MOUSE BEHAVIOR TRACKING USING 3D DEEP LEARNING

Shashank Nigam, Padmini Ramesh, Ramdas KrishnaKumar

Institute of Systems Science, National University of Singapore, Singapore 119615

ABSTRACT

Video forms the major digital data being consumed. From entertainment to medical research videos from key applications. Appropriate moderation and classification of data is required to ensure appropriate use. Complex data representation makes this task difficult for traditional algorithms. Manual classification which involve monitoring long duration video, poses many limitations. A high performance automated recognition system can reduce such limitations posed. Deep learning, has capability to learn complex features and build models which can be generalized for such activities. We propose a two streamed 3 D deep learning model for automated video behavior analysis. The model is trained on the mice data-set to classify different mice behavior such as drink,eat,groom,hang,micro-movements,rear,rest and walk. Model makes use of features derived from randomly sampled video frames and optic-flow using 3D convolution network. With appropriate data and fine tuning the model can be generalized for behavior analysis in video data

Index Terms— Behavior Analysis, Two stream Approach,Optic flow, 3D convolution.

1. INTRODUCTION

Videos are major form of digital data being consumed. With the advent of mobile networks videos form the major data being transmitted and consumed. It has been projected that video consumption rises 100% every year, with 55% watching videos online every day, 92% of the video consumers share the video with others. Social media is the major generator for such content [1]. With such huge amount of data being consumed, there is a larger need for moderation. Any triggering and offensive content can prove to be dangerous to both minor and adults. It is important to screen videos containing violence, illicit,illegal or inappropriate content online. With about handful of people monitoring such content globally, the task is difficult to process [2]. With job being emotionally taxing, with poor working conditions it poses high human limitations [3]. Most of the content moderation involves understanding the object behavior context in the content. A high performing automated process can help reduce and eliminate the human limitations posed in such activities.

Video analytics can be used in medical research. Animal behavior recognition is the key tool for accessing the effects of disease,drugs,gene mutation and perturbations in neural circuits. Such task are basic imitation and human behavior and allows interpolation of different behavior [4]. Such task were traditionally limited to human analysis, which posed serious drawbacks especially in classification of fast moving animals.Researchers widely make use of many animals such as fruit files,mice or primates for studying the biology,psychology or developing new therapies or medicines. In most of the research observing such behavior are crucial for answering research questions. Annotation of behavior involve hours of hard work. An automated task can help delegate such activities to computers. A well performed system can make such research much easier to reproduce [5]. The proposed deep learning system is used for the classification of eight stereotypical mouse behaviors such as eat, drink groom,hang, micromovement, rear, rest and walk. The dataset is prepared by continuous monitoring of the mouse behavior in the caged environment. The video was captured at 360x240 pixel resolution and consist of a 10 hour of continuous labeled monitoring. The data was collected at various lighting condition and light sources. The environment contains a drinking tube and a feeder on top of the cage. From the full annotated dataset 4200 short clips containing unambiguous single behavior were extracted [4]. The model was trained on this clipped dataset.

The model is inspired from the Two stream Convolution Network for Action Recognition in Videos [6] and R(2+1)D[7] networks. The model combines the feature extracted from video frames and its corresponding optic-flow data. The feature extractor network makes use layers consisting R(2+1)D convolution, which is a combination of spatial feature extractor using 2D convolution and temporal feature extractor using 1D convolution. Each of the layer is interleaved with batch normalization and RELU which gives the model capability to understand additional complexity as compared to simple 3 dimensional convolution, with similar number of parameters.

The model is trained on a limited dataset and can be further generalized based on availability of good annotated dataset.

2. RELATED WORK

One of the system implementation by H.Jhuang et al. [8] involves traditional machine learning approach for the feature extraction and further classification. The proposed system consisted of a feature computation module and a classification module.

The feature computation module involves following steps[8]:

- A background subtraction procedure is applied to compute a foreground mask for pixels belonging to the animal versus the cage.
- A subwindow centered on animal is cropped based on the location of mouse.
- Space-time motion features are derived from combinations of the response of afferent units that are tuned to different directions of motion as found in the primate primary visual cortex.
- Position and velocity based features are derived from the instantaneous location of the animal in a cage. These features were computed from a bounding box tightly surrounding the animal in the foreground mask.

Classification module consisted of following elements:

- The output of the feature computation module consisted of a 310 features per module that were passed to a statistical classifier namely SVMHMM (Hidden Markov Model Support Vector Machine)

Following is the vision based model for the system:2

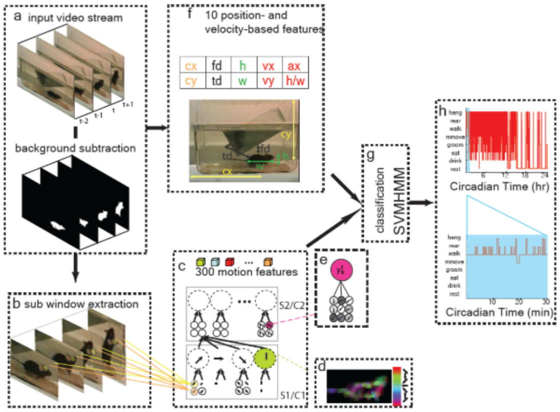

Fig. 1. Vision-Based System for Automated Mouse Behaviour Recognition System[8]

The model depends highly on the manually extracted features for the learning the behavior of the model. It depends on the background modelling of the environment where the experiment was conducted.The model could obtain a general accuracy of 77.3% for the clipped dataset and 78.3% for full dataset. The model however has following limitations:

- The model depends mostly on background modeling, such as Single mouse in home cage, white background, Video recorded from side view, Process one cage at a time [8]
- The model behavior cannot be generalized to other animals like raccoon even when both the animals share similarities in appearance [5]
- The model can not be generalized to multiple object behavior recognition in the same environment

Another approach proposed by Nguyen Ngoc et al [5] makes use of state of art deep learning model for the identification of the behavior.The behavior classification is done constructing an i3D model and R(2+1)D model for the classification. I3D model implemented is an extension of inflated Inception v1. The model is initialized using the weights of Inception v1,inflated and applied across all the channels. The model is trained on different data augmentation and fused with the optic flow of the frames using weighted average. A similar approach is followed for training of R(2+1)D model. The model is initialized with the weights derived from the R(2+1)D model trained on the ImageNet data. Although the system provides a better performance than the previous proposed approach it still lacks in classifying behavior that are ambiguous to other behavior. Overall the system provides an accuracy of 96.9% for I3D model and 96.9% for R(2+1)D model

3. PROPOSED APPROACH

For predicting mouse behaviour from the given clipped dataset we propose a Two streamed convolution network. Traditionally a two stream network consist of a spatial convolution network and a temporal convolution network. The spatial convolution network allows extracting the key image features thus allowing the model to be trained on large annotated dataset like ImageNet. Temporal stream allows extracting temporal feature of a video stream[6].We propose a two stream network where one stream is used to extract feature of video(both spatial and temporal) and other stream is used for extracting features from optic flow of the corresponding frames. The feature maps are combined later fusing using averaging. Optic flow captures the apparent motion between 2 consecutive frames caused by object or camera. It is a 2D vector filed where each vector is a displacement vector showing the movement of points from the first frame to second[9]. This allows additional motion features to be integrated with the original network which supplements the temporal features extracted using the first stream.

3.1. Data Preparation

The data is prepared by capturing mouse movements using JVC digital video camera (GR-D93) with a frame rate of 30fps [4]. The environment consist of a home cage containing a single mouse.The home cage consist of the feeder and drinking tube located at the top of the home cage.The daily activity of the mouse was captured at different lighting conditions. Stereo typical behavior of mouse were labelled from the captured video

[width=7.7cm]

Table 1. label description

label	Proposed approach
drink	Mouse drinking from the drinking tube
eat	Mouse touching his head
groom	Mouse eating from the feeder
hang	Mouse hanging from the top of the cage
micromovement	Mouse slightly moves around
rear	Mouse rears to the side of the cage
rest	Mouse stays stable or sleeps.
walk	Mouse walks or run inside the cage

Video recorded from each of the source is labelled per segment. A clipped data set consisting of 4200 short clips of the original data is prepared each of the short clip describing a single mouse movement is prepared. The number of frames per label, number of segments per label and average number of frames per segments were calculated.

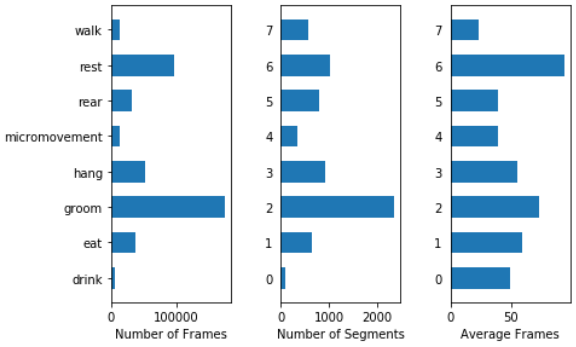

Fig. 2. Short clipped data segment histogram

Collectively groom has the highest number of frames and segments followed by rest. Rest corresponds to have highest number of average frame per segment followed by groom. Walk has minimum number of average frame.For walk label there were 24 frames per segment on average. In order to prevent loss of the possible data in the data set the for training 16 frames were selected at random, as a depth map from the available video segments to train the video stream.

The data was prepared at a resolution of 360x240. For limiting the parameters of the model to be trained each of selected frames was resized into 190x120 pixels for training.Further for reduction of dimensions the individual frames were converted to gray-scale before feeding into the network. Data set is collected from different video source in separate folder. A cross validation training is carried out on the data set where the test set is changed after every 10 epochs

3.2. Optic Flow

For enhancing the feature maps an optic flow for each of the selected frames were calculated. Optic flow gives the pattern of the apparent motion of image objects between 2 consecutive frames. The movement can be caused by movement of the camera or the object. In the current scenario, our camera remains stable while the object(mouse moves). Optic flow makes following assumptions [9]

- The pixel intensities of an object do not change between the consecutive frame
- Neighboring pixel have similar motion.

Considering each pixel $I(x, y, t)$ an object moved by dt, optic flow with the above assumption can be given as

$$I(x, y, t) = I(x + dx, y + dy, t_d t) \quad (1)$$

[9]

Approximation by Taylor series gives

$$f_x u + f_y u + f_t = 0 \quad (2)$$

[9]

The equation gives optic flow in f_x and f_y directions respectively.

As frames are selected at random both the condition might not be satisfied in some video. For appropriate classification of the optic flow images we make use of Lucas Kanade algorithm for calculation of optic flow between the frames.

The optic flow gives 2 frames for each of the corresponding input frame. For processing throughout the network the optic flow obtained in both the directions are summed together to give an input feature map.

3.3. R(2+1)D

Video has a complex data structure which includes both spatial feature including images per frame and temporal feature representing each frame at time t.A 2D convolution can only be used to extract feature from spatial component while loosing the temporal component in the process.

A 3D convolution is well suited for spatial-temporal feature learning and are well suited for such data structures [10]. In a 3D convolution the operation are performed spatially and temporally. This allows multiple images to be considered together thus preserving the temporal feature.

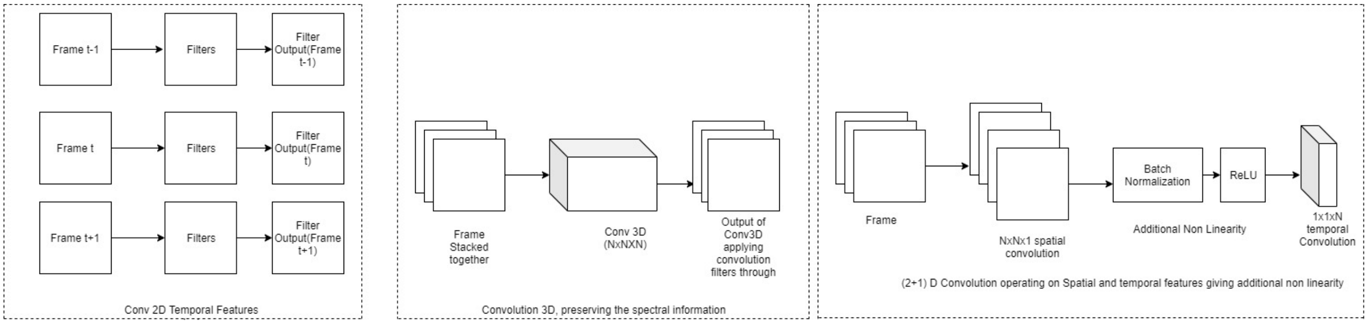
Frame t-1
Filters
Filter Output(Frame t-1)
Frame t
Filters
Filter Output(Frame t)
Frame t+1
Filters
Filter Output(Frame t+1)
Conv 2D Temporal Features
Frame Stacked together
Conv 3D (NxNXN)
Output of Conv3D applying convolution filters through
Convolution 3D, preserving the spectral information
Frame
NxNx1 spatial convolution
Batch Normalization
ReLU
Additional Non Linearity
1x1xN temporal Convolution
(2+1) D Convolution operating on Spatial and temporal features giving additional non linearity

Fig. 3. Operating on Video data.

Although C3D networks solves the issue for the temporal feature an additional non linearity can be added through the use of a Convolution 2D applied on the spatial component followed by a 1D convolution on temporal feature [7]. Each of the layers are separated by an additional Batch Normalization and ReLU activation.Although the structure as a whole represent a convolution 3D ,the additional Batch Normalization and ReLU gives network an additional non linearity component. This additional non linear linearity allows network to learn more complex features as compared to a simple 3D convolution network. Also the separation of layers allows efficient back propagation of the weights there by allowing the network to learn faster. Figure shows the difference between Conv2D,Conv3D , Conv(2+1)D 3.

R(2+1)D approximates a full 3D convolution by a 2D convolution followed by a 1D convolution, thus decomposing the spatial and temporal component into 2 separate steps. A 3D convolution filter of size $N_ixtxdxd$ is decomposed as $N_(i-1)x1xdxd$ and $M_ixtx1x1$. M_i is a hyperparameter that determines the intermediate subspace where the signal is projected between spatial and temporal convolutions. The parameter calculation for M_i can be given as [7]

$$\frac{td^2N_(i-1)Ni}{d^2N_(i-1)+tN_i} \tag{3}$$

With the above equation we can approximate the number of parameters in Conv3D with R(2+1)D with additional non linearity.

We constructed our network with a series of R(2+1)D layers. A simple R(2+1)D layer can be given by 4,While a residual network defined using the R(2+1)D. The structure of the same is denoted by 5 . Each of the simple convolution layer is followed by a residual block followed which is further down sampled by a residual block denoted using **??**. The structure for down sample residual can be given by

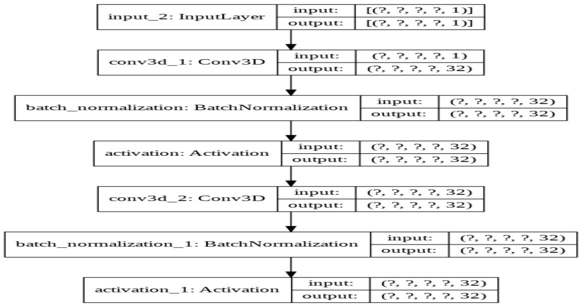

Fig. 4. Simple Convolution 2D 1D network

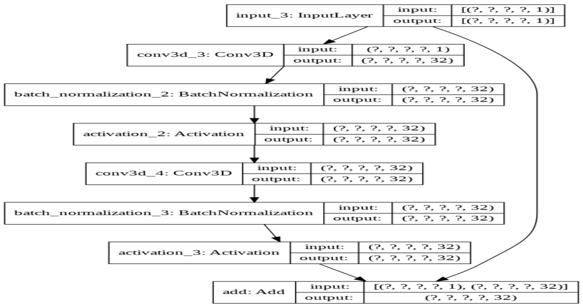

Fig. 5. Simple Residual Block

3.4. Two Stream Network

2 stream network was originally built as an extension towards still image recognition and action recognition. It involved 2 components a spatial component to recognize the set of features and identify the image and a temporal component to identify the optical flow for action recognition. The feature map obtained from both the network are fused together to give a unified class score. Based on application the feature maps can be fused early or can be fused later[6]. While stacking frames together give a good feature extraction for the action recognition it increases the search area to investigate the motion. Optic flow gives a motion description occurring between two consecutive frames. Histogram of Flow (HoF) was a handcrafted feature traditionally used for action recognition. Although HoF calculation is complicated we propose developing a deep learning feature map to learn these feature and recognize the action from the optic flow data.The feature map extracted from optic flow can further supplement the feature

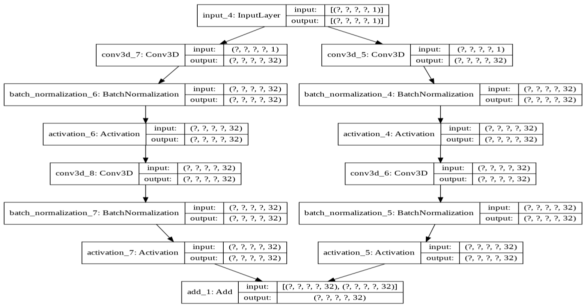

Fig. 6. Down sample Residual Block

extraction from frame stacking.

We propose a two stream network with late fusion 7

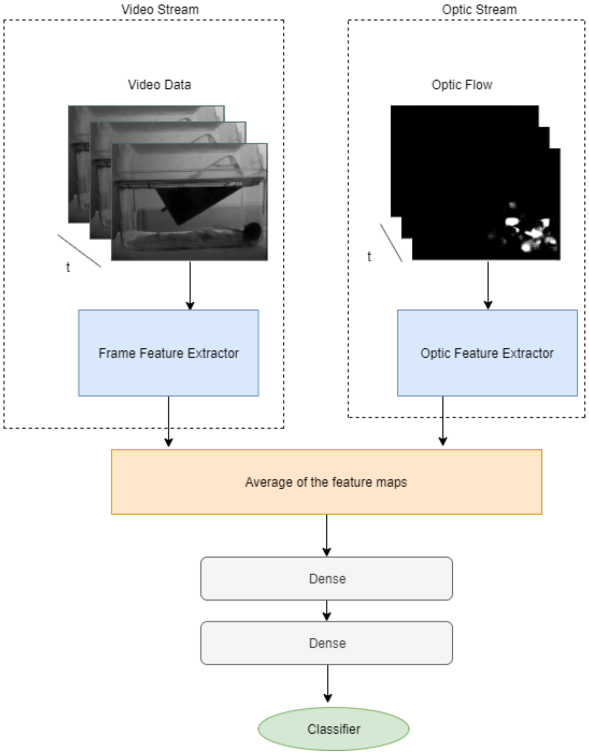

Fig. 7. Two stream network classifier

The first stream extracts the feature map from the stacked frames. The second stream extract the feature map from the corresponding optic frames stacked together. The feature maps obtained from both the averaged and passed to a classifier for classification of the behavior from the given input clip.

3.5. Feature Extractor

The Feature extractor is a series of Conv(2+1)D layers. It follows pattern of simple Conv(2+1)D followed by Residual convolution layer followed by residual down sampling. The model consist a total of 129 million parameters.The feature map is represented in 8.

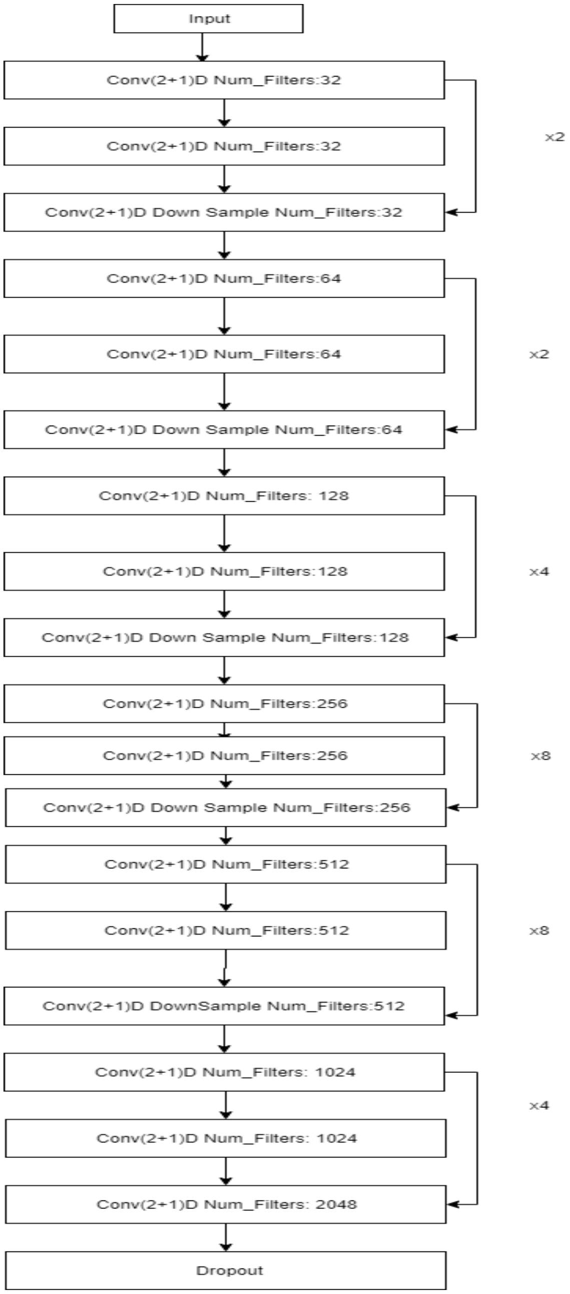
Input
Conv(2+1)D Num_Filters:32
Conv(2+1)D Num_Filters:32
Conv(2+1)D Down Sample Num_Filters:32
x2
Conv(2+1)D Num_Filters:64
Conv(2+1)D Num_Filters:64
Conv(2+1)D Down Sample Num_Filters:64
x2
Conv(2+1)D Num_Filters: 128
Conv(2+1)D Num_Filters:128
Conv(2+1)D Down Sample Num_Filters:128
x4
Conv(2+1)D Num_Filters:256
Conv(2+1)D Num_Filters:256
Conv(2+1)D Down Sample Num_Filters:256
x8
Conv(2+1)D Num_Filters:512
Conv(2+1)D Num_Filters:512
Conv(2+1)D DownSample Num_Filters:512
x8
Conv(2+1)D Num_Filters: 1024
Conv(2+1)D Num_Filters: 1024
Conv(2+1)D Num_Filters: 2048
x4
Dropout

Fig. 8. Feature Map

The model consist of alternating residual and downsample layer before its converged with the other layer. The Input image is of size 190x120x16. The down sampling allows the the size of the input feature to reduce to 6x1x1 till the point it reaches dropout.

4. EXPERIMENTAL RESULTS

The model was trained for about 40 epochs on Tesla K80 GPU. The learning rate was kept at 0.0001 and Adam optimizer was used for the optimization of the training data. Batch normalization was initialized with a momentum of 0.9 and an epsilon of 0.0005.Due to limitation of the data set cross validation approach was followed for training and testing after every 10 epochs. Data set consist of video clips from different recorders. For each of the 10 epochs one of the folder was considered as test data while rest were used as a train data. The result of the training can be summarized in the form of confusion matrix 9

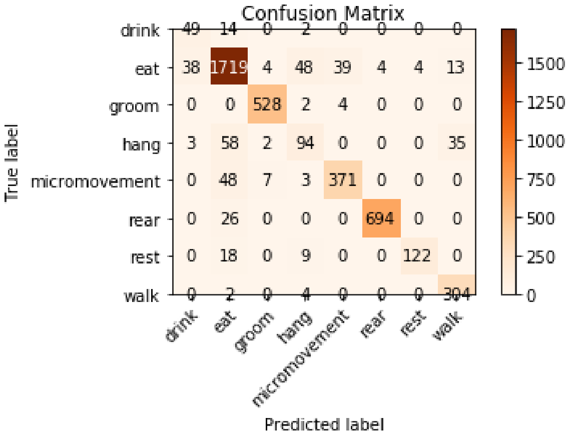

Fig. 9. Confusion

It can be seen from the confusion matrix the model performs good on most of the classes however it does confuses on classes drink and eat. This is due to the proximity of the feeder and the water dropper which makes it difficult to predict for such classes. The model gives around 90% accuracy on the clipped data set. The mode performs better in comparison of the the base model which makes use of the hand picked features . Below table summarizes the result of the same.

Table 2. The performance comparison.

Model	Clipped Database Accuracy
Reference Model1[4]	76.7%
Reference Model2[5]	96.6%
Proposed Model	90.93%

While reference model2 was build on a pre-trained model the proposed model was built from scratch. The model performance is better than the base model which made use of traditional handcrafted features to infer the behavior from the data

5. CONCLUSIONS

Although the model performance is better than the traditional approach of using handcrafted features it is still not as accurate as the state of the art models for behavior analysis such as i3D or R(2+1)D. With additional training data the model can be generalized to other action recognition with limited data and short time to deployment.The two stream based approach made use of the optic flow as a feature, the model can further include other computer vision motion based features to make accurate prediction. While the model is robust against different lighting condition in contrast to traditional machine based approach it is not as robust to other data augmentation technique which might represent camera position change. Further with different view point a depth 3D map can be used for better prediction of the behavior across the frames.

6. REFERENCES

[1] “17 stats and facts every marketer should know about video marketing :https://www.forbes.com/sites/miketempleman/2017/09/06.17-stats-about-video-marketing//48ae565567f,” .

[2] “Content moderation :https://www.accenture.com/ie-en/$_acnmedia/pdf-47/accenture-webscale-new-content-moderation-pov.pdf,''$.

[3] “The underworld of online content moderation :https://www.newyorker.com/news/q-and-a/the-underworld-of-online-content-moderation,” .

[4] “Vision-based system for automated behavior recognition :https://cbmm.mit.edu/mouse-dataset,” .

[5] M Reza Faisal Bendy Purnama Nguyen Ngoc Giang, Favoirisen Rosyking Lumbanraja, “Applying deep learning models to mouse behavior recognition,” in *Journal of Biomedical Science and Engineering*, 2019.

[6] Andrew Zisserman Karen Simonyan, “Two-stream convolution networks for action recognition in videos,” in *https://papers.nips.cc/paper/5353-two-stream-convolutional-networks-for-action-recognition-in-videos.pdf*.

[7] Lorenzo Torrensansi Jamie Ray Yann LeCunn Monohar Paluri Facebook Research Dartmount College Du Tran, Heng Wang, “A closer look at spatiotemporal convolution for action recognition,” in *CVPR*, 2018.

[8] X.Yu V.Khilnani T.Poggio A.Steele H.Jhuang, E. Garrote and T.Seere, “Automated home-cage behavior phenotyping of mice,” in *Nature communications*, 2010.

[9] “Optic flow :https://opencv-python-tutroals.readthedocs.io/en/latest/py$_tutorials/py_video/py_lucas_kanade/py$

[10] ,” .