Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 06, June - 2024

Infantza: Computer Vision and Deep Learning Enabled Infant Surveillance System

Anagha Suresh, Jitta Amit Sai, Jeevan Anil, Immadisetty Sai Jayanth, R. Bharathi
PES University, Bengaluru, India

***Abstract*—With the rise in complexity of the job roles of today's parents and their hectic schedules, the need for infants to be observed frequently when left in the care of a caretaker to avoid any kind of injury and to constantly have an eye upon them all day becomes a tedious task. In today's world, small infants and children being subjected to abuse from caretakers and others has become a serious issue, due to which parents are unable to entrust them with the safety of their child. We seek to provide a novel approach for infant monitoring that sends alerts to parents with respect to few cases of prospectively dangerous situations that an infant might be exposed to. Infantza is a framework that has been developed with three phases: a module for detecting infant activity, a module for detecting objects, and a module for alerting parents when their infants are in danger.This would eliminate the need for the parents to constantly watch over the baby because they would only need to do so when an alarm had been given indicating the existence of an odd circumstance.**

***Index Terms*—CNN, Computer Vision, Deep Learning, Flutter, FaceNet , Haar Cascade,Ngrok, YOLOV3.**

I. INTRODUCTION

In today's world, danger lurks around every corner, and concerns with respect to infant safety have become a growing issue. Daily news report states that most of the time infants are physically abused by their own caretakers.Due to anxieties about their children's safety, a large percentage of mothers end up abandoning their career aspirations to care for their infants. An effective alarm system is needed to identify any potentially unsafe situations because parents cannot spend all of their time at their desks checking on their children's well-being.

Numerous existing studies are prevalent in the IoT domain which have a lot of hardware usage and different types of sensors[8],[12],[13].The proposed methodology will have minimal hardware and more software usage, making it more efficient and cost-effective.The limited exploration in the topic of infant surveillance using computer vision and deep learning served as motivation to use such techniques for the infant surveillance system.The framework not only addresses multiple areas of concern with respect to infant safety but also provides alerts to parents when the infant is in dangerous situations.

The paper has been organized in the following manner.The next section deals with the literature survey, the third section is the infant surveillance framework and the fourth section explains the implementation of the framework.The fifth section consists of results and discussions and the sixth section is the error analysis.The seventh section has the conclusion and future work.

II. LITERATURE SURVEY

A deep learning system has been used to recognize the occurrence of crimes captured through CCTV and further to notify police stations as well[1].It also identified the perpetrator and the weapons through image segmentation using YOLO.The data set was prepared using the Ybat annotation tool, and the boxes have been manually annotated to verify accuracy and regularity.The data set is trained using the Darknet framework.The face detection occurs at a rate of 45 frames per second because of the incredibly quick architecture that has been utilised.

Chang and Chen [2] used a deep learning neural network to detect the baby's face and the SSD+ Mobilenet network architecture in Tensorflow for infant vomit detection using public face data set-WIDERFACE.The model had infant face detection, vomit detection, and finally the classification results. After detection of the infant's mouth, the noise is removed using gaussian filtering. The average pixel value of the mouth region and the difference between the previous and subsequent frames are determined.Vomit or a covered mouth is recognised if the value of r (mouth area) is less than 0.5.With this method, the baby's face can be recognised in a range of lighting conditions or complicated backgrounds.

The input video has been used to detect any suspicious activity using YOLOv3[3].The dataset considers three instances of suspicious activity—wallet theft, bag snatching, and lock breaking.The videos are converted to frames at 30 fps. Frames are annotated to focus on the region of interest followed by model training and evaluation.The performance of YOLOv3 is superior to FASTER R-CNN .With an accuracy of roughly 95 percent, the model is capable of detecting each image at an incredibly rapid rate.Only in a controlled environment does the feature extraction technique utilised yield correct findings. There were discrepancies between the test results and the actual data since a limited amount of training data was available.

Multi Task Cascaded Convolutional Networks has been used for facial recognition and detection to identify criminals from a data set of 200 images[4].It is a three-part CNN that can identify facial landmarks like the nose, forehead, and eyes.The normalising technique allows for the identification of the face landmarks. Feature vectors are created by extracting the features from the face. FaceNet is used to recognise and verify the images.The model has achieved a high accuracy in facial classification. An accuracy of 92 percent has been obtained for training and 90 percent for testing.

Ali, Khatun,Turzo,Nakib [5] used neural networks to identify facial emotions from a dataset composed of real-world data of human facial expressions. The input images are preprocessed and the noise is removed. The image has been segmented, and facial features have been extracted. CNN models as well as Keras, Tensorflow, and pretraining principles have been applied. The Viola Jones algorithm has been used to detect the eye and mouth areas, and machine learning, deep learning, and neural network algorithms are used to recognise emotions. The method's advantages include excellent accuracy, which was determined using decision trees, and the detection and classification of seven emotions. To enhance accuracy, a large amount of test data and keywords are required.

Deep learning and IOT edge computing has been used to do real-time facial expression recognition.[6]The face detection model identifies the presence and position of the face.The face segment is cropped, and the 128d face embedding is generated before model classification for the three emotions happy, sad, and sleeping . The Caffe deep learning framework is used to train the deep learning face detector, which is built on the Single Shot Detector framework with a ResNet base network. Another DNN-based model is used to combine the face into a 128-D unit hypersphere, which quantifies the face. The Deep Convolutional Neural Network (DNN) model is the foundation of the DNN model (CNN).Following training, the fine-tuned deep learning model is hardware optimised and deployed for production on a low-cost Jetson Nano embedded device.The deployed deep learning model functions as a web service, with the picture being given to the web server through REST API and the model predicts the image's category .The edge device has memory and computational constraints, therefore the size of the deep learning models must fulfil all constraints in order to function, and insights are only sent to the cloud if the internet is accessible.

Identification of emotions has been done using Deep learning to build an artificial intelligence (AI) system capable of detecting emotion through facial expressions[7].Face detection, feature extraction, and emotion classification are the three primary steps of the approach.The data sets utilised were Japanese Female Face Expression (JAFFE) and Facial Expression Recognition Challenge (FERC-2013). The model utilised is a convolutional neural network-based deep learning model.The findings show that the suggested model outperforms earlier models in terms of emotion detection results. The FERC dataset was 70.14 percent accurate, whereas the JAFFE dataset was 98.65 percent accurate.

A smart baby cradle has been developed using IOT in order to help parents monitor the activities of the infant[8].The framework is composed of three sensors which are the temperature, sound, and gas sensors. The live images of the infants were captured using a camera module from the Wide Area Network (WAN). The humidity and temperature sensing module helps identify if the cradle is wet. The data from the sensors is stored in the cloud and then mailed to parents on a regular basis. On detecting any abnormal activity the cradle swinging mechanism is enabled by using motors.

Kulkarni and Talele [9] have performed tracking and detection of faces and objects from video input using the Viola Jones algorithm which carries out detection and cropping of the image.The image is then pre-processed and image segmentation is applied to it. Object detection and recognition are performed on the image. The discovered items will subsequently be tracked using the KLT algorithm, followed by the data fusion process. The method is resilient even in the presence of noise and clutter. Choosing a face frame from real-time surveillance cameras and analysing it reduces human effort and also eliminates human errors. However, the Viola Jones algorithm is very slow in training and is mostly effective when the face is in frontal view.

Convolutional neural networks have been used to detect and track moving objects from video input[10].This is followed by the extraction of frames and the Tensorflow library has been used for robust object detection. Once the object has been located, the tracking of the object is done using CNN that improves the performance significantly as it is trained in millions of classes. The model tracks objects at a speed of 150 frames per second. This is also able to remove the barrier of occlusion. The approach achieves an accuracy of 90.88 percent, 92.14 percent sensitivity, and 91.24 percent specificity.

Two techniques Haar Cascade and LBP Classifier have been compared for the purpose of facial recognition[11].A real time camera is used in order to capture the image.The input images are then converted to grayscale .The two techniques are then applied to the image using which the face and eyes are detected.There is a comparison between the two based on the factors of accuracy and time.The results indicate that Haar Cascade has a much higher level of accuracy and has the ability to detect a higher number of faces than the LBP classifier but LBP classifier is much faster than Haar Cascade. The background investigation reveals that no proposal has been made yet for a baby surveillance system that would notify parents. Hence, it is proposed to create an infant surveillance framework that will inform parents if the child suffers any injury when left with the caretaker.

III. INFANT SURVEILLANCE FRAMEWORK

The infant surveillance framework consists of three modules:infant activity detection module , object detection module and alert module as shown in Fig.1.The infant activity detection module is responsible for detecting critical situations such as infant crying, choking, or being hit. The object detection module is composed of two sub parts. The first part is concerned with harmful object detection. It is used to detect sharp and harmful objects like knives, scissors, and forks in the vicinity of the infant . The second part of the

Object Detection module is used to identify strangers who may pose a potential threat to the infant. This module works by comparing the image of the stranger to a database of photographs initially fed into the application by the user.The two modules work simultaneously, continuously monitoring the infant's environment.The alert module is concerned with displaying the alert on the mobile application when any of these five situations are encountered.Fig. 3 shows the alert on detection of stranger and Fig. 2 shows the alert on detection of sharp objects being displayed on the mobile application. Live-streamed video via webcam will be used as the input to the framework.

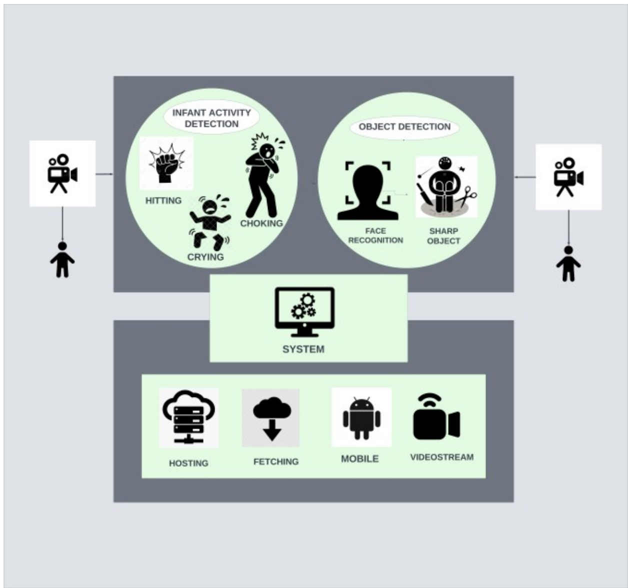


Fig. 1. Infant Surveillance Framework

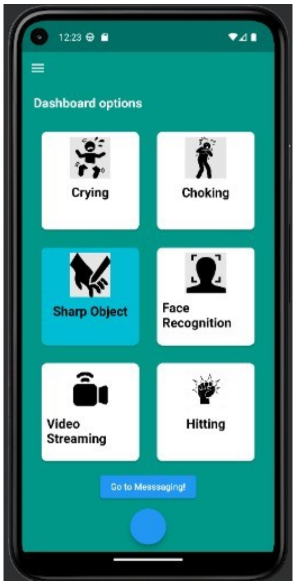


Fig. 2. Alert indicating detection of sharp objects on the Application

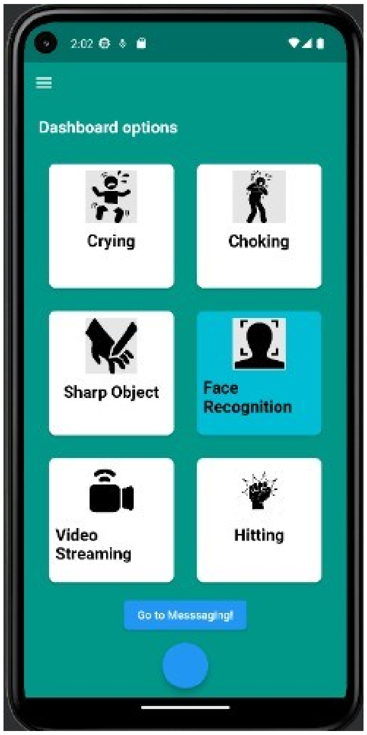


Fig. 3. Alert indicating detection of Stranger on the Application

IV. IMPLEMENTATION OF INFANT SURVEILLANCE FRAMEWORK

The implementation of the infant surveillance framework has three modules:infant activity detection module , object detection module and alert module.

The infant activity detection module utilizes a sequential model composed of InceptionV3 and Gated Recurrent Units(GRU). The model is used to identify instances of infant crying, infant being hit and infant choking. InceptionV3[16] is a deep convolutional neural network that is pre-trained on a large image dataset, and is used to extract features from the input video frames. These features are then fed into the GRU layer, a recurrent neural network to model sequential data. The GRU layer captures the sequence in the input and makes predictions about the infant's activity based on the patterns it has learned from the input sequence. The model is trained using a dataset of labeled videos of infants in different activity states.The sequential model summary is as shown in Fig.5.The object detection module uses the YOLO[15] framework to detect objects in images and video frames. The YOLO model consists of several convolutional and pooling layers, followed by a set of fully connected layers that is used to detect any harmful objects near the infant like knives, scissors and forks.YOLO is designed to be fast and efficient, and is able to process the images in real-time.The model is trained on a large dataset of labeled images.It is also used to detect the presence of any stranger in the infant's room.When a person is detected by the YOLO model, the input is first sent to the Haar Cascade classifier, which is a machine learning algorithm that detects objects in images using a set of pre-defined features. If only a particular part of the body is detected ,the YOLO algorithm shall specify that a person has been detected but in order to identify a stranger the face is the important aspect for which the Haar Cascade classifier has been used.If a face is detected by the Haar Cascade classifier, the input is then passed to the FaceNet[17] model, which is a deep neural network that is trained to recognize faces. The FaceNet model compares the input face to a pre-existing database of face embeddings, and makes a prediction about the identity of the person in the input image.The two models will work parallelly. The input to the models will be the live streamed video via webcam.The third module is the alert module to display the alerts on the application.The results computed by the model are hosted using Ngrok as an API. Flask has been used as the web framework for the API. The API is used by the application to fetch the results which are displayed as alerts on the mobile application.

The application has been developed using Flutter and has three pages:the dashboard page, the messaging page and video streaming page.The dashboard page is an overview of all the situations for which the alert will be raised.Through the dashboard page, parents can check the current condition of the baby and the description of this condition will be displayed on the messaging page .There is also a video streaming page where the parents can view and monitor their infant in real-time.

Overall, the Infant Surveillance Framework provides caregivers and parents with an additional layer of safety and security when caring for an infant. By continuously monitoring the infant's environment, the system can detect potential threats and provide caregivers and parents with the information needed to take appropriate action.

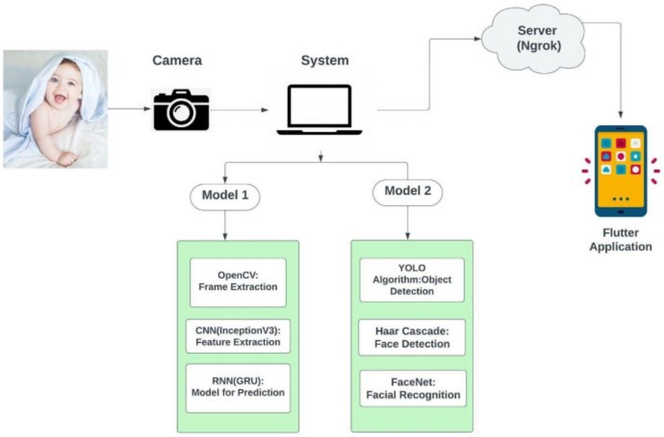
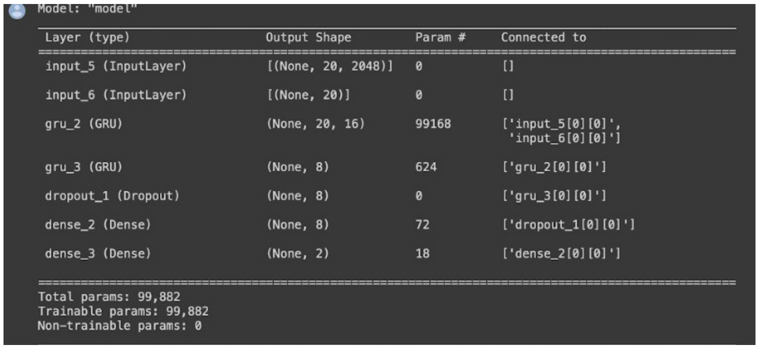


Fig. 4. Infant Surveillance Framework Implementation

V. RESULTS AND DISCUSSION

The infant activity detection model has been trained on several videos. The model hyper parameters consist of sparse categorical cross entropy as the loss function and Adam as the optimizer.The model has been trained for 100 epochs with a batch size of 64.



```
Model: "model"
________________________________________________________________________________________
 Layer (type)                Output Shape          Param #     Connected to
========================================================================================
 input_5 (InputLayer)        [(None, 20, 2048)]    0           []

 input_6 (InputLayer)        [(None, 20)]          0           []

 gru_2 (GRU)                 (None, 20, 16)        99168       ['input_5[0][0]',
                                                                'input_6[0][0]']

 gru_3 (GRU)                 (None, 8)             624         ['gru_2[0][0]']

 dropout_1 (Dropout)         (None, 8)             0           ['gru_3[0][0]']

 dense_2 (Dense)             (None, 8)             72          ['dropout_1[0][0]']

 dense_3 (Dense)             (None, 2)             18          ['dense_2[0][0]']

========================================================================================
Total params: 99,882
Trainable params: 99,882
Non-trainable params: 0
```

Fig. 5. Sequential Model Summary

The performance of the infant activity detection model is evaluated based on several metrics such as accuracy, precision, recall, and F1 score.The performance of infant activity detection model is shown in Table 1. and has an accuracy of 77%.The situation of infant being hit has a precision of 88%, recall is 64% and f1 score is 74%. For infant choking the precision is 88% ,recall is 70% and f1 score is 78% and infant crying has a precision of 67%, recall of 100% and f1 score of 80%.

The confusion matrix for the infant activity detection model is shown in Table 2.

Table 1. Performance of infant activity detection model

	precision	recall	f1score	support
hitting	0.88	0.64	0.74	11
choking	0.88	0.70	0.78	10
crying	0.67	1.00	0.80	10
accuracy			0.77	31
macro average	0.81	0.78	0.77	31
weighted average	0.81	0.77	0.77	31

Table 2. Confusion matrix for infant activity detection model

		Predicted Label		
		choking	crying	hitting
True Label	choking	7	1	3
	crying	1	7	2
	hitting	0	0	10

VI. ERROR ANALYSIS

The error analysis for stranger detection and for sharp object detection are in Table 3. and Table 4. respectively.

Table 3. Error Analysis for Stranger Detection

		Predicted Label		
		Person1	Person2	Stranger
True Label	Person1	120	0	15
	Person 2	5	110	30
	Stranger	8	3	174

Table 4. Error Analysis for Sharp Object Detection

		Predicted Label		
		Scissors	Fork	Knife
True Label	Scissors	98	0	2
	Fork	0	99	1
	Knife	1	2	97

VII. CONCLUSION AND FUTURE WORK

A novel and revolutionary approach for infant surveillance has been presented that is the need of the hour. Live streamed video is provided as input and on encountering situations like infant crying, infant choking, infant being hit, harmful object detection and stranger detection ,alerts are displayed in the mobile application in real time. These alerts would be extremely useful for parents and caretakers as it would help them to constantly monitor the infant even while being away from them.An accuracy of 77 % has been attained for the infant activity detection model.As a part of the future work, we seek to enhance the accuracy of the models as well as reduce the response time with respect to receiving the alerts on the mobile application.

REFERENCES

[1] P. Sivakumar, J. V, R. R and K. S, ”Real Time Crime Detection Using Deep Learning Algorithm,” 2021 International Conference on System, Computation, Automation and Networking (ICSCAN), 2021, pp. 1-5, doi: 10.1109/ICSCAN53069.2021.9526393

[2] C. -Y. Chang and F. R. Chen, ”Application of Deep Learning for Infant Vomiting and Crying Detection,” 2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA), 2018, pp. 633-635, doi: 10.1109/WAINA.2018.00158.

[3] N. Bordoloi, A. K. Talukdar and K. K. Sarma, ”Suspicious Activity Detection from Videos using YOLOv3,” 2020 IEEE 17th India Council International Conference (INDICON), 2020, pp. 1-5, doi: 10.1109/IN- DICON49873.2020.9342230.

[4] S. T. Ratnaparkhi, A. Tandasi and S. Saraswat, ”Face Detection and Recognition for Criminal Identification System,” 2021 11th International Conference on Cloud Computing, Data Science Engineering (Conflu- ence), 2021, pp. 773-777, doi: 0.1109/Confluence 51648.2021.9377205.

[5] Ali, Md. Forhad Khatun, Mehenag Turzo, Nakib. (2020). Facial Emotion Detection Using Neural Network. International Journal of Scientific and Engineering Research. 11. 1318-1325.

[6] R. Pathak and Y. Singh, ”Real Time Baby Facial Expression Recognition Using Deep Learning and IoT Edge Computing,” 2020 5th International Conference on Computing, Communication and Security (ICCCS), 2020, pp. 1-6, doi: 10.1109/ICCCS49678.2020.9277428.

[7] Jaiswal, Akriti, A. Krishnama Raju and Suman Deb. “Facial Emotion Detection Using Deep Learning.” 2020 International Conference forEmerging Technology (INCET) (2020): 1-5.

[8] S. Joseph, A. Gautham.J, A. Kumar and M. K. Harish Babu,”IOT Based Baby Monitoring System Smart Cradle,” 2021 7th In- ternational Conference on Advanced Computing and Communica- tion Systems (ICACCS), Coimbatore, India, 2021, pp. 748-751, doi: 10.1109/ICACCS51430.2021.9442022.

[9] V. Kulkarni and K. Talele, ”Video Analytics for Face Detection and Tracking,” 2020 2nd International Conference on Advances in Comput- ing, Communication Control and Networking (ICACCCN), 2020, pp. 962-965, doi: 10.1109/ICACCCN51052.2020.9362900.

[10] S. Mane and S. Mangale, ”Moving Object Detection and Tracking Using Convolutional Neural Networks,” 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS),2018, pp. 1809-1813, doi: 10.1109/ICCONS.2018.8662921.

[11] Anirudha B Shetty, Bhoomika, Deeksha, Jeevan Rebeiro, Ramyashree, Facial recognition using Haar cascade and LBP classifiers, Global Transitions Proceedings, Volume 2, Issue 2, 2021, Pages 330-335

[12] Ishak, Daing Abdul Jamil, Muhammad Mahadi Ambar, Radzi. (2017). Arduino Based Infant Monitoring System. IOP Conference Series: Materials Science and Engineering. 226. 012095. 10.1088/1757- 899X/226/1/012095.

[13] M. P. Joshi and D. C. Mehetre, ”IoT Based Smart Cradle System with an Android App for Baby Monitoring,” 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, 2017, pp. 1-4, doi: 10.1109/ICCUBEA.2017.8463676.

[14] Redmon, J. and Farhadi, A., 2018. Yolov3: An incremental improvement.arXiv preprint arXiv:1804.02767.

[15] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A., 2016. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779- 788).

[16] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In Proceed- ings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

[17] Schroff, F., Kalenichenko, D. and Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815- 823).