International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 03 | March - 2023 Impact Factor: 7.185 ISSN: 2582-3930

TRACKING OF HUMAN ACTIVITES

Mr.Md Imran

Department of Information Technology

Andhra Loyola Institute of Engineering and Technology

V. Abhinav Raj, M. Sravan, A. Veera Gopi Chand

Department of Information Technology

Andhra Loyola Institute of Engineering and Technology

Abstract--Is it feasible to accurately understand what is happening at a certain location when we are not physically present without watching video footage. Since we are all currently busy with other tasks, we don't have a lot of time to dedicate to watching the entire film in order to understand what is happening. But, there is another choice for this, namely an audio clip with a person actually 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 that may lead to a major medical issue for elderly people. Therefore to prevent such emergencies, it will also feature an alarm system to detect human falls. This is made feasible by utilising cutting-edge technology like computer vision and image processing to record live events, RNN with LSTMs to process and analyse the recorded ones, and natural language processing to provide a description of what is happening. Users receive audio clips that are created using the Google Text to Speech API.

Keywords: Image processing, Computer Vision, Natural Language Processing, Convolution Neural Networks.

I. INTRODUCTION

We frequently encounter CC cameras installed in and around our surroundings in daily life. They film everything that is happening in the location 24/7 but we don't have enough time to watch

everything that is being recorded. Is it possible to accurately comprehend what is happening at a certain location while we are not physically present there without watching video clips?

There may be occasions when you must leave for work, leaving elderly family members alone at home. In this In some circumstances, it can be difficult for us to monitor their situation, and they themselves might be unable to contact with us. Then there is our application called "Tracking of Human Activities," where we can monitor a variety of human behaviours, create an audio clip from the input recording, and even recognise a person's unexpected fall and send out an alert. We can listen to something that has been captured and converted to audio while driving or even while seated in a conference. This helps us save a tonne of time while also letting us know what is going on in the area.

Early approaches for creating picture descriptions assemble image data using the image's static object class libraries, which are then characterised using statistical language models. The query expansion method, which pulls similar images from a huge dataset and uses the distribution stated in association with the obtained photos, are some indirect approaches to solving the problem of image description that have also been presented.. The common drawback of all the brainstorming techniques mentioned is that they neither provide an end-to-end mature general model to address this issue nor do they make intuitive feature observations on items or actions in the image.

Deep neural networks with encoder-decoder components are used to generate natural language descriptions of images on-the-fly. The use of an attention-based technique to determine where to focus in the image or video has been attempted in some later worksOn the other hand, they continue to ignore the difference between sentences that describe low-level video elements and those that clearly express high-level video concepts. Recent works include explicit high-level semantic ideas of the input image/video to overcome the issues. The most likely nouns, verbs, situations, and prepositions that make up the sentence can also be used to predict the visual description.

Because to the convolution kernel size constraint, 3D-CNN can only capture data over a brief period of time. Unfortunately, it was discovered that using this strategy led to an exponential accumulation of grammatical errors and decreased word association as video length increased. A "discriminator" module is introduced to the system architecture in order to address this LSTM flaw, acting as an opponent to the sentence generator.

A huge interest in photographs and videos has been sparked via caption generating. In high-level vision tasks, it is challenging for the models to choose appropriate subjects in a complicated context and produce desirable captions. We propose an unique image captioning model based on high-level image attributes in light of current efforts. The senior monitoring system's most crucial component is automatic human fall

detection. Several fall detection strategies have recently been proposed. In order to assess a person's movement using vibration and pressure-based systems, which place sensors on the floor, comes first. Here, computer vision-based technologies offer promising and practical solutions.

Second, wearable technology based on accelerometers was demonstrated. However, because they are wearable devices, they must be worn constantly, which can be painful for elderly persons. The third group includes radar-based systems that employ "Doppler effects" from backscattered waves. Although reliant on radar signals, this technology frequently generates false alarms since falls are mistaken for other human actions, such as laying down or sitting up. The final one is a vision-based system, which has become extremely important in the past ten years for a variety of reasons, including the fact that it does not need to be worn, can cover large spaces, and can employ various camera sensors.

II. LITERATURE SURVEY

1. **Neural Image Caption Generation with Weighted Training and Reference.**
Ding, G., Chen, M., Zhao, S. et al.

This paper mainly focused on automatically generating captions for an image. They have used the encoder-decoder framework to generate a more descriptive sentence for the given image. They have used different weights for the words according to the correlation between words and images during the training phase. They maximized the consensus score between the captions generated by the captioning model and the reference information from the neighboring images of the target image, which can reduce the misrecognition problem. They mainly used computer vision and natural language processing domains. They have conducted experiments and comparisons on the datasets MS COCO and Flickr30k.

2. **Image Captioning with Bidirectional Semantic attention-Based Guiding of Long Short-Term emory.**
Cao, P., Yang, Z., Sun, L. et al. [

In this paper they have used an end-to-end approach to propose a bidirectional semantic attention-based guiding of long short-term memory (Bag- LSTM) model for image captioning. The proposed model consciously refines image features from the previously generated text. By fine-tuning the parameters of convolution neural networks, Bag-LSTM obtains more text-related image features via feedback propagation than other models. This model dynamically leverages more text-conditional image features, acquired by the semantic attention mechanism, as guidance information. They have used bidirectional LSTM as the caption generator, which is capable enough of learning long term relations between visual features and semantic

information by making use of both historical and future information.

3. **An Overview of Image Caption Generation Methods", Computational Intelligence and Neuroscience**

Haoran Wang, Yue Zhang, Xiaosheng Yu

In this paper, they used 2 models statistical probability language model to generate handcraft features and a neural network model based on an encoder-decoder language model to extract deep features. They used several attention mechanisms to improve the effect of image captioning. In this model, they used MSCOCO, Flickr8k, Flickr30k, PASCAL 1K, AI Challenger Dataset, and STAIR Captions datasets. BLEU and METEOR are for machine translations similarly ROUGE, CIDEr, and SPICE are used for several other evaluation criteria.

4. **Multimodal Feature Learning for Video Captioning**

Sujin Lee, Incheol Kim

In this paper, visual features of the input video are extracted using C3D and ResNet, and semantic features are obtained using RNN such as LSTM. Semantic feature learning is used to identify actions, objects, persons, and background in the input video whereas Attention-based caption generation is used for effective caption generation using multimodal features. Part-Of-Speech (POS) tag function in Natural Language Toolkit (NLTK) was used to separate nouns and verbs, while plural nouns and tenses of verbs, past, continuous, and so on, were converted back to their root forms using the lemmatize function in NLTK.They have trained their model using MSRVTT, MSVD datasets.

5. **Video Captioning by Adversarial LSTM**

Y. Yang et al

They proposed a novel approach for video captioning and adopted a standard generative adversarial network (GAN) architecture, characterized by an interplay of two competing processes. For generator module, they took an existing video captioning concept using LSTM network. For discriminator, they proposed a novel realization specifically tuned for the video captioning problem.

6.Fall detection algorithm for the elderly based on human posture estimation

G. Sun and Z. Wang

They used a falil detection model based on Open Pose

Human Posture estimation algorithm. It is based on Open Pose human key point detection, combined with SSD Mobile Net object detection framework. It used SVDD classification algorithm to classify them.

7.Elderly fall detection system based on multiple shape features and motion analysis

K. **Sehairi, F. Chouireb and J. Meunier**

Silhouette of a person is extracted using a background subtraction technique. They presented

a technique to estimate the head position, and a finite state machine (FSM) to compute the vertical velocity of the head. They tested on the L2ei dataset. It contains more than 2700 frames have been labeled in order to train 3 different classifiers.

III. METHODOLOGY

This software contains two modules.

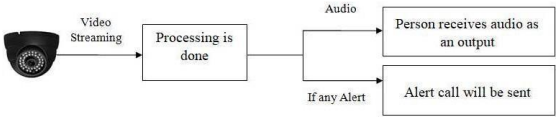
They are.

a. Audio Generation Module

b. Alert System Module

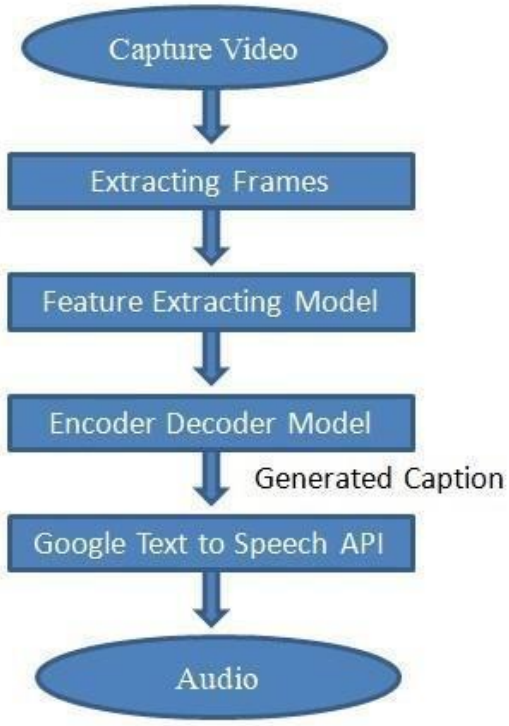
Audio Generation module will be running and generates audio clip for every 1 hour throughout the life cycle of this software. Alert system module monitors the actions and if there is any unusual activity then it alerts the user. To design this software, we are using agile methodology.

Block Diagram for Tracking of Human Activities



Flow chart for Audio Generation

Audio generating model is developed by using CNN for extracting the features and 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.



Steps:

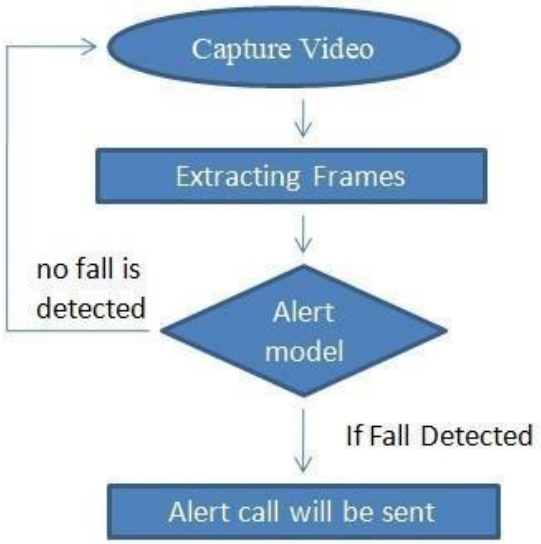
- Capturing video:
  - using **cv2 module** video streaming coming from cctv is captured.
- Extracting frames:
  - To extract frames from the captured video **Video Capture ()** function of cv2 module is used.
- Feature Extracting model:
  - After extracting the frames, those frames are given input to this trained model, which is used to extract the features in the given input frame.

- Feature Extraction Model is constructed by convolution neural network (Inception Resnet V2 architecture).
- **keras API** is used.

- Encoder - Decoder Model:
  - This model finally generates description for the given frame.
  - Encoder-Decoder Model is constructed by using **RNN with LSTM**.
  - **keras API** is used.
- Audio generation:
  - The generated caption is sent to **Google Text to Speech API** which will convert text to audio.

Flow chart for Alert System

An alert model is developed using CNN. If this model classifies it as a fall then an alert call will be send to the user.



Steps:

Capturing video:

- using **cv2 module** video streaming coming from cctv is captured.
- Extracting frames:
- To extract frames from the captured video **Video Capture ()** function of cv2 module is used.

Alert model:

- After extracting the frames, those frames are given input to this trained model, which will classify whether the given frame is fall or not fall.
- If its classifies as fall, then an alert call will be sent.

IV.PROPOSED MODEL

In this proposed system it contains two modules.

1. Audio Generation module : This module captures video streaming from Cctv and the

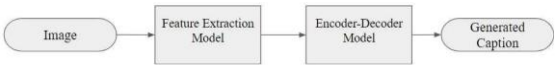
necessary processing is done and it finally generates an audio clip. This clip contains all the actions that are done by the person in the captured video. And that generated audio clip sent to the user.

2. Alert System module: 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

This module generates an audio clip every 1 hour .Here first it takes video and then it preprocesses it .This preprocessed data is then sent to the caption generation model. And that output is converted to an audio clip using API and send to the user .

In this caption generation model, it generates a caption for the captured once. These generated captions are stored and for every one hour, it converts those captions into an audio clip. This audio clip contains the actions performed by the person.



This architecture consist of three models:

- Feature Extraction Model
- Encoder Model
- Decoder Model

Feature Extraction Model

This model is basically responsible for acquiring features from an image for training .It finally gives a vector as output which consist of features of the input image .This vector is sent to an encoder model as input.

Features are parts or patterns of an object in an image that help to identify it. Traditional Computer Vision techniques for feature detection include:

a. **Harris Corner Detection —** Uses a Gaussian window function to detect corners.

b. **Shi-Tomasi Corner Detector —** The authors modified the scoring function used in Harris Corner Detection to achieve a better corner detection technique.

c. **Scale-Invariant Feature Transform (SIFT) —** This technique is scale invariant unlike the previous two.

d. **Speeded-Up Robust Features (SURF) —** This is a faster version of SIFT as the name says.

e. **Features from Accelerated Segment Test (FAST) —** This is a much more faster corner detection technique compared to SURF.

f. **Binary Robust Independent Elementary Features (BRIEF) —** This is only a feature descriptor that can be used with any other feature detector. This technique reduces the memory

usage by converting descriptors in floating point numbers to binary strings.

g. **Oriented FAST and Rotated BRIEF (ORB)** — SIFT and SURF are patented and this algorithm from OpenCV labs is a free alternative to them, that uses FAST keypoint detector and BRIEF descriptor.

ALTERNATE FEATURE EXTRACTION MODELS

Convolutional neural networks (CNNs) can replace conventional feature extractors since they are significantly more effective and have a strong ability to extract complicated characteristics that express the image in greater detail. On this, numerous works have been produced. This is a list of a few of them:

a. **SuperPoint: Self-Supervised Interest Point Detection and Description** — The authors suggest a fully convolutional neural network that computes SIFT like interest point locations and descriptors in a single forward pass. It uses an VGG-style encode for extracting features and then two decoders, one for point detection and the other for point description.

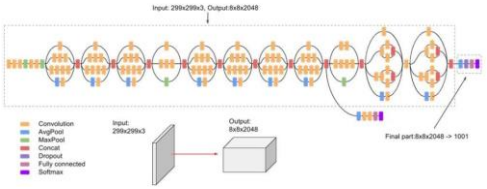
b. **D2-Net: A Trainable CNN for Joint Description and Detection of Local Features** — The authors suggest a single convolutional neural network that is both a dense feature descriptor and a feature detector.

c. **LF-Net: Learning Local Features from Images —** The authors suggest using a sparse-matching deep architecture and use an end-to-end training approach on image pairs having relative pose and depth maps. They run their detector on the first image, find the maxima and then optimize the weights so that when run on the second image, produces a clean response map with sharp maxima at the right locations.

The InceptionV3 architecture was used in the creation of this model. Inception v3 is a widely-used image recognition model that has been demonstrated to obtain higher than 78.1% accuracy on the ImageNet dataset. The model is the sum of several ideas developed by different researchers over the years. Rethinking the Inception Architecture for Computer Vision by Szegedy et al. served as its foundation. Convolutions, average pooling, max pooling, concatenations, dropouts, and fully linked layers are some of the symmetric and asymmetric building components that make up the model itself.

The model makes considerable use of batchnorm and applies it to activation inputs. Softmax is used to calculate loss.

A high-level diagram of the model is shown below:



Encoder-Decoder Model

The encoder-decoder architecture for recurrent neural networks is the standard neural machine translation method that rivals and in some cases beats classical statistical machine translation methods.

Google's translate service uses a recurrent neural network architecture with an encoder-decoder at its heart.

a) **Sutskever model** - for direct end-to-end machine translation.

b) **Cho model** - that extends the architecture with GRU units and an attention mechanism.

a) Sutskever NMT Model:

It was one of the first neural machine translation systems to outperform a standard statistical machine learning model on a significant translation problem, making it a key model in the field of machine translation.

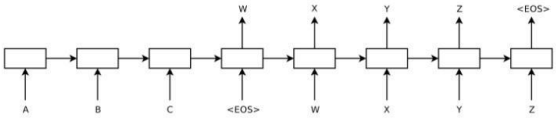
 A portion of the dataset's 12 Million phrases, which included 348 Million French and 304 Million English words, were used to train the model. This set was selected as it had already been tokenized.

 The source vocabulary was condensed to the 80,000 most often used target French terms and the 160,000 most frequently used source English words. The "UNK" token was used to replace any words that were not common terms.

Model:

A full input sequence was read and encoded to a fixed-length internal representation using an encoder-decoder architecture. Then, until the end of the sequence token was reached, a decoder network output words using this internal representation. The encoder and decoder both made use of LSTM networks.

Five deep learning models were used to create the final model. The translations were inferred using a left-to-right beam search.



Model Configuration:

- Input sequences were reversed.
- A 1000-dimensional word embedding layer was used to represent the input words.
- The input and output models had 4 layers with 1,000 units per layer.

- The model was fit for 7.5 epochs where some learning rate decay was performed.
- A batch-size of 128 sequences was used during training.
- Batches were comprised of sentences with roughly the same length to speed-up computation.

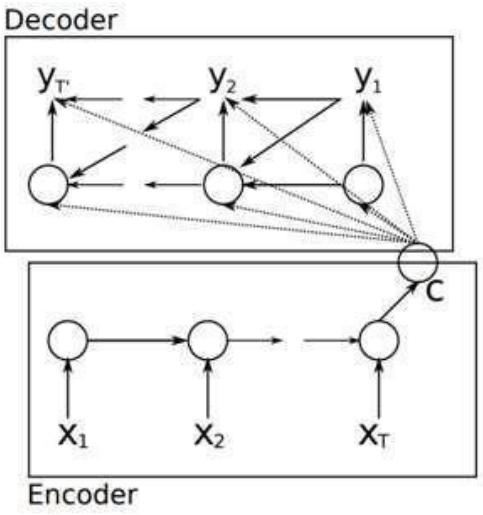
b) **Cho NMT Model:**

As opposed to the Sutsk ever model mentioned above, the Cho Model is solely used to evaluate potential translations. Although the model is used directly and only for translation, there have been extensions to the work to better diagnose and improve the model.

 The source and target vocabulary were restricted to the 15,000 most common French and English words, which account for 93% of the dataset. Words that were not in the source or target vocabulary were substituted with "UNK".

Model :

The model employs the same two-model strategy, explicitly referring to the encoder-decoder architecture in this instance.



Instead of using LSTM units, a more straightforward recurrent neural network unit known as the gated recurrent unit, or GRU, is devised for the implementation.

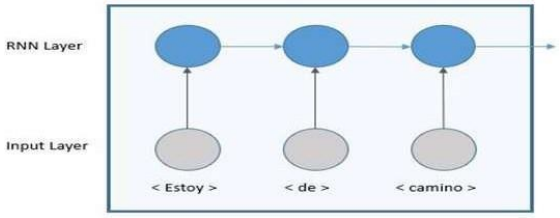
Model Configuration:

- A 100-dimensional word embedding was used to represent the input words.
- The encoder and decoder were configured with 1 layer of 1000 GRU units.
- 500 Maxout units pooling 2 inputs were used after the decoder.
- A batch size of 64 sentences was used during training.
- The model was trained for approximately 2 days.

Encoder:

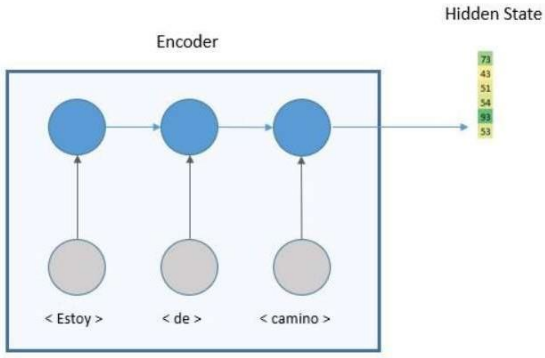
Data must be encoded in order to be in the desired format. In the context of machine learning, we

transform a list of Spanish words into a two-dimensional vector, sometimes referred to as the hidden state. Recurrent neural networks are stacked to create the encoder (RNN). We employ this kind of layer since the model can comprehend the context and temporal relationships of the sequences thanks to its structure. The previous RNN timestep's state is the concealed state, the encoder's output.



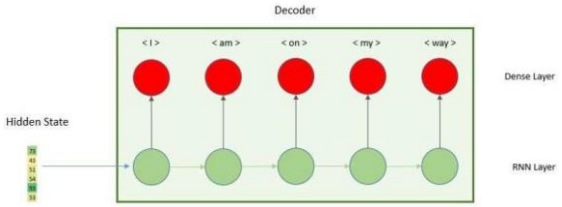
Hidden State:

A two-dimensional vector representing the entire meaning of the input sequence is the encoder's output. The number of cells in the RNN determines the length of the vector.



Decoder Model:

A message that has been encoded must be decoded before it can be understood. The Pictionary team's second member will transcribe the image into a word. The decoder will transform the two-dimensional vector into the output sequence, which is the English phrase, in the machine learning model. In order to predict the English term, it is also constructed with RNN layers and a thick layer.



- The possibility of different input and output sequence lengths is one of this model's key advantages. This makes way for some really intriguing uses, such question-and-answer sessions or video captioning.
- For extended input sequences, this simple encoder-decoder model's main limitation is that all the data must be condensed into a one-dimensional vector, which can be quite challenging to do.

ALERT GENERATION MODULE

This module detects fall of a person and sends an alert call to the concerned user. Here, video is captured first, and then it is pre-processed. The Alert model receives this pre-processed data after that. The user will receive an alert call if the output generated by that model is declining, for example.

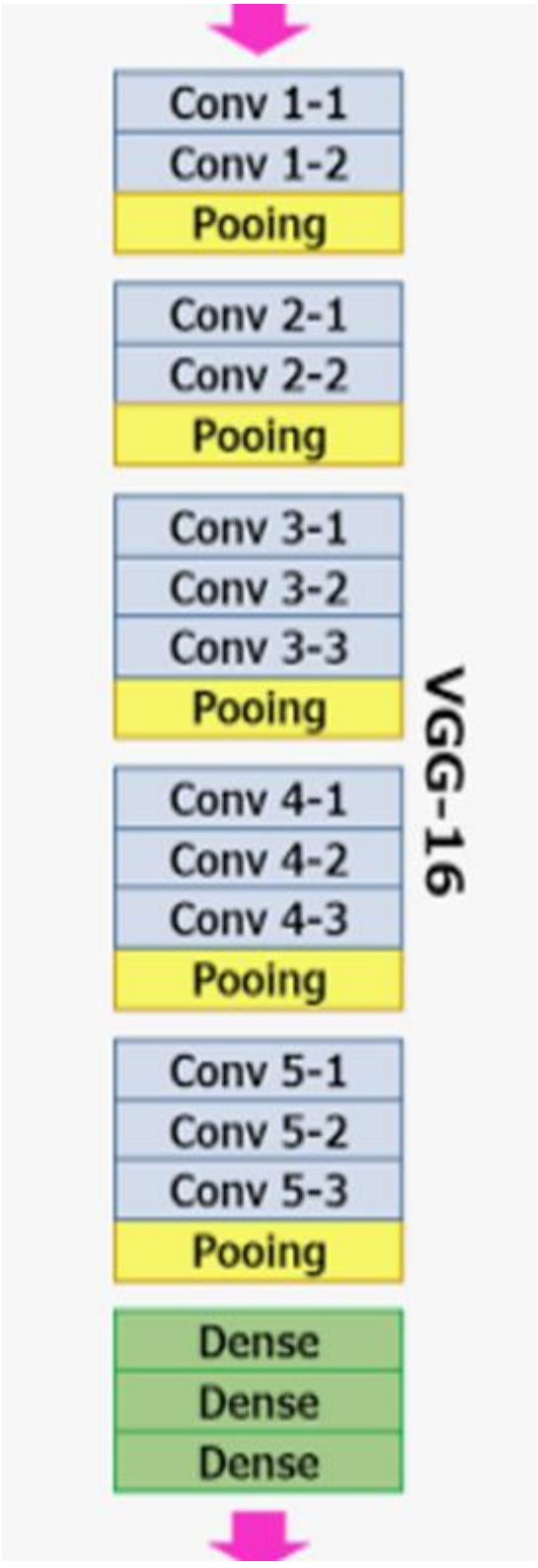
Whether the captured individuals are a fall or not is determined by this Alert model. The user will be

informed by a call if it is a fall.

Architecture

The pre-trained model VGG16 was used to create this alert model. The network receives a dimensioned image as input (224, 224, 3). The first two layers have the same padding and 64 channels with a 3*3 filter size. Following a max pool layer of stride (2, 2), two layers with 256-layer convolutions and a filter size layer are added (3, 3). This was followed by a stride (2, 2) max pooling layer that was identical to the layer before it. The following two convolution layers have 256 filters and a filter size of 3 and 3. There are then two sets of three convolution layers followed by a max pool layer. Each has the same padding and 512 filters of size (3, 3). The stack of two convolution layers then receives this image. Instead of using 11*11 in Alex Net and 7*7 in ZF-Net, we employ 3*3 filters in these convolution and max pooling layers. It also uses 1*1 pixels in some of the layers, which is utilised to control the number of input channelsTo prevent the spatial characteristic of the image, 1-pixel padding is applied after each convolution layer

padding is applied after each convolution layer.



Architecture of VGG16 Model

Dataset Collection

We are using Flickr8k_Dataset. After downloading the dataset from official website two zip files will be downloaded.

a) Flickr8k_Dataset.zip

b) Fickr8k_text.zip

Download the datasets and unzip them into your current working directory. You will have two directories:

a) **Flickr8k_Dataset**: Contains 8092 photographs in JPEG format.

b) **Flickr8k_text**: Contains several files containing different sources of descriptions for the photographs.

The dataset has a pre-defined training dataset (6,000 images), development dataset (1,000 images), and test dataset (1,000 images).

V.RESULTS

In this project we have created two modules namely

1. Audio Generation Module
2. Alert System Module

The following are the result analysis for the above-mentioned modules.

Alert Generation Module

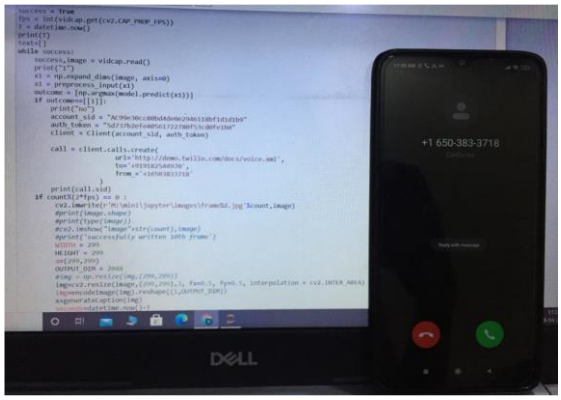
The generated captions are displayed in the terminal andgenerated audio file will be saved at the workin directory of our project.

The main goal of the study is to predict the action being performed in the given input picture. The model is built by using LSTM and CNN algorithms. The model is later trained using approx. 8000 pictures. The resultant trained model gave the above output while testing.

The generated caption is then transferred to the google text to speech API which then converts the text into an audio file. Which will then be sent to the end user

Alert **Generation Module**

This module generates an alert call if any fall is detected in live monitoring.



In the above Fig an alert call is generated by the model(Alert model) created because it detected a fall during live monitoring.

VI. CONCLUSION

In this research, we suggested an intelligent system that monitors human activity automatically and alerts users when it notices unexpected activity. It creates an audio clip and delivers it to the user, containing all the tasks they completed that day. The proposed technology can be utilised to monitor elderly persons. The purpose of this project is to design such a system which tracks the actions of the person, converts it into an audio clip and provides it to the user and can readily notice an unusual activity by warning them.

VII.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 .

[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] JeffinGracewell, 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]. A. Torralba, R. Fergus and W. T. Freeman, "80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 11, pp. 1958-1970, Nov. 2008, doi: 10.1109/TPAMI.2008.128.

[10]. 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)

[11]. Dash, S.K.; Saha, S.; Pakray, P.; Gelbukh, A.: Generating image captions through multimodal embedding. J. Intell. Fuzzy Syst. 36(5), 4787–4796 (2019)

[12]. Zhou, C.; Mao, Y.; Wang, X.: Topic-specific imagecaption generation. In: Proceedings of Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, pp. 321–332 (2017)

[13]. 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)

[14]. 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)

[15]. Simonyan, K.; Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556(2014)

[16]. 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)