A Survey on Video Surveillance for Object Detection and Classification

Krishna kumar Hegde
Computer Science Department
PES University
krishnash.hulimane@gmail.com

Karthik R
Computer Science Department
PES University
kar97r@gmail.com

Guide
Dr D Uma
Professor, PES University
umaprabha@pes.edu

Abstract **.**Video surveillance in dynamic scenes, particularly for people and vehicles, is at present a standout amongst the most dynamic research subjects in Computer vision using various methods such as CNN,RNN and statistical analysis. It has a large range of promising applications, identification of a human at a distance, congestion analysis and crowd flux statistics , interactive surveillance using multiple cameras, detection of anomalous behaviors etc. This paper surveys several approach for video surveillance and shows pros and cons of each approach used to address the following problem. In general video surveillance contains the following stages: detection of motion, modeling of environments , classification of moving objects and scene, tracking those objects, understanding, measuring and description of behaviors of motion, human recognition, and merging of data from multiple cameras. We review recent achievements and popular approaches of all these stages in this paper. Finally, we conclude possible future work and all the results of different approaches, a combination of motion analysis, CNN efficient calculation, RNN technique for action classification, other methods for surveillance videos, and staic surveillance.

I. INTRODUCTION

Detection of different actions in the videos is very tough task and has got tremendous notice in research area. In the field of video processing video indexing becomes one of the big tasks. So many works address the extraction of feature from videos to explain their actual topic. So in many applications “Event-based” and “Action-based” classification approaches are preferred compared to low level approaches this is shown using recurrent neural networks with long term dependencies called long short term memories.

Video surveillance try to detect certain events and motions. The aim is to find the optimal and accurate method for surveillance. The techniques are as follows, activity detection using Frame differencing, activity recognition and characteristic correlation algorithms permits awareness of exceptional interaction sample based on easy stats computed on tracked among a certain amount of group of human beings trajectories and built up a format that recognizes special cooperation sketch among a gathering of individuals by means of distinguishing special marks trajectories and built up a format that recognizes special cooperation sketch among a gathering of individuals by means of distinguishing special marks. They have exhibited their strategy utilizing certifiable video information to distinguish and perceive practices in parking area setting and could track conceivably suspicious , perilous ,stalking practices which can help security observation. Novel model for human recognition, human following and activities progressively and does not make any suppositions about conditions utilizing three CNNs cooperate, layers are consolidated to speak to numerous sorts of activity Temporal pictures are utilized in three layers of various leveled activity structure.

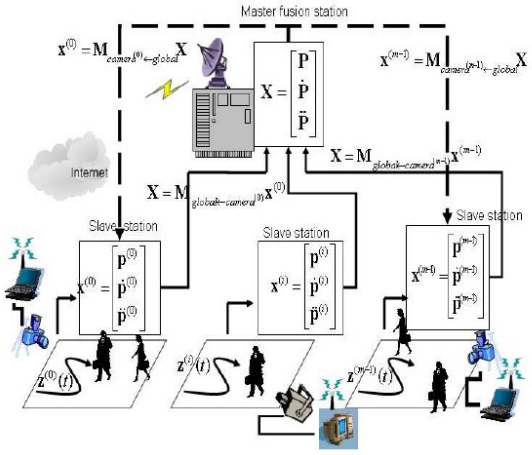
Large scale video classification with convolution neural network which is deep learning branch for classification of images is inspired from presence of large amount of images and videos on internet. Paper consist of video classification of ‘Sports-1M dataset of 1Million YouTube vide- os belonging to 487 classes are made. Based on the training of the model videos are classified into different classes. The technique used is convolution neural network which is best suitable for video classification. To increase the run time performance of CNN engineering is altered into two separate floods of handling those are a high-goals fovea stream and setting stream that learns includes on low-goals outlines

The temporal component of the video can provide much more information about the action compared to still frames of the, as many actions are depend on many frames in the time. Moreover, video gives natural data augmentation for individual image classification.

1.ACTIVITY DETECTION AND TRACKING USING MOTION DETECTION ALGORITHMS

Open air reconnaissance all things considered falls in the far-field situation, which is additionally accepted for this exploration. At the point when individuals in the scene are adequately far away, we can roughly check them as 'blob' and utilize single-protest state to quantify direction, to at present continuous, control and flop over component over low-level frame differencing and relationship based following to manage commotion, scene mess, brief times of nonattendance and converging of outlines, and extensive stretches of impediment of exercises from a camera's field of view. Definition depends on the incredible speculation

and-confirmation worldview, which has been on the other hand dedicated as particle filtering ,Monte Carlo filtering , genetic algorithm, condensation and other similar tecnologies.



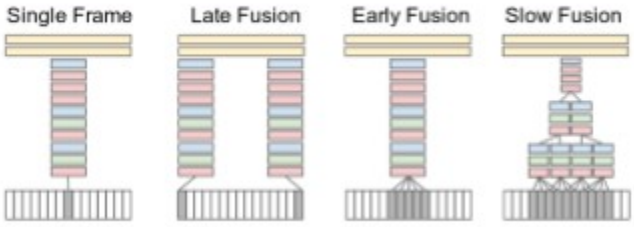
2.ACTIVITY REPRESENTATION AND RECOGNITION USING STATISTICAL MODEL:

Noteworthy measure of research has been done in basic portrayal and acknowledgment of human exercises utilizing Markov chains. Concealed Markov models and coupled shrouded Markov models yet here they have perceived individual and gathering exercises and association dependent on measurable properties registered utilizing recouped directions without parsing against pre-set up morkov display. They have concentrated on following, after and picking up conduct and stalking conduct dependent on the specific attributes for instance following conduct is described by relatively consistent development and almost zero relative velocity.

II. HUMAN ACTION RECOGNITION USING CNN.

The structure incorporate three layers Action , Motion layer and posture layer, Each layer includes diverse class The three layers together convey an entire arrangement of data for human actions.The approach portrayed in this paper performs movement location utilizing Guassian Mixture(GMM) The framework recognizes people in the movement area utilizing a Histogram of Gradient(HOG) To expand speed racking system is utilized in created calculation, distinguished in past however lost in current edge are identified utilizing Kalman filter.

4.LARGE SCALE VIDEO CLASSIFICATION USING CNN



TIME INFORMATION FUSION IN CNNS

Here first Single-frame CNN is discussed and then it's extensions in time according to different types of fusion is discussed. Number of ways are there for fusing information across temporal domain. The converging of data should be possible by changing the underlying layer CNN channel to stretch out in time, or it tends to be accomplished late by putting two individual single-frame network some separation in time separated and combining their yields later in the preparing.

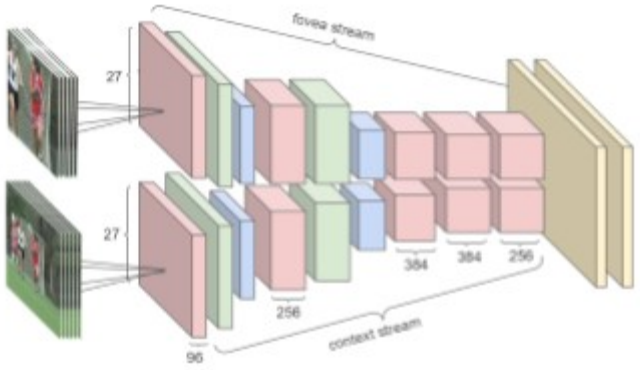
Single-frame Each frame is taken at once and contribution of that frame to include feature is talked about. In design C(f,n,s) means a layer with 'f' filters of measurement ' n*n', with stride 's' applied to input

Early Fusion. The Early Fusion system joins data over a whole time window instantly on the pixel level.By expanding them to be of size 11×11×3×T pixels this should be possible with the assistance of filters of the early CNN layer in the single-frame architecture.

Late Fusion. Late Fusion places 2 different single frame at 15 frames apart and then combine the information from those at the end of each layer

Slow Fusion. The Slow Fusion architecture is a reasonable blend between the two approaches. Higher layers gain admittance to great measure of vast element data in both spatial and worldly measurements filters of the early CNN layer in the single-outline design by gradually consolidates data over the system to such

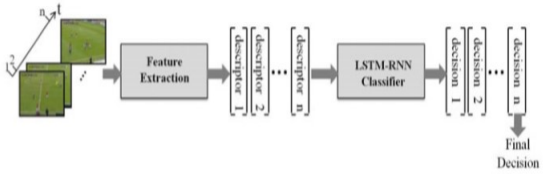
Fovea and context streams (figure *) .Fovea stream CNN architecture works on low resolution frames where as context stream architecture works on central part high resolution frames. It increases 2-4x runtime execution of the system because of the lower dimensionality input, while holding classification accuracy.



figure(*)

5.ACTION CLASSIFICATION BASED ON LSTM-RNN APPROACH

In this paper they are classifying the football video sequences which have descriptors (descriptor per image) corresponding to the set of features. For the proper classification the presence of those features are mandatory. By checking continuous descriptors in temporal manner the LSTM-RNN is trained to make continuous decisions according to descriptors.. accomplish that objective, collection of proper descriptors are given to the system of neurons in a fashion of a descriptor at any given moment, based on the storage of several individual results it generates a final decision as shown in below figure.



ACTION CLASSIFICATION USING LSTM-RNN

RNNs are a specific domains of ANNs which can produces output by keeping in mind previous state inputs. While handling the long term dependencies , their short-term memory causes problems To get the answer of RNN's short term dependencies Schmidhuber introduced recurrent architecture The LSTM.

BOW

Bag of words (BoW) is classical approach for image processing in which numerical vector is generated from the image and based on that feature is generated and finally these features set forms Bag of words (BoW). In this approach every image is represented by an histogram of visual words which are based on the set of local features calculated from numerical vectors of the image.

6.TWO-STREAM ARCHITECTURE FOR VIDEO RECOGNITION

Any video can be devided into spatial and fleeting segments. Spatial part is the static casing containing the data about items and the scene. The fleeting part, delineates the relative movement between the question and the scene dependent on the development of the eyewitness

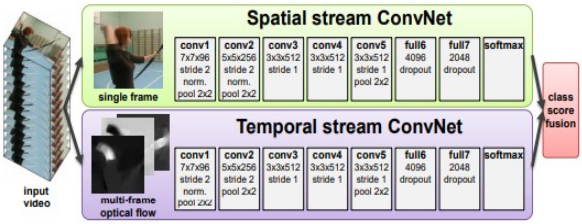


Figure 1: **Two-stream architecture for video classification.**

Spatial stream ConvNet will work on still frames to get the features of action recognition. Because few actions are greatly connected with specific frames the static appearance by itself is a helpful. Actually action classification base on static frames is also promising method for classification

Temporal stream ConvNet architecture, takes optical flows displacements as input. Optical flow is the method which is based on relative motion between the objects

RESULTS OF LARGE SCALE VIDEO CLASSIFICATION

Sports class	Δ AP	Δ AP	Sports class
Juggling Club	0.12	-0.07	Short Track Motor Racing
Pole Climbing	0.10	-0.07	Road Racing
Mountain Unicycling	0.08	-0.07	Jeet Kune Do
Tricking	0.07	-0.06	Paintball
Footbag	0.07	-0.06	Freeride
Skipping Rope	0.06	-0.06	Cricket
Rope Climbing	0.06	-0.06	Wrestling
Slacklining	0.05	-0.06	Modern Pentathlon
Tee Ball	0.05	-0.06	Krav Maga
Sheepdog Trial	0.05	-0.05	Rally Cross

As calculated by difference in per-class average precision.
In left Slow fusion is greater than single frame but in the right opposite is true.

Model	Clip Hit@1	Video Hit@1	Video Hit@5
Feature Histograms + Neural Net	-	55.3	-
Single-Frame	41.1	59.3	77.7
Single-Frame + Multires	**42.4**	**60.0**	**78.5**
Single-Frame Fovea Only	30.0	49.9	72.8
Single-Frame Context Only	38.1	56.0	77.2
Early Fusion	38.9	57.7	76.8
Late Fusion	40.7	59.3	78.7
Slow Fusion	**41.9**	**60.9**	**80.2**
CNN Average (Single+Early+Late+Slow)	41.4	63.9	82.4

Table 1: Results on the 200,000 videos of the Sports-1M test set. Hit@k values indicate the fraction of test samples that contained at least one of the ground truth labels in the top k predictions.

ACTION CLASSIFICATION BASED ON LSTM-RNN

Football actions classification based on RNNs is higly efficient .Models on MICC-Soccer-Actions-4 indicates that LSTM-RNN is better compared to SVMs and K-NN approach. Long terrn dependencies can be addressed using. LSTM-RNNs

Methods	Classification rate
BoW + k-NN [3]	52,75 %
BoW + SVM [3]	73,25 %
BoW + LSTM-RNN	76 %
Dominant motion + LSTM-RNN	77 %
BoW + dominant motion + LSTM-RNN	92 %

MULTI-TASK LEARNING OF TWO STREAM CONVOLUTIONAL NETWORKS

Training of the network is done with the help of UCF-101 and HMDB-51 datasets which has 9.5K and 3.7K videos respectively. Overfitting is reduced by combinig these datasets.

Training setting	Accuracy
Training on HMDB-51 without additional data	46.6%
Fine-tuning a ConvNet, pre-trained on UCF-101	49.0%
Training on HMDB-51 with classes added from UCF-101	52.8%
Multi-task learning on HMDB-51 and UCF-101	55.4%

The algorithm can automatically identify and track movement occasions and characterize them into being and conceivably exact classes. This capacity can be utilized in parking areas, walk lanes, subways and wherever there is no dense and active public. The model is tested with real time data captured from parking lot where people performed some actions and recognition rate was around 100% for this action

CONCLUSION

In large scale video classification using CNN has the accuracy of 82.4 % which is calculated by average CNN, the prediction of particular class is actually is the original class for 82.4 test samples.

In action classification based on LSTM-RNN the BoW +Dominant motion +LSTM-RNN method has classification rate of 92%.In two stream convolutional network multitask learning on HMD-51 and UCF-101 method has 55.4 % accuracy.

Paper Title with Methods	Accuracy Or Classification Rate(%)
Large scale video classification using CNN CNN Average(Single +early +late +slow)	ClipHit@1 VideoHit@1 VideoHit@5 41.4 63.9 82.4
Action classification based on LSTM-RNN BoW +Dominant motion +LSTM-RNN	92
Two stream convolutional network Multitask learning on HMD-51 and UCF-101	55.4
Human Activity Detection and Recognition for video Surveillance Monte carlo filtering+Genetic algorithm+Condensation+Relative motion & velovity	*Behaviors* | *# of Instance* | *F* | *FG* | *S* *F* | 30 | 30 | 0 | 0 *FG* | 28 | 1 | 25 | 2 *S* | 26 | 3 | 2 | 21
Real Time Human Action Recognition Using CNN over Temporal images for static video surveillance CNN{3}+Temporal images	precisions of the posture, motion, and action layers are 97.77 %, 85.99 %, and 71.29 %

Human activity detection based on statistical trajectory, using Monte carlo filtering, relative motion and velocity using Bayesian Estimator attains almost 100% accuracy for following behavior, 80% for the stalking and gaining.The confusion matrix for the following method has been shown in the table above.

Human action recognition using temporal images or motion history images which are very much significant in the results of this method trained on ICVL action dataset which gives great precision of the posture layer.As this reduces in thier superclass i.e motion and action.

So with the help of above table we can easily visualize that for video surveillance the method of LSTM-RNN works efficiently compared to other techniques

CNNs are the best techniques for image classification compared to other techniques such as feature based techniques. The difference between different CNN architectures are insignificantly small compared to feature based technique. In qualitative analysis also their model works really good with only small amount of errors and continuous activity acknowledgment method that does not utilize any presumptions about the conditions Easy to add new actions

In deep video classification model which consists of temporal and spatial model, temporal model with optical flow is one of the best techniques.

The LSTM-RNNs are better compared to other techniques such as K-NN based and SVM. LSTM-RNNs are able to learn to classify long term dependencie

FUTURE WORK

CNNs are the best techniques for image classification compared to other techniques such as feature based techniques. One can increase the accuracy with increasing

the hidden layers. With more training data also accuracy will increase

In current approach they have only used LSTM-RNNs which is basic approach for finding long term dependencies,

More robost approach such as GRU_RNN can be used for higher performance. This approach is better for finding long term dependencies with more accuracy

REFERENCES

[1] Andrej Karpathy, George Toderici, Sanketh Shetty,Thomas Leung,Rahul Sukthankar,Li Fei-Fei,2Computer Science Department, Stanford University "Large-scale Video Classification with Convolutional Neural Networks"

[2] Moez Baccouche, Franck Mamalet,"Action Classification in Soccer Videos with Long Short-

[3] Karen Simonyan, Andrew Zisserman "Two-Stream Convolutional Networks for Action Recognition in Videos

[4] Yuan-Fang Wang,Wei Nu,Jiao Long,Dan Han "Human Activity Detection and Recognition for Video Surveillance"

[5] Trung Dung Do, Chengbin Jin, Shengzhe Li , Hakil Kim"Real-Time Human Action Recognition Using CNN Over Temporal Images for Static Video Surveillance Cameras"

[6] Weiming Hu, Tieniu Tan, Fellow, IEEE, Liang Wang "A Survey on Visual Surveillance of Object Motion and Behaviors"