Depth and Movement Data Analysis for Fight Detection

Amol S Patwardhan[1]

[1]*Senior Researcher, VIT, University of Mumbai, 400037, India*

---***---

Abstract - *This research paper investigates human action recognition using color and depth data (RGB-D) and motion analysis of image frames. It proposes novel recognition methods of aggressive actions such as throwing, kicking, punching, and threatening by using 3D depth infrared based sensors. The color channel data was captured from video recordings of actions and infrared sensing device. 23 participants were enacted aggressive actions from a list of actions in a script. The SVM classifiers were trained using the features extracted from the sequence of frames and tested against violent actions simulated on the basis of real life action, violence and movie fight scenes. The results indicated that the performance for aggressive actions containing single individuals showed higher accuracy but the system performed at low recognition rate for crowded and group activity detection in terms of accuracy*.

Key Words: Emotion Recognition, Crowd Activity, Group Activity, Edge detection, Audi-Video data, 3D sensor, Rule, Kinect, Affect Recognition.

1. INTRODUCTION

Emotion recognition by computers has several applications in marketing and user profile specific advertisements. Many people watch live sports telecasts in groups. The football games in United States and international soccer games or basketball games are viewed by large crowds and groups of people gathering in a public place, sport bars and a friend's house. The various events during the games invoke several emotions among the supporters of each team. The emotions range from disgust, disappointment or joy depending on whether the supporter's team won or lost. There are incidents of controversial calls, surprise decisions and anguish when a referee makes a bad call or a favorite player gets injured and makes an excellent shot, which gets revoked. Multimodal emotion recognition has been studied using RGB-D data and audio-visual data. These studies have examined the videos mostly in controlled conditions using individual enactments or spontaneous emotional episodes. In this research, the focus is on spontaneous group emotion detection under indoor lighting. Researchers [1] have examined the side effects of emotional thinking on memory and judgement. Emotion representation has been discussed in the field of psychology [2] using basic emotions such as happy, sad, angry, fear, surprise and disgust. In addition to video based emotion recognition, researchers [3] have also used vocal features to examine expression of emotions. A survey [4] on emotion detection and modelling using speech, audio and vocal input data has evaluated significance of such modalities in recognition accuracy. Application of emotion estimation in educational settings has been studied by researchers [5]. Studies [6] have used biosensors to assist in emotion detection and analyze the psychological and physiological effects of emotions in humans. Researchers [7] have used face, voice and body data to evaluate emotion modelling. A study [8] investigated the co-relation between various input channels in estimation accuracy using neural networks. Researchers [9] used dynamic Bayesian networks for monitoring crowd activity. Researchers [10, [11], [12] have studied intelligent surveillance systems using biosensors and bio-inspired devices. Studies [13] have evaluated the connection between body expressions and emotional states. Some studies [14], [15] have focused on emotion maps and cognitive states and their relation with emotion modelling and expression. Studies [16], [17], [18], [19] have investigated dimensionality issues in emotion modelling, effect of cross-cultural influence of emotion, pedestrian behavior and crowd simulation for emotion estimation. The order of reference in the running text should match with the list of references at the end of the paper.

Research [20], [21] has been done on human gait, pedestrian dynamics especially under influence of alcohol. Studies [22], [23] have focused on specific emotion such as detection of fear, implementation strategies for automatic systems. Researchers have studied [24], [25] the view-invariant emotion detection for specific human behavior such as crowd analysis, fatigue and tiredness and sleepiness prediction. In addition to audio-visual modalities studies [26], [27], [28] have focused on text based sentiment analysis. Several supervised learning techniques (neural network, DBN, HMM) [29], [30], [31], [32], [33] for emotion recognition in various settings (closed spaces, indoor and outdoor) have been used to determine the accuracy of each method. Studies [34], [35], [36], [37], [38] on software implementation of automated continuous computer vision based multimodal emotion recognition techniques have been studied in detail. These studies have implemented novel methods to solve problem of modality fusion using hybrid methods and also provided architectural methods for real time detection. Researchers [39] have developed novel algorithms and behavioral rule based features for multimodal emotion recognition using continuous data and supervised learning.

2. METHOD

Four groups of five people for 20 individuals participated in the study. The individuals were all dressed in casual attire and were aged in the range of 22 to 45. Eighteen participants were male and two were females. The groups gathered for different games of basketball finals. Three different cameras were used to record their reactions for 5 min at the beginning of each half and 5 min before the end of each half. The games usually get interesting at the beginning and towards the end. Tight games invoke more emotions compared to one-sided matches. The non-intrusive way of data capture allowed the subjects to watch the event and express their reactions without consciousness about being recorded on camera.

After the data was captured, the canny edge detection was applied on every third frame from the video sequence. This caused down sampling of a 24 frame per second to an eight frames per second sequence. For each image frame, the edge detection filter was applied. Then the frame was divided into 20 x 20 meshes and the intersection of the grid lines with the edges was considered as feature co-ordinates. For a consistent feature vector length, each line was further divided into 5 divisions and the features were counted as one for an intersection and zero for no intersection. Thus, 20 x 5 features for vertical lines and 20 x 5 features for horizontal lines on the mesh, for a total of 200 feature co-ordinates were obtained to form the feature vector. Additionally, the temporal features were also tracked.

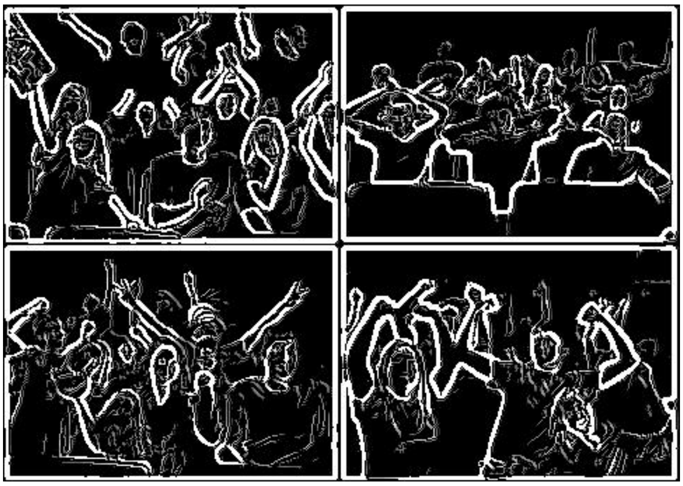

Fig -1: Step 1 with edge detection applied to the frames

In the pre-processing step shown in Fig. 1. edge detection is applied to extract the series of high intensity change edges. These edges assisted in extracting further set of more discriminating feature points. This strategy was useful in extracting potential edges without depending on specific human behavioural actions, gestures and simply relied on view-independent image processing technique. Additionally, the occlusion caused by multiple people moving in the scene did not affect the process.

As the next step in the feature extraction process, the optimized grid was applied and super-imposed on each frame to find the intersections between the grid lines and the detected edges. Each intersection point was used as the feature for the frame. Next, the movement of each of these features was measured across the eight frames to find the temporal, kinetic and motion pattern of the feature. This allowed the method to rely on a limited yet discriminating set of features instead of tracking individual human activities. The method also allowed overcoming the limitation of wearing tracking devices to detect each human in the scene. It also allowed view-independent and occlusion resistance mechanism and purely relied on the visual layout of the scene based on the available data from the extracted features across various video frames and series of images. In addition, to limit the number of intersections between the edges and the lines of the grid only the N equally spaced points were considered. This number N was set to the same threshold as the number of blocks within the grid. For instance, if the grid size was 20 then the spacing was set to 20 as well. As a result, any other intersections were discarded. The settings for grid threshold at 5, 10, and 15 up to 50 showed that this strategy did not affect the overall accuracy results and the N simply contributed in limiting the feature vector size and eliminating the redundant tracked points.

The above figure shows the super-imposed grid and the intersection of the grid lines with the edges extracted from the first pre-processing step. The next figure shows the processing workflow to extract the edges, then the application of grid, mesh and extraction of potential points for tracking. This step is followed by tracking the motion of the points across the series of video frames.

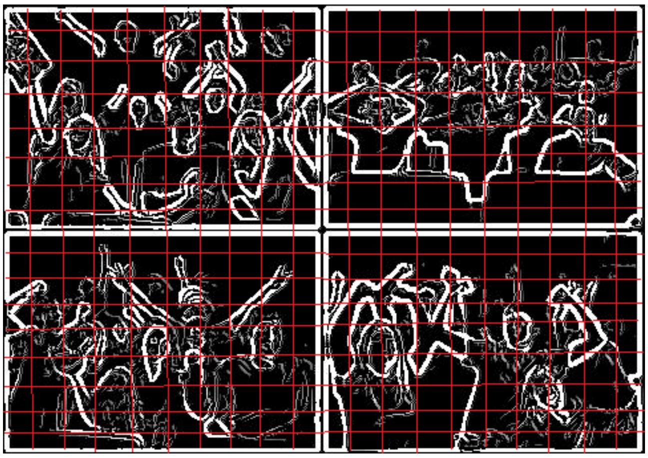

Fig -2: Step 2 with the mesh applied to the edges.

The movement of each co-ordinate was tracked across the eight frames. Thus, the final feature vector consisted of 200 static and 200 velocity values across eight frames for 400 features. The best first search technique was used for feature selection. This resulted in reduced dimensionality of feature vector size and 23 discriminating features were chosen. The sequence of all actions was annotated using three observers to avoid inter-annotator disagreement.

The classes used were happy, angry, surprised, sad, disgust, fear and neutral. After the annotation was done, the feature vector was used to train the classifiers using support vector machine and radial basis function as the kernel function with 0.4 as the slack variable.

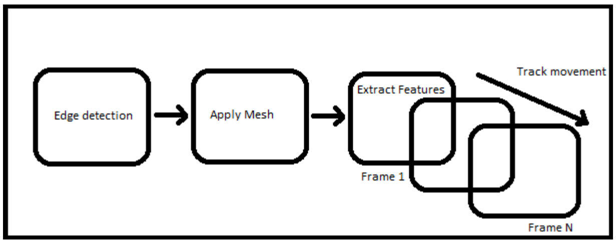

Fig -3: Feature Extraction Process Workflow

The optimized slack variable was calculated using grid forward search method. The data was split into 70% training and 30% test data. The training was performed using 10-fold cross validation.

3. RESULTS

The threshold of 0.2 resulted in better accuracy for anger and neutral emotion. The threshold of 0.4 resulted in best accuracy for most of the emotions except fear and surprise. The thresholds for 0.6 and 0.8 performed poorly. This was because very few edges emerged after applying these thresholds resulting in fewer discriminating features in subsequent processing steps of grid based super-imposed feature extraction.

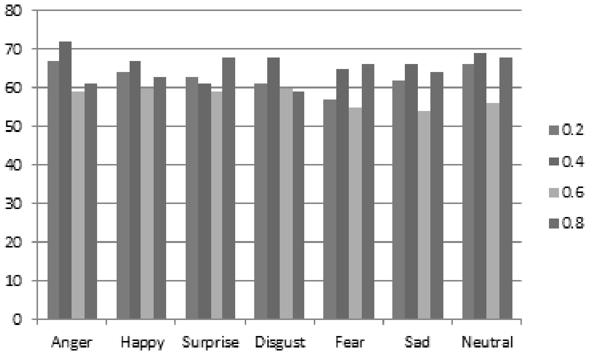

Fig-4: Thresholds for grid size

The grid size of 20 showed a clear improvement in accuracy as compared to the performance from grid size of 10 or 50. Except for surprise and sadness, the grid size of 20 showed the best emotion recognition accuracy. This was because the size was optimal. The size of 10 was too low to provide enough features while 50-grid size provided too many redundant and overlapping features to result in any significant improvement and at times even caused the accuracy to degrade.

The highest recall rate was observed for estimation of happy class label. It was mostly confused with surprise. The next best recall rates were recorded for surprise and disgust. The least accuracy was scene for anger. This was because in some cases the movement of individuals was so high in the scene, that they sometimes even walked out of the viewing frame or collapsed on the couch or hid their faces with objects such as gloves, helmets, caps or pillows. The above results were all recorded for the optimized edge detection thresholds of 0.4 and the optimized grid size threshold of 20 blocks per grid.

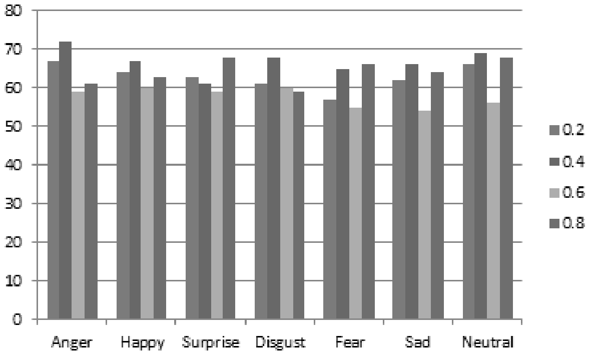

Fig-5: Thresholds for grid size

4. CONCLUSIONS

The threshold for edge detection that yielded the best recognition results was 0.4. The classification results for happiness emotion class label was the highest at 76.6% followed by disgust 72.9% and surprise 72%. The recognition rate for fear was the lowest with 67.2%. The overall accuracy of the group emotion recognition process was 70.9%. The grid size of 20 resulted in the best accuracy for 5 out of 7 emotion classes (including neutral class). This study evaluated a novel technique that implemented image-processing steps to extract the edges and then extract the features for chaotic scenes resulting from expression of emotions in crowded spectator settings. The group of people in the scene resulted in occluded view where many other techniques are not accurate because of the lack of patterns.

As a result, the action based recognition techniques cannot be applied in such scenarios. The techniques mentioned in the paper showed promising results to overcome this limitation of view-dependence and lack of sufficient training data. This paper mostly looked into indoor scenes and a limited set of outdoor spontaneous scenes with crowds of people in the scene reacting emotionally expressive manner for sports events. As a future scope, the study could be extended in outdoor bigger crowd settings and a comparative study could be done between various techniques with the processing steps described in this paper.

REFERENCES

[1] A.Veenendaal, Elliot Daly, Eddie Jones, Zhao Gang, Sumalini Vartak, Rahul S Patwardhan, Drunken Abnormal Human Gait Detection using Sensors, Computer Science and Emerging Research Journal, vol 1, 2013.

[2] A.Veenendaal, Elliot Daly, Eddie Jones, Zhao Gang, Sumalini Vartak, Rahul S Patwardhan, Fear Detection with Background Subtraction from RGB-D data, Computer Science and Emerging Research Journal, vol 1, 2013.

[3] A.Veenendaal, Elliot Daly, Eddie Jones, Zhao Gang, Sumalini Vartak, Rahul S Patwardhan, Code Definition Analysis for Call Graph Generation, Computer Science and Emerging Research Journal, vol 1, 2013.

[4] A.Veenendaal, Elliot Daly, Eddie Jones, Zhao Gang, Sumalini Vartak, Rahul S Patwardhan, Multi-View Point Drowsiness and Fatigue Detection, Computer Science and Emerging Research Journal, vol 2, 2014.

[5] A.Veenendaal, Elliot Daly, Eddie Jones, Zhao Gang, Sumalini Vartak, Rahul S Patwardhan, Group Emotion Detection using Edge Detecttion Mesh Analysis, Computer Science and Emerging Research Journal, vol 2, 2014.

[6] A.Veenendaal, Elliot Daly, Eddie Jones, Zhao Gang, Sumalini Vartak, Rahul S Patwardhan, Polarity Analysis of Restaurant Review Comment Board, Computer Science and Emerging Research Journal, vol 2, 2014.

[7] A.Veenendaal, Elliot Daly, Eddie Jones, Zhao Gang, Sumalini Vartak, Rahul S Patwardhan, Sentiment Analysis in Code Review Comments, Computer Science and Emerging Research Journal, vol 3, 2015.

[8] A.Veenendaal, Elliot Daly, Eddie Jones, Zhao Gang, Sumalini Vartak, Rahul S Patwardhan, Temporal Analysis of News Feed Using Phrase Position, Computer Science and Emerging Research Journal, vol 3, 2015.

[9] A.Veenendaal, Elliot Daly, Eddie Jones, Zhao Gang, Sumalini Vartak, Rahul S Patwardhan, Decision Rule Driven Human Activity Recognition, Computer Science and Emerging Research Journal, vol 3, 2015.

[10] A.Veenendaal, Elliot Daly, Eddie Jones, Zhao Gang, Sumalini Vartak, Rahul S Patwardhan, Depression and Sadness Recognition in Closed Spaces, Computer Science and Emerging Research Journal, vol 4, 2016.

[11] A.Veenendaal, Elliot Daly, Eddie Jones, Zhao Gang, Sumalini Vartak, Rahul S Patwardhan, Dynamic Probabilistic Network Based Human Action Recognition, Computer Science and Emerging Research Journal, vol 4, 2016.

[12] A.Veenendaal, Elliot Daly, Eddie Jones, Zhao Gang, Sumalini Vartak, Rahul S Patwardhan, Fight and Aggression Recognition using Depth and Motion Data, Computer Science and Emerging Research Journal, vol 4, 2016.

[13] A.Veenendaal, Elliot Daly, Eddie Jones, Zhao Gang, Sumalini Vartak, Rahul S Patwardhan, Sensor Tracked Points and HMM Based Classifier for Human Action Recognition, Computer Science and Emerging Research Journal, vol 5, 2016.

[14] A. S. Patwardhan, 2016. “Structured Unit Testable Templated Code for Efficient Code Review Process”, PeerJ Computer Science (in review), 2016.

[15] A. S. Patwardhan, and R. S. Patwardhan, “XML Entity Architecture for Efficient Software Integration”, International Journal for Research in Applied Science and Engineering Technology (IJRASET), vol. 4, no. 6, June 2016.

[16] A. S. Patwardhan and G. M. Knapp, “Affect Intensity Estimation Using Multiple Modalities,” Florida Artificial Intelligence Research Society Conference, May. 2014.

[17] A. S. Patwardhan, R. S. Patwardhan, and S. S. Vartak, “Self-Contained Cross-Cutting Pipeline Software Architecture,” International Research Journal of Engineering and Technology (IRJET), vol. 3, no. 5, May. 2016.

[18] A. S. Patwardhan, “An Architecture for Adaptive Real Time Communication with Embedded Devices,” LSU, 2006.

[19] A. S. Patwardhan and G. M. Knapp, “Multimodal Affect Analysis for Product Feedback Assessment,” IIE Annual Conference. Proceedings. Institute of Industrial Engineers-Publisher, 2013.

[20] A. S. Patwardhan and G. M. Knapp, “Aggressive Action and Anger Detection from Multiple Modalities using Kinect”, submitted to ACM Transactions on Intelligent Systems and Technology (ACM TIST) (in review).

[21] A. S. Patwardhan and G. M. Knapp, “EmoFit: Affect Monitoring System for Sedentary Jobs,” preprint, arXiv.org, 2016.

[22] A. S. Patwardhan, “Embracing Agile methodology during DevOps Developer Internship Program”, IEEE Software (in review), 2016.

[23] A. S. Patwardhan, “Edge Based Grid Super-Imposition for Crowd Emotion Recognition”, International Research Journal of Engineering and Technology (IRJET), May. 2010.

[24] A. S. Patwardhan, “Human Activity Recognition Using Temporal Frame Decision Rule Extraction”, International Research Journal of Engineering and Technology (IRJET), May. 2010.

[25] A. S. Patwardhan, “Low Morale, Depressed and Sad State Recognition in Confined Spaces”, International Research Journal of Engineering and Technology (IRJET), May. 2011.

[26] A. S. Patwardhan, “View Independent Drowsy Behavior and Tiredness Detection”, International Research Journal of Engineering and Technology (IRJET), May. 2011.

[27] A. S. Patwardhan, “Sensor Based Human Gait Recognition for Drunk State”, International Research Journal of Engineering and Technology (IRJET), May. 2012.

[28] A. S. Patwardhan, “Background Removal Using RGB-D data for Fright Recognition”, International Research Journal of Engineering and Technology (IRJET), May. 2012.

[29] A. S. Patwardhan, “Depth and Movement Data Analysis for Fight Detection”, International Research Journal of Engineering and Technology (IRJET), May. 2013.

[30] A. S. Patwardhan, “Human Action Recognition Classification using HMM and Movement Tracking”, International Research Journal of Engineering and Technology (IRJET), May. 2013.

[31] A. S. Patwardhan, “Feedback and Emotion Polarity Extraction from Online Reviewer sites”, International Research Journal of Engineering and Technology (IRJET), May. 2014.

[32] A. S. Patwardhan, “Call Tree Detection Using Source Code Syntax Analysis”, International Research Journal of Engineering and Technology (IRJET), May. 2014.

[33] A. S. Patwardhan, “Walking, Lifting, Standing Activity Recognition using Probabilistic Networks”, International Research Journal of Engineering and Technology (IRJET), May. 2015.

[34] A. S. Patwardhan, “Online News Article Temporal Phrase Extraction for Causal Linking”, International Research

Journal of Engineering and Technology (IRJET), May. 2015.

[35] A. S. Patwardhan, "Online Comment Processing for Sentiment Extraction", International Research Journal of Engineering and Technology (IRJET), May. 2016.

[36] A. S. Patwardhan, "Analysis of Software Delivery Process Shortcomings and Architectural Pitfalls", PeerJ Computer Science (in review), 2016.

[37] A. S. Patwardhan, "Multimodal Affect Recognition using Kinect", ACM TIST (in review), 2016.

[38] A. S. Patwardhan, "Augmenting Supervised Emotion Recognition with Rule-Based Decision Model", IEEE TAC (in review), 2016.

[39] A. S. Patwardhan, Jacob Badeaux, Siavash, G. M. Knapp, "Automated Prediction of Temporal Relations", Technical Report. 2014.