

Tracking Of Human Activities

**Jyothirmai Sai Sri Gelli¹, Yuva Sri Vemulapalli², Geethika Nimmagadda³, Sai Nagini
Vallurupalli⁴, Lakshmi Hima Bindu Lahori⁵ and G. Krishna Kishore⁶**

^{1,2,3,4,5,6}Department of Computer Science and Engineering, VR Siddhartha Engineering College,
Andhra Pradesh, India

¹jyothigelli@gmail.com, ²yuvasrivemulapalli@gmail.com, ³geethika200010@gmail.com,
⁴nagini186@gmail.com, ⁵himasai55@gmail.com, ⁶gkk@vrsiddhartha.ac.in

ABSTRACT:

Is it possible to know what is happening at a particular place precisely without seeing video clips where we are physically absent. Nowadays all of us being engaged with other works we don't have much time to spend on seeing the full-length video to know what is happening. But there is an alternative option for this, i.e in the form audio clip which is exactly like a person narrating the scene. The main advantage of this is we can simultaneously save time and multi-task i.e doing our work by listening to the audio clip that is generated by getting the up to date information and also if any person suddenly falls which may cause heavy injuries which can indirectly lead to a major medical issue for elderly people. So to prevent such emergencies, it will also provide an alarm system to detect human falls. This is possible by using current trending technologies like image processing and computer vision to capture the live moments, RNN with LSTMs to process and analyze the captured ones and by using natural language processing we can describe what is going on. Audio clips are generated by using Google Text to Speech API and they are sent to users.

Keywords: Human activity tracking system ,alert detection , LSTM , VGG16, Inception V3 .

1. INTRODUCTION

In our everyday daily life we see many CC cameras fixed in and around our surroundings. They record everything that is happening in the place 24/7 but we don't have enough time to watch everything that is being recorded. Is it viable to understand what is happening at a particular place precisely without watching video clips when we are not actually present in the scene.

Lately, the computerized approach of creating an interpretation in the form of a sentence for an image, has engrossed ample amount of researcher's attention in Artificial Intelligence which is entitled as image captioning, is very predominant in computer vision, i.e., allowing our PC's to acknowledge images, that is being utilized over numerous areas. Although this is undemanding to a person in-order to explain the content on the subject of a picture, which happens to be very demanding for advice to deal with. The task is arduous as they need a desktop for recording the items and their characteristics that are present in the snapshots, besides it also ought to convey the meaningful association that is there among them in the language that is easily understood. Preliminary attempts on captioning a picture primarily take on the model-based ways. This issue can be generally solved by using deep neural network with an encoder-decoder approach. These networks are built up by combining two sub

networks i.e. Convolution neural networks to portray the images and recurrent neural networks for language modeling.

Early image description generation methods aggregate image information using static object class libraries in the image and are modeled using statistical language models. Some indirect methods have also been proposed for dealing with image description problems, such as the query expansion method, retrieving similar images from a large dataset and using the distribution described in association with the retrieved images. All the methods described are brainstorming and have their own characteristics, but also have the common disadvantage that they do not make intuitive feature observations on objects or actions in the image, nor do they give an end-to-end mature general model to solve this problem [9].

There is some still disregarding the vent among low-level video featuring and sentence illustration, in the absence of evidently making use of advanced video concepts. For tackling such hardships, latter projects connect explicit advanced semantic concepts to the input picture or the recording [10]. The image explanation can also be procured by forecasting the major parts of the speech such as nouns, verbs, prepositions, adverbs etc. which obtains a meaningful sentence. [13] 3D-CNN solely holds the knowledge for a very little amount of time for the reason to which the convolution kernel size limits. This proved to pile up the semantic flaws vastly and created a decrease in relations among the produces words that has risen up the length of the tape. [12] LSTM is used as a decoder for obtaining the representation of sentences by enhancing the likelihood of an actual sentence. A “discriminator” module is added to the device frame work that is used as an opponent for the sentence generator. Automatic human fall detection is an important component in elderly fall recognition devices. Currently, various fall detection approaches came into the picture. Initially it was the vibration and pressure based systems, where sensors were put on the surface, for examining the person’s gestures where computer vision methodologies offered most encouraging and potent results.

Caption generation has raised a vast interest in images and videos. However, it is demanding for the models to select proper subjects in a complex background and generate desired captions in high-level vision tasks. From recent works, we suggest a novel image captioning model based on high-level image features [11].

Next was the accelerometer-based devices that were used as wearable devices. However, wearing such devices all day long made elderly people excruciating kind of sense with them. Thirdly, comes into picture the radar from of systems that makes use of Back-Scattered waves called ‘Doppler’ effects. Though relying on these radar signals, the scientific advancement restrains against many fake alerts, due to puzzled falls with the extra human actions such as walking or jogging. The last one is a vision based device that obtained huge prominence over the last 10 years for many possible grounds like it doesn’t need to be worn, this covers great surfaces and can be utilized indistinct camera sensors.

Around 703 billion people are aged 65 years and above among the world population as per 2019 records. The number is expected to double by 2050, i.e they constitute nearly 9% of the world's population. As of the WHO, nearly six hundred forty six thousand lethal collapses are identified annually all over the world, most of the falls are reported by adults who are aged over 65 years of age. The result of a fall can differ acutely depending on multiple factors like falling while walking, standing, sleeping or sitting in a chair. All these might relate to close traits but they also have a lot of dissimilarities among them. So it is a need of the situation to generate an alert system when an unexpected fall has been detected to diminish the effect of injuries and also the rate of fatal falls. Detection approaches are primarily treated with detecting collapses just after they have occurred and

activate an alert to their respective caretakers, while these strategies focus to estimate involuntary fall events in advance or at the time of occurring, thereby prompting for swift measures. At initial phase of detecting accidental falls they used vision based tactics which makes use of RGB cameras, infrared cameras and depth cameras because of their inexpensive and easy installation but they lack interpretation of other fall detection devices that are dependent upon non-vision sensors.

Swinging improvements in science and technology have paved a way for more tiny and economic electric gadgets such as low priced cameras and accelerometers lodged into smart phones might be more advantageous for identifying falls. Machine learning plays an important part in the analysis of data. It is important to consider both the assets and liabilities, this also has three major challenges to tackle with, they are real-world deployment performance, usability and acceptance. There is a dearth to acceptance to live in an environment that is being under surveillance by sensors. Another important point that is to be taken into consideration is that older people might not carry smart phones.

2. SIGNIFICANCE OF THE STUDY

People nowadays are unable to take care of their elderly ones and manage their work schedules as well. To prevent these kind of situations they are tending to quit on either of the things and even taking too much stress in managing things. Although Institutions are making use of CCTV cameras to identify any suspicious activities they have to allocate a person for knowing about the incident and Even though there is someone they can only oversee during their working hours. It might not be possible for any human being to watch each and every nook and corner 24/7. If this should be surveilled continuously the institution should recruit in large number to keep track anywhere in anytime. In hospitals, patient's health plays most important role for the management. So Patients risk should be assessed automatically every second and they should be notified immediately regarding this.

3. RELATED STUDIES

[1] A method is proposed by using encoder decoder framework in order to frame a clear caption for the given image. So to achieve this they considered the connection between words and credited distinct weights while training the model. They identified the main key advantage of the model they developed as their model can understand the crucial information in the image and can represent it in clear and unique way. [2] developed a model using an end-to-end approach which can sensibly treat features of image from the text which is already generated before. For image captioning they also proposed a bidirectional semantic attention-based guiding of LSTM which is also called BagLSTM. [15] mentioned about a system which helps in preventing detecting and reporting of any fall that is uncontrollable which might cause frequently in case of elderly people who are not having caretaker and in turn may lead to high risks of incidents. For this framework they used smartphone, triaxial acceleration of particular person's movements and some other features. They used a deep belief network for developing this model. training and testing for this system is done using 2 datasets which are having 10 classes in total i.e 9 classes having fall frames and 1 class with frames of daily activities. [7] OpenPose human posture estimation algorithm is used for developing fall detection model which is used for identifying fall in case of elderly people and to ensure safety for them. This is achieved by finding key points in human and joining those keypoints together to estimate the posture of the person using OpenPose human key point detection framework and uses SVDD classification algorithm for classifying whether it is fall or not SSD MobileNet object detection framework is also used along with open pose keypoint detection framework which helps in removing the non-key points detected. This also reduces the false positives of the algorithm and is robust even in a complex environment.

4. OBJECTIVES OF STUDY

- To produce an audio clip which can be heard at any time which is far more easy rather than watching the recorded video clip.
- To detect the sudden fall of a person and being able to alert the user instantly by producing an alarm to the end-user at the time of fall which helps in preventing complicated injuries.

5. PROPOSED SYSTEM

We proposed an intelligent system, which automatically detects the human activities, sends alert when unusual activity is detected. It generates an audio clip and sendsto the user which contains the activities the person preformed throughout the day. The proposed system can be used for old age people monitoring purposes. The goal of this studyis to develop such a system which tracks the actions of the person, converts it into an audio clip and sends it to the user and can easily recognize an unusual activity by alerting them. This software contains two modules. They are :

- Audio Generation module (AG) : This module captures video streaming from CCTV and the necessary processing is done and it finally generates an audio clip. This clip contains all the actions that are done by the person in the captured video. And that generated audio clip sent to the user.
- Alert System module (AS) : This module monitors a person's actions. If any unusual activity is detected then it immediately sends an alert call to the user.

Audio Generation module will be running and generates audio clip for every 1 hour throughout the life cycle of this software. Alert system module monitors the actions and if there is any unusual activity then it alerts the user.

This software provides users with real time updates. Monitoring system is installed at their houses (users choice) which will analyze the content that is recorded. For every one hour it generates an audio clip which contains the individual activities and it will be sent to the user. If any alert is detected by this monitoring system it will send an alert call.

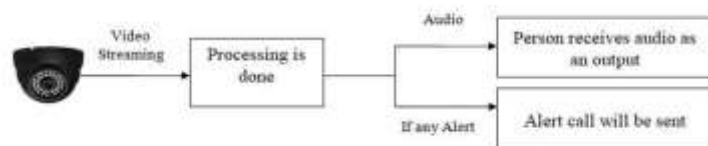


Figure 1 : Block Diagram for the proposed software

In the above (Figure 1), camera will be monitoring the persons in the room and these recordings are sent to trained models. In this architecture it contains two models one for generating audio and another one is to detect anomalous activities.

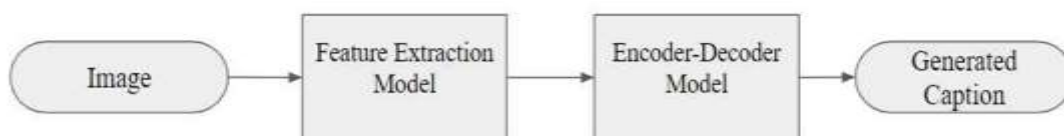


Figure 2 : Architecture for Audio Generation

In the above (Figure 2), this interface takes frames of the recorded video and sends it to the feature extraction model to extract features of that frame and then that model's output is sent to the Encoder - Decoder model which will generate captions. For every one hour all the generated captions as a whole are converted into a single audio clip by Google Text-To-Speech API.



Figure 3 : Architecture for Alert System

In the above (Figure 3), same as the Audio generation module, the same input is sent to this model and the purpose of this model is to detect unstable activities of the individuals. And if detected it sends an alert to the user by using Twilio API.

6. METHODOLOGY

As discussed in the proposed system part, this software has two distinct independent functionalities (Audio Generation and Alert System). To make it simple it is divided into two modules and they are loosely coupled.

The audio generation module contains a model called caption generation model which is a combination of two sub-models: namely Feature extraction model, which is developed by using InceptionV3 architecture for extracting the features, and encoder-decoder model, which is developed by using LSTM for generating the captions. And then generated captions are converted to audio by using Google Text to Speech API. This output is sent to the respective users.

Features are parts or patterns of an object in an image which help to generate text based on the correlation of these extracted features. InceptionV3 uses 1x1, 3x3, and 5x5 filters to learn where 1x1 learns patterns across the depth of the input image and 3x3 and 5x5 learn spatial patterns across all dimensional components (height, width, and depth) of the input image. While combining all the patterns learnt from varying filter sizes, representational power increases. This Inception module consists of a concatenation layer, where all the outputs and feature maps from the conv filters are combined into one object to create a single output of the Inception module. It finally gives a vector as output which consists of features of the input image. This vector is sent to an encoder-decoder model. In this encoder-decoder model, the encoder is a model which takes features of the input image and encodes the content into a fixed-length vector (34) using an internal representation and passes it to the decoder. The decoder is a model where it takes the encoded vector and generates the textual description of that image as output. The encoder-decoder model is built by using an inject model architecture where it combines the encoded vector of the given input image with each word from the text description that is generated so far, meaning it uses a sequence of both encoded vector and word information (embedding information) as input in order to generate the next word in the sequence.

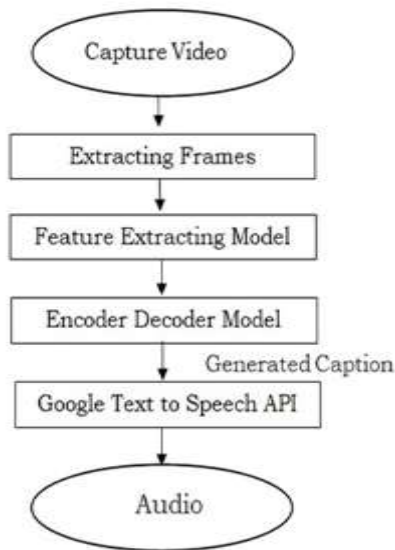


Figure 4 : Flow Diagram for Audio Generation Module

The above (Figure 4) represents the flow diagram of the Audio generating model. First and foremostly we collect the live recording video from a camera which is used for monitoring the person in a room. After extracting frames from video we send these frames to the Feature Extracting Model. Then the output from this model is sent as input to the Encoder-Decoder model which generates sentences describing the features present in the input frames. These generated captions are sent as input to Google Text to Speech API to convert them into Audio.

As software developers we should maintain both production cost and product accuracy. In this module, it should find out the relation between the objects which is only achieved by deep analysis and it should send to the encoder-decoder model. So this functionality is provided by inception V3 where it mainly focuses on extracting deep features inside an image in a particular area. So that means we must try out to select a model which reduces the cost of computational cost rather than computational accuracy. So the feature extraction model is developed by using inception v3 architecture which its main goal is to reduce computational cost instead of Resnet where it mainly focuses on computational accuracy and increases computational cost due to its layer's equation. Here in Figure 5 it is residual block where it takes input from the above layer's output and also at ending of that module previous layer's output is also added to result and then it is send to the down layers where its computational increased. But in this case Figure 6 in inception Module it choses filters based on the object size it should learn i.e without adding any additional layers at that level, it is deciding which one is better for learning and which reduces both time and cost.

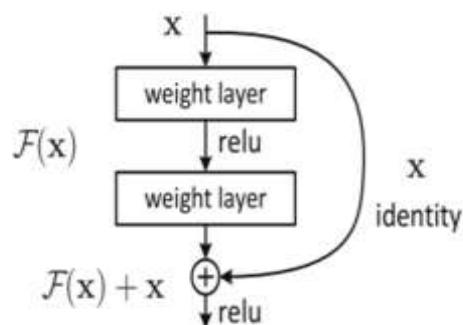


Figure 5 : Residual Block

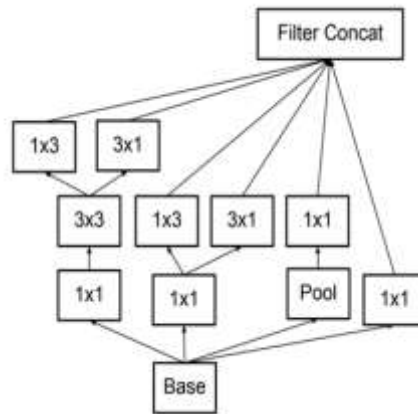


Figure 6 : Inception Module

Alert System module's main goal is to detect falls during monitoring. Here the main task lies in classification rather than extracting features of the frame deeply. In order to achieve this we used vgg16. In this neural network we used Adam optimizer and Categorical crossentropy loss function.

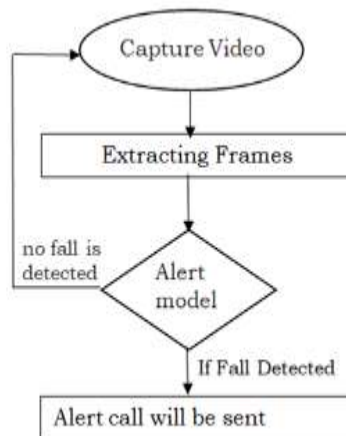


Figure 7 : Flow Diagram for Alert System

The above (Figure 7) it is the flow chart for alert system. In emergency situations giving an Alert is more important than generating an audio for the guardian. So the frames extracted from video are first sent to the Alert model to identify whether there is any fall detected or not. If there is any fall identified then immediately an alert call will be sent to the concerned person else these frames will be sent to the Audio generation model to produce an audio clip. For generating alert calls, the Twilio package is used.

7. RESULT ANALYSIS

When the model is given 4 epochs then the accuracy is 0.91 and for 5 epochs it predicted with 0.93 accuracy but when 6 epochs is considered it lead to overfitting. So if we choose 6 epochs then the model will not be able to classify accurately. So 5 epochs is chosen to run the model accurately.

Frame No	Extracted Frame	Caption generated	Alert call
----------	-----------------	-------------------	------------

Tracking Of Human Activities



1		Two children are playing on see-saw	Since there is no fall in the frame alert call won't be sent
2		One of the child is falling down from the see-saw	Immediately alert call is sent

Table 1: Test case description for Frames extracted from 1st input video clip.



Frame No	Extracted Frame	Caption generated	Alert call
1		A man standing backside of the car	Since there is no fall in the frame alert call won't be sent
2		A man fallen down nearby car	Immediately alert call is sent

Table 2: Test case description for Frames extracted from 2st input video clip.

The above tables (Table1 and Table 2) represents how the input video is being divided into frames and will be passed to models. So that after the frames are passed to model it generated the corresponding captions as mentioned in the table and if there is any fall detected alert call is sent immediately.

8. CONCLUSION

In this paper we proposed an intelligent system, which automatically detects the human activities, sends alert when unusual activity is detected. It generates an audio clip and sends to the user which contains the activities the person preformed throughout the day. The proposed system can be used for old age people monitoring purposes. The goal of this study is to develop such a system which tracks the actions of the person, converts it into an audio clip and sends it to the user and can easily recognize an unusual activity by alerting them. In the future, the system can be extended to multiple people tracking. It can also be made to recognize the person whom it is tracking and it can be made in such a way that it can be tracked continuously the same person even with the change of room.

REFERENCES

- [1] Ding, G., Chen, M., Zhao, S. et al. "Neural Image Caption Generation with Weighted Training and Reference". *CognComput* 11, 763–777 (2019). <https://doi.org/10.1007/s12559-018-9581-x> .s
- [2] Cao, P., Yang, Z., Sun, L. et al. "Image Captioning with Bidirectional Semantic Attention- Based Guiding of Long Short-Term Memory". *Neural Process Lett* 50, 103– 119 (2019). <https://doi.org/10.1007/s11063-018-09973-5>
- [3] Haoran Wang, Yue Zhang, Xiaosheng Yu, "An Overview of Image Caption Generation Methods, Computational Intelligence and Neuroscience". vol. 2020, Article ID 3062706, 13 pages, 2020. <https://doi.org/10.1155/2020/3062706>
- [4] Sujin Lee, Incheol Kim, "Multimodal Feature Learning for Video Captioning", *Mathematical Problems in Engineering*, vol. 2018, Article ID 3125879, 8 pages, 2018. <https://doi.org/10.1155/2018/3125879>
- [5] Jeffin Gracewell, J., Pavalarajan, S. "Fall detection based on posture classification for smart home environment". *J Ambient Intell Human Comput* (2019). <https://doi.org/10.1007/s12652-019-01600-y>
- [6] Y. Yang et al., "Video Captioning by Adversarial LSTM". in *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5600-5611, Nov. 2018, doi: 10.1109/TIP.2018.2855422.
- [7] G. Sun and Z. Wang, "Fall detection algorithm for the elderly based on human posture estimation," 2020 Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), Dalian, China, 2020, pp. 172-176, doi: 10.1109/IPEC49694.2020.9114962.
- [8] K. Sehairi, F. Chouireb and J. Meunier, "Elderly fall detection system based on multiple shape features and motion analysis," 2018 International Conference on Intelligent Systems and Computer Vision (ISCV), Fez, 2018, pp. 1-8, doi: 10.1109/ISACV.2018.8354084.
- [9] Ordonez, V.; Kulkarni, G.; Berg, T.L.: Im2text: describing images using 1 million captioned photographs. In: *Proceedings of Advances in Neural Information Processing Systems*, pp. 1143–1151 (2011)
- [10] Dash, S.K.; Saha, S.; Pakray, P.; Gelbukh, A.: Generating image captions through multimodal embedding. *J. Intell. Fuzzy Syst.* 36(5), 4787–4796 (2019)
- [11] Ding, S.; Qu, S.; Xi, Y.; Sangaiah, A.K.; Wan, S.: Image caption generation with high-level image features. *Proc. Pattern Recognit. Lett.* 123, 89–95 (2019)
- [12] Gan, Z.; Gan, C.; He, X.; Pu, Y.; Tran, K.; Gao, J.; Carin, L.; Deng, L.: Semantic compositional networks for visual captioning. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1141–1150 (2017)
- [13] Simonyan, K.; Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*(2014)
- [14] Karpathy, A.; Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128– 3137 (2015)
- [15] Jahanjoo, A., Naderan, M. & Rashti, M.J. Detection and multi-class classification of falling in elderly people by deep belief network algorithms. *J Ambient Intell Human Comput* 11, 4145–4165 (2020). <https://doi.org/10.1007/s12652-020-01690-z>