The Samaritan: Surveillance AI

Jayesh Gadewar,University of Mumbai (jayesh6399@gmail.com)

Abstract— The latest advances in hardware technology and state of the art of computer vision and artificial intelligence research can be employed to develop an autonomous threat detection system that accurately assesses the threat and provides relevant support. This paper suggests a multi-module state of the art surveillance system. The proposed system is intelligent and is capable of event discrimination by developing a systematic approach by solving the problem with its visual and audio LSTM modules. This system promotes a decentralized intelligence to evaluate the dynamic scene to account for all the factors which can influence activity and thus, can deliver a robust and concrete prediction with an optimal interpretation of the scene.

Index Terms— Artificial intelligence , Action Detection, Computer Vision, Surveillance,Emotion,Speech

1 INTRODUCTION

We are surrounded by cameras and to our surprise always under surveillance. Video surveillance is important for security. Up until now, we are just storing this surveillance as data. And as quoted popularly “Data is the new oil.'' We can use this Data to our advantage. Around the globe countless criminal activities occur on streets. Furthermore, law enforcement services are not always alerted in time when threats/crime occurs. With breakthroughs in Image recognition classifiers which have improved real-time performances significantly. They can detect anywhere from 5 to 100 or more objects in a single frame. However, constrained by its hardware and processing power and training data. But, Surveillance needs more semantic detection to understand a dynamic scene. The proposed system Samaritan takes this one step further and looks at this problem as an audio and visual problem. Often every intent can be mapped to a particular set of activities which are followed in order and thus we can predict the threat beforehand. This paper suggests using an LSTM cell for its video activity predictor and speech recognition. Furthermore, the system also uses Emotion recognition to perceive the dynamic scene better with more information and make a learned decision .This novel methodology proposes a new approach and a deep semantic analysis.

2 RELATED WORK

Human activity Detection is a wide problem and researchers have successfully divided this into two parts. Human activity Classification[11] which deals with classifying the action after it takes place and human activity prediction[1],[2] which actually predicts the action before taking place. human activity prediction by utilizing local Spatio-temporal [4], [6],[10],[11].features is thoroughly researched .This approach mainly focuses on the complete execution of a single activity and recognizes them based on video segments. But they are less effective in predicting and classifying complex interactions which may or may not involve more than one body parts and objects which is an ambiguous problem.

On the other hand, an innovative methodology like dynamic bag-of-words was developed, which considers the sequential nature of human activities while maintaining the advantages of the bag-of-words to handle noisy observations[11] however still uses spatiotemporal features to predict the action and needs at least 60 % of the action to be completed.

There are also notable State-of-the-art strategies which are employed for object detection[12] and Pose estimation[13]. The whole video is looked at as a pictorial structure method and uses a sliding window which is one of the most successful techniques to determine labels for objects. They introduce “atomic poses” to tackle the problem and recognize Human object interactions in still images. However, this approach is solely dependent on correlations between the object and the “atomic pose”.

Another notable line of work is a convex learning formulation based on the structured SVM to enforce their label consistency on videos containing Temporally incomplete action executions.

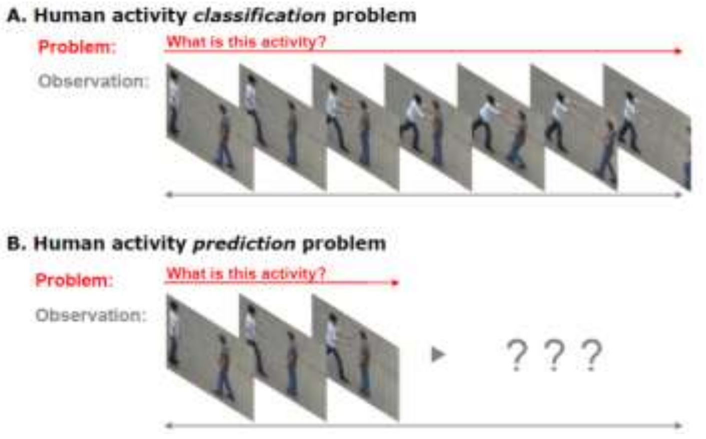

FIG 1. Differentiation between activity prediction and

classification.

3 SAMARITAN

The Samaritan System is modeled to rightly predict activities which are termed as “Threat”. This novel system exploits all types of data at its disposal and creates a complex correlation map. This mapping technique serves two purposes.Firstly, It helps samaritain to retain memory of the actions.Secondly,if the action predicted is wrong then samaritan can refer to erroneous maps and using regression method reiterate the model until a developer perfection is achieved.

4 SAMARITAN ARCHITECTURE

The superficial architect shown in the figure explains the outline of the system. This method can be considered a multi-module informative network. The core function of the system is to predict a threat and not the completion of the activity. For better implementation and resource management, similar functions are performed within each module. The output of each module gives an individual feature assessment which is concatenated with output of other modules and gives the overall “Threat index”.This Threat index is the value of how “dangerous” or lucid the scene is. The main advantage of this system is that it accounts for various variables to predict the action successfully.

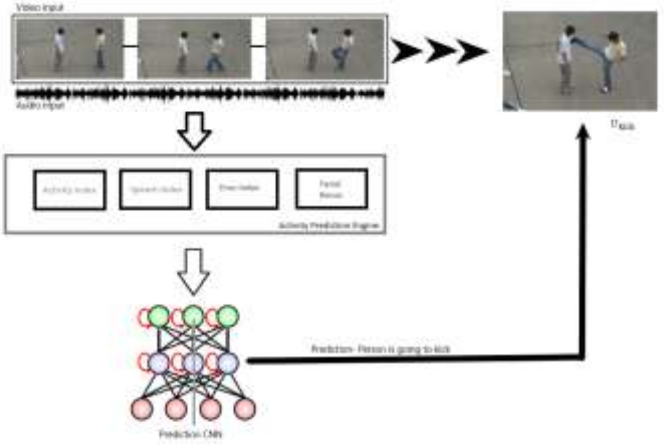

Figure 2 - Samaritan Architure

4.1 Action Prediction module

This is the core module of the system. This system uses a mem-LSTM model with a convolution neural network (CNN) to predict with partial real-time observation. The Surveillance feed is divided into a flow stream and RGB stream. Both have similar LSTM architectures however,the flow stream uses LSTM without a residual system. the memory module integrated with LSTM and is used to remember early observations. the memory operations as computed by lei et al [14] are divided into three operations mainly query, update, lose. their memory is defined as

M = (K, V,A)

Where m stores the samples dimensional k, and V = {v} = {(y, z)} indicates their corresponding action classes y and progress levels z.

Memory query is used to compute similarity using similarity metrics i.e dot product and gsauusuin kernel..

$$d=q.K[i],d=\exp(\frac{-\|q-k[i]\|^2}{2\partial^2}).$$

Memory update is performed when query is new corresponding to v

$$q:n=NN(q,M).$$

Memory loss is performed when the computed dot product is similar to the “memorized” samples. The purpose of this function is to minimize the similarity to incorrect keys.

$$max([q\cdot K[nb]+q\cdot K[nc]-q\cdot K[na]+\xi],0),$$

The memory module is integrated with bi-directional LSTM. It's widely known that LSTM is unable to make use of future references. This paper suggests a method to calculate two layers distinctively with their own hidden features, independently .The outputs are then summed from t=1 to t=T moment. A RES-18 model; is used to train the flow images, Different from [14]. The output is summed and passed through a “Threat Function“ which correlates with sigmoid values of summed outputs of RGB and Flow layers. The output of this module “activity index” is used with outputs of other modules and fed to a CNN for further analysis.

4.2 Emotion Detection module

The visual information of the person is reduced into a 4-dimensional feature vector. This vector is passed through a facial classifier to predict facial expression. the classifier outputs quantified facial expression. The facial data is normalized with respect to (1)identifying nose marker and making it the coordinate center of the masked frame(2)define rigid markers to and local coordinates of the face(3) identify each local coordinates and create a block planes of the forehead, eyebrow, low eye.right cheek, left cheek .using Principal component analysis the data is reduced to 10-dimensional vector for each identified block plane. The 10-dimensional frame is then again classified using k-nearest neighbor as different emotions belong to different clusters. Apart from this, a separate SVC classifier is used to give a better emotional disparity

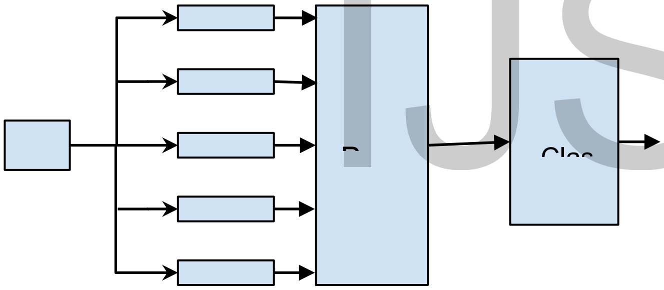

Figure 3 - Emotion detection architecture

The visual information of the person is reduced into a 4-dimensional feature vector. This vector is passed through a facial classifier to predict facial expression. the classifier outputs quantified facial expression. The facial data is normalized with respect to (1)identifying nose marker and making it the coordinate center of the masked frame(2)define rigid markers to and local coordinates of the face(3) identify each local coordinates and create a block planes of the forehead, eyebrow, low eye.right cheek, left cheek .using Principal component analysis the data is reduced to 10-dimensional vector for each identified block plane. The 10-dimensional frame is then again classified using k-nearest neighbor as different emotions belong to different clusters. Apart from this, a separate SVC classifier is used to give a better emotional disparity

4.3 Speech module

This system uses prosodic information as acoustic features which also incorporates the duration of voiced and unvoiced segments. These features work closely with the Emotion module and together they estimate the emotion of the person in the dynamic scene. Firstly single level classifiers are used to classify the speech elements. The audio and emotion features a then fused together to reduce the modalities of the system. Prosodic features such as pitch and intensity are used as features for the classifier. The standard deviation, the minimum values, medians of the pitch were computed using Praat speech processing software. Using a sequential backward selection process an 11-dimensional feature vector for each block of audio is generated. This vector is then used as input for the classifier. The classifier outputs a correlated threat index which is used with the emotion module threat index.

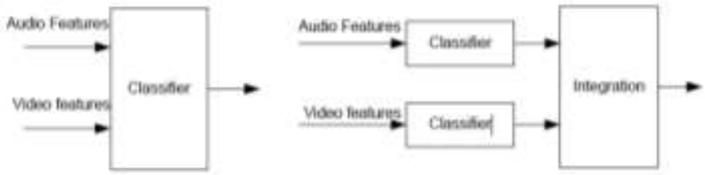

Figure 4 - Speech detection architecture

4.3 Facial Recognition module

An object detection model is trained on a person's facial data points.CNN is employed to extract features from the scene and classify objects and people in the scene. Moreover, this system also accounts for objects like guns, knives and other harmful objects along with the person in the scene. Upon successful detection, the records are fetched from a pre-built citizen database.

4.5 Integration

A CNN model is trained with different threat index values and interpretations of the scene to further assess the scene which the system encounters. The model is trained on different wide-ranging samples to accommodate every possible scenario. even if a variant scenario takes place due to its multimodal system the agent can successfully predict the activity. The integration phase evaluates the threat index of each module and finds a correlation between then to classify the threat as dangerous.

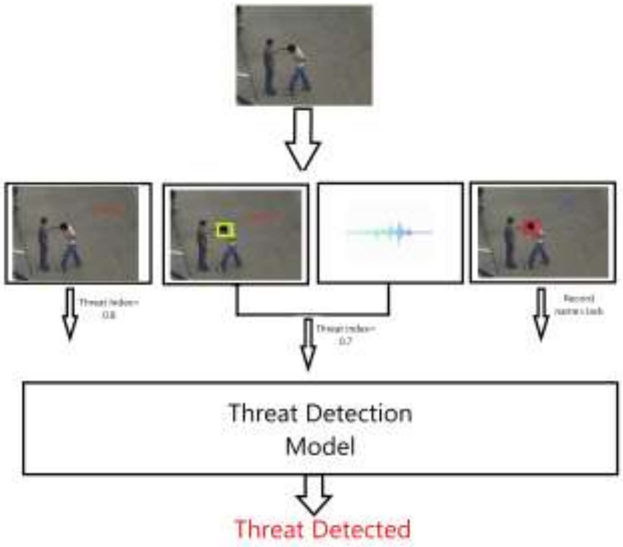

Figure 5 - integrated interpretation of the dynamic scene

5 EXPERIMENTS

Experiments were done on a custom dataset that contained "threat by a kick "," pulling out a gun'', "hug'', ``suspicious activity``.Different emotions were tested under each threat category in the sample. Table 1 shows the performance of the system and respective threat indexes achieved based on a controlled sample data. The table also buttresses to 75% average accuracy using samaritan methodology

	Anger	Happiness	Sadness	Neutral	Overall	Threat index	Threat assessment
Kicking	0.81	0.23	0.57	0.33	0.87	0.77	YES
Hugging	0.1	0.85	0.77	0.56	0.67	0.33	NO
Pulling out a gun	0.86	0.15	0.49	0.56	0.73	0.89	YES

Table 1- Integrated Threat assessment

Furthermore, testing on UT-dataset2 with custom audio samples produced satisfactory results.Upon testing on UT2 dataset bimodal module (speech and emotion) was integrated and a confusion matrix was observed with product combining rule.

	Anger	Sadness	Happiness	Neutral
Anger	0.84	0.08	0.01	0.08
Happiness	0	0.7	0.01	0.10
Neutral	0	0	0.85	0.15
Sadness	0	0.1	0.23	0.91

Table 2- Emotional assessment values

6 CONCLUSION

In order to understand a dynamic scene, the samaritan methodology trumps unimodal systems with significant differences. Although there are many variables that are still not discernible due to current technical limitations and hardware limitations. Furthermore, more known scenes could be fed to the network and an increased activity database can be implemented. The samaritan successfully differentiated the scene objects and created a correlation between objects, pose, speech Observed results prove that an intelligent surveillance system can be deployed and which can be improved in the coming future.

7 REFERENCES

[1] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, p. 1627, 2010.
[2] P. Viola and M. J. Jones, "Robust real-time object detection," Int. J. of Comput. Vision, vol. 57, no. 2, p. 87, 2001.
[3] K. K. Sung and T. Poggio, "Example-based learning for view-based human face detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 1, pp. 39–51, 2002.
[4] E. Ohn-Bar and M. M. Trivedi, "To boost or not to boost? on the limits of boosted trees for object detection," in ICPR, 2016.
[5] C. Wojek, P. Dollar, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 4, p. 743, 2012.
[6] H. Kobatake and Y. Yoshinaga, "Detection of spicules on mammogram based on skeleton analysis." IEEE Trans. Med. Imag., vol. 15, no. 3, pp. 235–245, 1996.
[7] X. Bai, X. Wang, L. J. Latecki, W. Liu, and Z. Tu, "Active skeleton for non-rigid object detection," in ICCV, 2010.
[8] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in ACM MM, 2014.
[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in

NIPS, 2012.

[10] M. Andriluka, L. Pishchulin, P. V. Gehler, and B. Schiele, “2d human pose estimation: New benchmark and state of the art analysis,” in CVPR, 2014.

[11] M. S. Ryoo “Human Activity Prediction: Early Recognition of Ongoing Activities from Streaming Videos” in IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, Nov. 2011.

[12]Joseph Redmon,Ali Farhadi , “YOLO9000: Better, Faster, Stronger” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7263-7271

[13] Alexander Toshev, Christian Szegedy ”DeepPose: Human Pose Estimation via Deep Neural Networks” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1653-1660

[14] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, Xiaohui Xie “Co-Occurrence Feature Learning for Skeleton Based Action Recognition Using Regularized Deep LSTM Networks” in Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)