and Significance

With violent attacks occurring more frequently these days, public safety has also grown in importance, as evidenced by the pattern of rising crime incidences worldwide year over year. Incomplete data indicates that over the previous five years, public safety incidents have resulted in an average of over 300,000 unnatural deaths and 2.5 million disabilities, with associated economic losses of 700 billion yuan. Annual statistics based on real-world incidences indicate an increase in crime.

Advanced video surveillance, human behavior recognition, and crime prediction are required in light of the rising trend in crimes and the transition in crimes from monolithic to diverse (Altamimi et al. 2018). It is obvious that traditional video surveillance systems could not satisfy the demands of real-world applications. In the meantime, deep learning, pattern recognition, image recognition, and computer vision are all included in intelligent surveillance systems. According to Bengio, Simard, and Frasconi's 1994 evaluation of the literature, deep learning-based human behavior recognition produces higher-quality research results and considerably lowers social costs.

In recent years, there has been a steady increase in the frequency of criminal acts. The primary objective of the widespread usage of surveillance devices is to identify and profile individuals in public areas. Conventional video surveillance falls into two types: One can be identified in real time with the naked sight of a person. This approach has low detection efficiency and accuracy, making police officers more vulnerable to eye fatigue. Examining and looking for recordings as proof following an unusual event is another approach, but there is a delay in information that prevents a prompt real-time response. It is evident that modern practical needs are not being met by classic video surveillance systems; intelligent surveillance systems represent a new generation of video surveillance technology.

Research Content and Difficulties:

Rapid detection and identification of hazard behaviors is required in a wide range of settings. The inability of intelligent surveillance to promptly address and resolve security issues stems from the caliber of recorded footage and the prevalence of risky conduct. We require a quick and precise way to identify dangerous behaviors in the midst of an increase in crimes.

To identify and detect risky human behavior, a Spatial Temporal Graph Convolutional Network (ST-GCN) and TRN techniques are used. By establishing the degree of danger of the associated line of criminal activity, In order to identify illegal activity, we can construct a spatial temporal association network with the assistance of human behavioral skeleton sequences that are extracted from surveillance footage. Using visual feature information, a spatial temporal network model based on optical flow and the spatial temporal relationship found in video frames is constructed in the spatial temporal relationship network.

Research Questions:

The primary goal of this research is to support investigations by presenting a novel method for automatically identifying suspects and/or victims using soft biometric clues like gender and age. Address the digital forensic backlog, which has grown to be a well-known problem in LEAs across the world, while providing strong, forensically sound tools and methodologies for digital forensics (DF) that can support the evidence that is presented in court in a timely manner.

i. What are the driving forces behind the enhancement of face age prediction algorithms' efficacy?
ii. In what ways may digital forensic investigators benefit from and minimize their exposure to sensitive material through the application of Depp Learning?
iii. How can the backlog in digital forensics be helped by the creation and application of an age-prediction-based DL model?
iv. What effects will the integration of Deep Learning and Machine Learning techniques have on the forensic soundness, admissibility in court, and case throughput capacities of digital forensic processing?

a. Contribution of Research

This thesis proposes a method for extending spatial graph convolutional neural network to spatial temporal CNN for aberrant behavior categorization, using the SoftMax classifier, which has good robustness and high real-time performance. To study human behavior detection, attentional models are integrated with RGB-color films and 3D skeletal information.

A hazard recognition method based on TRN is proposed in this thesis to identify human behaviors, improve

efficacy, and accelerate the hazard behavior recognition process. The temporal association between the frames before and after the behavior is inferred using this method by using the TRN network.

2.1 Current Status of Behavior Identification

In contrast to human action recognition, which employs the observation of human actions contained in the supplied films, which uses video frames supplied as streaming data, human behavior prediction uses the full sequence of observations as input. Gradient boosted decision trees (GB-DT) (Martinezetal, 2018), artificial neural networks (Gunay, 2016; Chaawlaetal, 2019), logistic regression (Luetal, 2014; Cadelieri, 2017), and so on are the main prediction algorithms.

Crime mapping and forecasting have been used by law enforcement to take social security into account. The use of intelligent surveillance significantly updates the conventional techniques. Large volumes of data are typically used by crime analysts to assess whether a certain kind of crime was committed. The knowledge could be used to stop such incidents from occurring in the future.

Predictive policies are a tactic designed to lower crime rates. Crime management and prediction are improved by applying data-driven strategies. Using machine learning techniques, Reddy et al. investigated the development of crime predictions (Shah, Reddy, 2013). With the help of the idea of machine learning, we can forecast future outcomes based on the connections between distinct criminal case patterns and visualize enormous volumes of data.

A sparse representation was used to recreate the training set in order to propose a probabilistic model for video representation using a bag-of-words technique. The training sets are combined using the gathered security footage. Predicting human behavior in lengthy movies proved to be challenging since it requires a lot of processing power to identify actions, making it tough to deal with human behavior in intricate outdoor settings.A support vector machine-based multiscale discriminative prediction model was presented. Predicting human behavior necessitates a great deal of in-depth study in areas like model construction and behavioral representation.

A sparse representation was used to recreate the training set in order to propose a probabilistic model for video representation using a bag-of-words technique. The training sets are combined using the gathered security footage. Predicting human behavior in lengthy movies proved to be challenging since it requires a lot of processing power to identify actions, making it tough to deal with human behavior in intricate outdoor settings.A support vector machine-based multiscale discriminative prediction model was presented. Predicting human behavior necessitates a great deal of in-depth study in areas like model construction and behavioral representation. Making a decision in the prediction problem is the first issue to be solved.

Deep Learning techniques to recognize Human Behavior

The idea of deep learning was first presented in 2006, when the network's weights were initialized by unsupervised training techniques, and the parameters were then adjusted for model training.
Joint point behavior recognition in deep learning refers to the process of extracting characteristics from edited video footage. The CNN, Recurrent Neural Network (RNN), and GCN deep learning techniques are used to process articulation data. These techniques relate to the pseudo-image representation of articulation data. The networks comprise three types of joint point behavior recognition (JPBR): hybrid network (HN)-based, GCN-based, and RNN-based.

In computer vision applications including visual object categorization, object segmentation, posture estimation, and pedestrian detection, basic network models now yield good results. Natural language processing and digital image processing were the main sources of inspiration for the deep learning-based behavior recognition techniques currently in use in the field of human behavior recognition.

CNN (Convolutional Neural Networks)

The four distinct components of a convolutional network are the convolutional layer, pooling layer, classification layer, and fully connected layer, each of which serves a specific function. Feature extraction from the convolutional layer is the main application of the convolution process, which entails a weighted summation of the input data. The pooling layer samples the input data, reduces the amount of data, and modifies the parameters to reduce overfitting. The fully connected layer serves as a classifier, converting the input data into feature vectors.

The convolutional layer contains convolution kernels. CNN's convolution function entails superimposing a 15-window box of convolutional kernels over the source picture. The final element of the output feature's value is obtained by multiplying CNN's elements by corresponding elements more times the number of convolutional layers it possesses. In the convolutional layer, a smaller filter is applied before the selected image. The product is computed using weights generated during the convolution process. As a result, the filtered region and weights

have a relationship. The nodes in the output layer are connected to the output nodes from the input nodes via the hidden nodes.

Graph Convolution Network

In order to create a feature map for spatial feature extraction, the convolution in CNN is a type of discrete operation that basically uses a kernel with shared parameters to compute a weighted sum of central and neighboring pixel points. The weight of the convolution kernel serves as the weighting factor in this process.

Most data, or non-Euclidean data, don't have a regular spatial organization. Recommendation systems and graphs produced from electronic transactions are two examples of this kind of data. Each node in the network has a unique connection; some have three, while many just have one. Conventional convolutional networks do badly on these unpredictable inputs. It is challenging to choose a fixed convolution kernel that can accommodate the abnormalities of the entire network when non-Euclidean spaces are taken into account, such as the ambiguity around the number and order of nearby nodes.

Graph theory and deep learning are combined in a network structure known as a GNN. The most widely used ones at the moment are Graph Correlation Networks (GCN), Graph Attention Networks (GAN), Graph Autoencoder (GAE), Graph Generative Networks (GGN), and Graph Spatial Temporal Networks (GST). In original GNN networks, the features of points and edges are passed into the network together.

GCN is a combination of graph and convolution. The spatial approach, which is separated into point classification and graph classification, defines the convolution directly on the graph and works on nodes that are closely related. The Laplacian matrix, on the other hand, is the foundation of the spectral approach and has weak generalizability despite its tight relationship to the graph.

Similar to CNN, GCN is a feature extractor that enhances the quality of features extracted from graphs for use in prediction, graph classification, and node classification. Generally speaking, GCN is divided into two groups: A spectral viewpoint, such as a spectral approach, investigates convolutional filters directly in graph nodes and evaluates the local character of graph convolution. Following the second premise, the suggested ST-GCN models build a GCN using each node's neighborhood distance as a basis before building the CNN kernel in the spatial domain

III. RESEARCH METHODOLOGY:

First, the dataset is gathered for the methodology. Second, an ST-GCN-based approach for recognizing human behavior is put forth. Next, strange conduct a technique for the identification of deviant behavior based on temporal relational networks (TRNs) is presented.

Data Preprocessing

UCF101, a dataset with 40 classes, has been chosen. Of the samples, about 2,400 (or 80%) are selected for training, and another 300 video clips are selected for testing. The samples that are still present are categorized as non-criminal cases. The clips have a resolution of 340x256 and are captured at 30 frames per second (FPS).

The four main stages of data preprocessing are data cleaning, data integration, data statistics, and data conversion. The primary goal of data cleaning is to fix issues with the data by incorporating missing numbers, reducing noise in the data, or eliminating contours.

Because of the aims, there are many different causes of data loss. The distribution and significance of the variables will be the main determinants of how to handle these missing data. Methods of padding or deletion are used, depending on the frequency of missing values. When populating continuous variables, the mean and random difference approaches are utilized; for discrete variables, the median or dummy variable is employed.

Using box plots and MAD statistical techniques along with a rather easy and understandable method of identifying outlines for variables, outlines are considered as anomalies that compromise the quality of the data during data processing. The number and impact are taken into account while deciding whether to delete the data.

Noise is the difference between the observed and actual data values for a variable, or its random volatility. Using the mean, center, or boundary values of each container—rather than all of the numbers within—the frequency within each equal-width container is separated throughout the process of dividing the data. For the data, this acts

as a smoothing filter. One method is to create a regression model of that variable and the predictor variable using an approximation based on the regression coefficients and the predictor variable.

Behavior of Criminals

Criminal activity can be divided into three categories: (1) Vandalism; (2) Abuse, assault, burglary, and shoplifting; and (3) Shooting and Arrest. Table 1 provides a definition for various human actions.

In this book chapter, the dataset UCF101—which contains 540 crime classes—is examined. Of the 300 films used for testing, 240 were used for training. We refer to the remaining clips as non-criminal cases. The clips have a resolution of 340x256 and a frame rate of 30 frames per second. The human actions in the dataset are categorized as Grade I, Grade II, and Grade III in this book chapter, as Table 1 illustrates. False alarms are less likely when criminal actions are graded and alarming interventions are implemented based on the levels of criminal activity.

Table 1: Criminal Behavior Levels

Grades of criminal behaviours	Behaviour performance	Basis of classification
Grade I	Vandalism	Destruction of public goods with purpose and intent, using violent means.
Grade II	Abuse, Assault, Burglary, Shoplifting	Using violent means with purpose and intent to harm another person behaviour.
Grade III	Shooting, Arrest	Transgressions, which are acts that may result in the safety of another person's life.

Behavior Identification using ST-GCN

ST-GCN is created by combining TCN and GCN. On visual data in the temporal dimension, TCN performs convolutional processes; on data in the spatial dimension, GCN does the same operations. Graph theory is the foundation of GNN, while GCN is a subset of it. Structured data with Euclidean distance reserves, like 2D graphics, 1D sounds, and so forth, is usually handled by neural networks. Non-Euclidean distance data, like that from social networks, transportation networks, etc., cannot be handled directly by the normal network structure; instead, GNN is utilized to handle this type of data. The existing method for target skeletal activity detection adds the connection of nearby important sites to improve accuracy.

Three steps are usually involved in producing a spatiotemporal map of a skeletal sequence. Making a spatial map of the natural joints of the human body for each frame in a video clip is the first stage. We also combine the same points from two surrounding frames to form the temporal boundaries. In the end, all of the edges from Steps 1 and 2 (which constitute the set of edges E) and all of the crucial points from the input video frames (which form the set of nodes) combine to form the required spatiotemporal map. There are two subsets of the edge set. The concatenation subset, represented by the symbol Eq., is found within the human skeleton.

TRN Based Behavior Identification

The article proposes a sparse temporal sampling technique that splits the input video into K segments, independent of the video's length. CNN is then used to extract spatial characteristics from each segment and perform feature-level fusion to discover a random time segment 42. Dense sampling of video frames is required for temporal segmentation networks. Lastly, the formula (3.7) is used to do SoftMax categorization.

$$TSN(T_1, T_2, \ldots T_K) = H(G(F(T_1; W), F(T_2; W), \ldots F(T_K; W))). \qquad (3.7)$$

The ability to understand the connections among people or visual objects in the temporal domain is referred to as temporal relational reasoning. TRN is a real-time temporal relational reasoning framework for inferring temporal

links between frames at the video level, building on the TSN framework. The main contributions of TRN are the development of new fusion functions that characterize the interactions between different temporal segments and the improvement of video-level robustness through multiscale feature fusion in the temporal dimension.
This chapter presents a TRN approach to human behavior recognition. The method uses a TRN network to analyze the temporal relationships in the frames before and after the movie by learning and reasoning about the temporal relationships of every frame in the video. By sparsely sampling the video frames, the TRN network achieves considerably better results in distinguishing human postures and can accurately recognize human interactions. Our research indicates that 3D convolutional networks and dual-stream networks do not generally yield as intuitive visual perception as TRN-based network models do.

In the real world, humans use temporal relationships to reason through the past, present, and future events.

Figure 1: Temporal Relationship in a Video

Figure 2 shows how the poses of human bodies reflect the temporal relationship between two states in a realistic scenario, and how the relationship between two video frames infers the temporal relationship between temporal states. Many temporal linkages, including long and short temporal relationships, are present in a single human behavior.

Recognizing human behavior in digital movies is a hot topic in computer vision, but neural networks have a tremendous barrier when it comes to the ambiguity of temporal behavior description.

ST-GCN Results

The position of skeletal joints is used as an input parameter by the ST-GCN arranged. All video resolutions were initially shrunk to 340 by 256, and the video frame rate was normalized to 30 frames per second in order to acquire the joint nodes. 18 joints' locations on each frame of the clip were subsequently assessed, using the coordinate system's 2D coordinates to reflect the locations of human joints. The human skeleton is thus formed by combining an array of 18 nodes, each of which is applied to represent a joint. In terms of multi person detection and skeleton extraction, each video clip's two joints that had the highest degree of confidence in the human skeleton were chosen and turned into a skeleton sequence.

The UCF101 crime dataset, which includes a wide range of crimes, is used in this thesis. We begin with the ST-GCN environment's setting. Because Microsoft Windows 10 is the operating system, we must use Cmake to compile OpenPose and AlphaPose when setting the ST-GCN model. There have been issues with the target build folder's data while downloading and installing Cmake. Furthermore, none of the necessary files are produced by the bins listed under the build file.

There are numerous benefits associated with Google Colab. First off, there are free GUP options in the cloud, but it's crucial to remember that Colab only allows a certain amount of GUP usage; Secondly, the environment configuration on Colab is less problematic; and last, it is faster. We conceive of Colab as a virtual computer running Ubuntu with GPUs; the only difference is that we can only access them via the command line. We treat the Google drive as a hard disk and construct a drive folder in the virtual system after mounting it. If Colab is utilized nonstop for more than 12 hours, the system will terminate the active program 52 and return the virtual machine.

The ST-GCN method for recognizing human behavior is displayed in Figure 4.1. The text in the movie depicts the outcome of detecting and recognizing human behavior, while the white dots and lines in each image indicate the skeleton contour of the human body. Based on the experimental findings, we can conclude that the algorithm successfully completes behavior detection since it can reliably identify the human body and accurately estimate human activity.

In this thesis, we evaluate OpenPose with AlphaPose for the detection of anomalous behavior using the UCF101 dataset. Three parameters—prediction accuracy, recall, and accuracy—are included in the equations (4.1–4.3) that represent the evaluation criteria versus human posture. The metrics accept values between 0 and 1, which include.

$$a = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{4.1}$$

$$p = \frac{Tp}{Tp + Fp} \tag{4.2}$$

$$r = \frac{Tp}{Tp + Fn} \tag{4.3}$$

Figure 4.1: Risky Behavior Identification

Table 4.1: Evaluation results from OpenPose

	Grade I	Grade II	Grade III
Prevision	60.16%	61.91%	58.13%
Recall	73.52%	72.43%	79.62%
Accuracy	81.24%	72.34%	81.17%

Table 4.2: AlphaPose Evaluation Results

	Grade I	Grade II	Grade III
Prevision	73.21%	75.32%	71.26%
Recall	64.13%	61.71%	61.01%
Accuracy	83.05%	84.53%	82.84%

Our research indicates that OpenpPose's accuracy in identifying human actions is not perfect, leading to mistakes at critical junctures in the identification process. In contrast, the computed sequences can be accurately imported into the behavior recognition model by AIphaPose for categorization. The ST-GCN method has a high recognition accuracy and does not require pre-obtaining the background information from the video. Additionally, ST-GCN does well in multiperson behavior recognition; however, as Figure 3 illustrates, recognition takes longer.

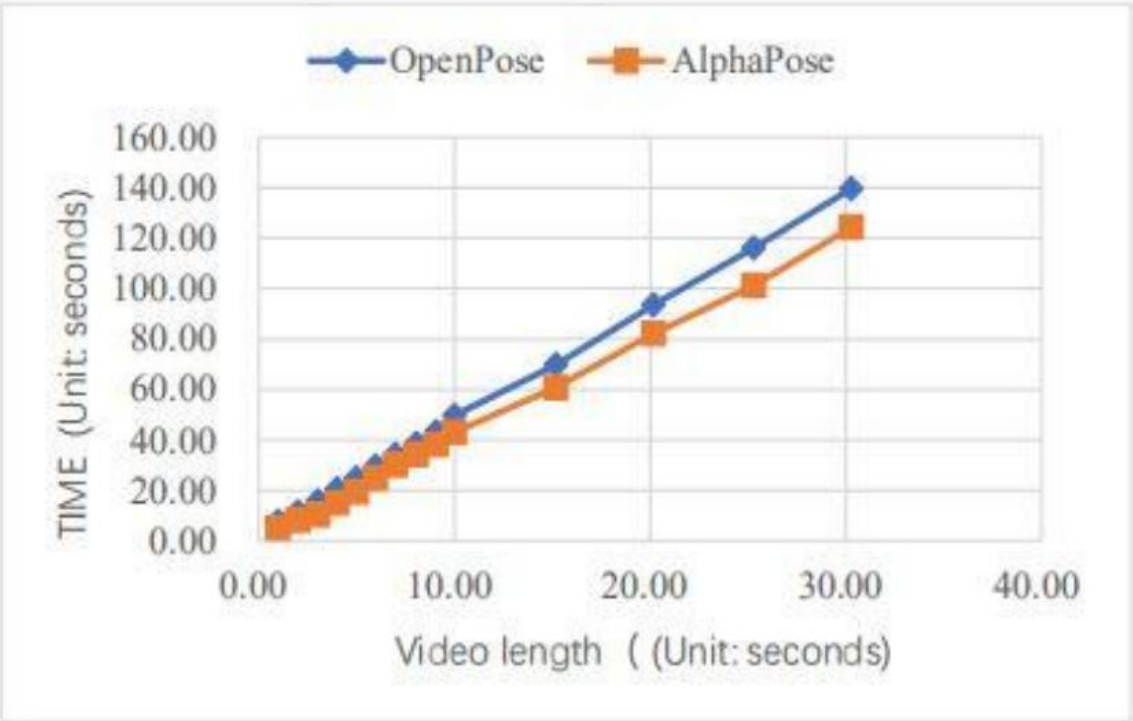

Figure 3: Timelines

IV. CONCLUSION AND FUTURE WORK:

In this chapter, we highlight the flaws and restrictions in our techniques; more work has to be done in this area in the future.

Recognition of human behavior has long been a significant area of study for computer vision and pattern recognition researchers. The ability to recognize human activity in videos is one of the most widely discussed issues. Deep learning has advanced quickly in the field of artificial intelligence recently, and this has led to the rapid growth of theory, which has produced more novel solutions to this issue. The main focus of this thesis is on how to more effectively recognize human activity in videos by utilizing deep learning theory and related techniques.

In this thesis, we focus on deep learning-based video human behavior recognition and obtain multiple study outcomes: To overcome the problems of overfitting and sluggish convergence of training on a small collection of human behavior video samples, a method of fusion of temporal and spatial features is developed. When dual-stream convolutional neural networks are used for human behavior on a small collection of human behavior video examples, these problems occur. Even if the network model and feature fusion technique created in this thesis may effectively finish the task of human behavior recognition and greatly boost the network's recognition skills, there are still certain issues that need to be solved and improved upon.

The literature indicates that TRN models, which forecast human criminal conduct in the ST-GCN, are essential for security prevention and control. They also help with resource allocation for law enforcement and other matters. However, several limitations exist in the literature.

The challenge of human behavior classification is essentially responsible for the inability of present behavior recognition techniques to anticipate human conduct. We must address the issue of accurately differentiating between behavioral prediction and behavioral classification.

Comparably, it is far more challenging to precisely separate human speech patterns in a video in order to make predictions. Many sub-activities can accompany a single human behavior; these sub-behaviors add to the randomness of the human behavior and have an impact on how well models identify behaviors. Since the length of the video clip is uncertain in genuine settings, the main concern becomes how to guarantee the models' long-

term memorability. The predictive behavior is much more urgent at the semantic level in complex acts with large time spans. We'll employ relational inference networks to forecast people's intents. speculating about what a person may do next.

The data and computer setup used in our trials meant that a significant amount of time was spent training the model. Simultaneously, the lack of sufficient data affected the detection findings, making the extraction of skeleton and optical flow-based information take longer. Simultaneously, the movies' poor noise management leads to a decrease in the accuracy of object recognition. In this thesis, human crime is categorized in a subjective manner without a set definition; in certain contexts, regular activities might also be considered deviant.

REFERENCES:

[1]. Al-Sarayreh, M., Reis, M., Yan, W., Klette, R. (2019) A sequential CNN approach for foreign object detection in hyperspectral images. Computer Analysis of Images and Patterns.

[2]. An, N., Yan, W. (2021) Multitarget tracking using Siamese neural networks. ACM Transactions on Multimedia Computing, Communications and Applications.

[3]. Aggarwal, J., Ryoo M. (2011) Human activity analysis: A review. ACM Computing Surveys, 43(3): 16.

[4]. Aggarwal, J., Xia, K. (2014) Human activity recognition from 3D data: A review. Pattern Recognition Letters, 48, 70-80.

[5]. Aggarwal, J., Xia L., Chen, C. (2012). View invariant human action recognition using histograms of 3D joint. IEEE Conference on Computer Vision and Pattern Recognition Workshops, 20-27.

[6]. Agrawal, P., Nair, A.V., Abbeel, P., Malik, J., Levine, S. (2016). Learning to poke by poking: Experiential learning of intuitive physics. Advances in Neural Information Processing Systems, pp. 5074–5082.

[7]. Ahmed, J., Rodriguez, M., Shah, M. (2008). Action matches a spatial temporal maximum average correlation height filter for action recognition. IEEE CVPR. Alice, G., Lai, G. (2014). A survey on still image based human action recognition. Pattern Recognition, 47(10), 3343-3361.

[8]. Altamimi, A., Ullah, H., Uzair, M., et al. (2018). Anomalous entities detection and localization in pedestrian flows. Neurocomputing, 290, pp.74-86.

[9]. Andonian, A., Zhou, B., Oliva, A., et al. (2017). Temporal relational reasoning in videos. IEEE CVPR.

[10]. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B. (2014). 2D human pose estimation: new benchmark and state of the art analysis. IEEE CVPR.

[11]. Anguelov, D., Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Erhan, D., Vanhoucke, V., Rabinovich, As. (2015). Going deeper with convolutions. IEEE CVPR.

[12]. Azizpour, H., Razavian, A., Sullivan, J., Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. IEEE CVPR Workshops.

[13]. Badii, C., Bellini, P., Ddfino, A., et al. (2019). Smart city IoT platform respecting GDPR privacy and security aspects. IEEE Access.

[14]. Bakshi, S. Guo, G. Proença, H. & Tistarelli, M. (2020) Visual surveillance, biometrics: Practices, challenges, and possibilities. IEEE Access.

[15]. Baradel, F., Wolf, C., Mille, J. (2017). Pose-conditioned spatial temporal attention for human action recognition. CoRR, abs/1703.10106.

[16]. Baradel F, Wolf C, Mille J. (2017). Human action recognition: Pose-based attention draws focus to hands. IEEE International Conference on Computer Vision Workshops (ICCVW), pp.604–613.

[17]. Begleiter, R., El-yaniv, R., Yona., G. (2004). On prediction using variable order Markov models. Journal of Artificial Intelligence Research, vol. 22, pp. 385-421.

[18]. Bengio, Y., Simard, P., Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2):157– 166.

[19]. Bengio, Y., Ducharme, R., Vincent, P., et al. (2003). A neural probabilistic language model. Journal of Machine Learning Research, 3(2): 1137-1155. Bengio, Y., Glorot, X. (2010). Understanding the difficulty of training deep feedforward neural networks. AISTATS.

[20]. Billings, D., Yang, J. (2006). Application of the ARIMA models to urban roadway travel time prediction-A case study, IEEE SMC, pp. 2529–2534.

[21]. Bischof, H., Zach, C., Pock, T. (2007). A duality-based approach for realtime TV-L1 optical flow. Joint Pattern Recognition Symposium, pp.214-223.

[22]. Blank, M., Gorelick, L., Shechtman, E., et al. (2005). Actions as space-time shapes. IEEE International Conference on Computer Vision (ICCV'05), pp. 1395-1402.

[23]. Bobick, A., Davis, J. (2001). The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis & Machine Intelligence, 23(3):257-267.

[24]. Boser, B., LeCun, Y., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L. 74 (1989). Backpropagation applied to handwritten zip code recognition. Neural Computation.