A Robust Approach to Increase Productivity and Safety in Garments Using Deep Learning and Image Recognition

Shafkat Waheed
Research Assistant
Electrical and Computer Engineering
North South University
shafkat.waheed@gmail.com

B. M. Raihanul Haque
Research Assistant
Electrical and Computer Engineering
North South University
raihanul.haque@northsouth.edu

Dr. Mohammad Ashrafuzzaman Khan
Assistant Professor
Electrical and Computer Engineering
North South University
mohammad.khan02@northsouth.edu

***Abstract*—This paper investigates and implements the application of image recognition and captioning in garments sector to increase production and ensure the safety of workers by minimizing workload imbalance and to tackle production line problems. Though Bangladesh is the second-largest garments manufacturing country, many times it fails to meet the production goals in due time. Moreover, in many cases, it is unable to meet the safety standards of the industry which causes losing contract from international buyers. To resolve this issue, we have proposed a model to detect complex work using non-invasive CCTV camera. Our model has incorporated three main algorithms which are VGG19, ResNet and Inception-V3 along with other methods like CNN, RNN-LSTM, Show_attend_tell methods and implemented them on three different datasets. Using the model, we can derive the work description from images to detect anomaly and time for completion of the task using graph matching. We used the BLEU score (bilingual evaluation understudy) to evaluate our model and by far Inception-V3 has produced better results.**

1. Introduction

Bangladesh is the second-largest exporter of garments in the globe and it is all due to the relentless works of the textile employees. In 2012 and 2013, the back to back devastating incident in garment industry left around 1200 workers dead. Since then external and internal safety measures have been taken to reduce the loss and increase a better working environment. Additionally, absence of easily on-hand middle management, few numbers of manufacturing methods, sluggish backward or forward blending procedure, time-consuming production have caused the industry to fall behind. Keeping all that in mind we have devised a system that will enhance the flexibility in the internal working place by ensuring safety and increasing efficiency.

It is estimated that by 2030, the advantage of cheap labour will be irrelevant due to fourth industrial revolution and thus, it is a concern how we can use the knowledge of AI that will benefit both the workers and the industry. Using the knowledge of image recognition and deep learning, we have initiated an automated approach which will generate the description of the process of the work and measure time for work completion. Using the description and timestamp, the management can detect whether there are any faults in workflow and take action immediately such as assisting the employee.

All the data we have worked on have been generated from camera feed i.e. CCTV footage and handheld cameras. We have preprocessed three datasets from different sources. The preprocessing involves breaking the video into single images, extracting features from them using CNN, manual captioning etc. We have used three models for our systems which are VGG19, ResNet and Inception-V3 which are transfer learning models having multiple convolution layers. All these model uses show, attend and tell methods, which is an LSTM based model, to generate captions from the images. In order to yield a work description from the images, graph matching is used where each individual task is described using two nodes and an edge joining them. Finally, we measured the time delay using equation 1.

The first dataset corresponds to a cooking activity captured by us while the second dataset has been obtained from YouTube video feed. In the final dataset, the task is divided into three sections i.e. Sowing, Cutting and Dyeing. We have used the BLEU (bilingual evaluation understudy) score, which uses N-gram language model to compare between generated captions and original captions using NLP (natural language processing), to measure the efficiency of our model. By far Inception-V3 worked better by exceeding other models in scores in all three categories.

2. Related Works

Over the years techniques have been developed to improve the efficiency and safety of production using tested industrial techniques. This paper [1] has analyzed the factors affecting lead time in ready-made garments. The authors have discovered that Bangladesh even being one of the second largest exporters of garments suffered greatly on its lead time. The researchers have proposed a three-stage step strategy that involves the management of production

in due time through supervision. In another work, M M Khatun [2] has tried to fine-tune the production process by diving the work and time management to unit components. The author has suggested time management process should be divided into worker capacity which involves a process-wise calculation of capacity. The researcher [3] has further analyzed her procedure and tried to find a relation to time and motion in the productivity of the garments industry. S. Jadhav et al. [4] have added on the thought of time-based productivity. They have researched the supply of garments piece and suggested that the proper timing of supplied components can improve productivity in the garments industry. Productivity is important for any garments to profit and produce high-quality products. But these products can also be hampered with sudden accidents occurring in the production process. These incidents not only hamper the production rate but also risk workers with injuries that are detrimental to health and life. A manual on safety issues and guideline is written to follow in garments in reference [5]. The guideline outlines the problems faced in the garments industry and how to tackle them. Most of the issues could be handle through management but guidelines of fatigue and rest are mostly ignored or not properly handled. The authors [6] have worked on a process called 6S study. His research has shown how 6S improved productivity by 27 per cent and multi-factor productivity by 13 per cent in the work process. The safety issues suggested in his research puts liability on unorganized workforce and long hours of work in these environments.

These issues of safety and productivity are directly likened to time and its proper utilization in the governance of the factory. Technology such as machine learning and deep learning allows us to optimize the use of time and push the boundaries for these old traditional methods of optimization. As we are moving towards the Fourth industrial revolution technologies such as IoT, Big Data and deep learning are becoming more and more necessary [7]. One of the ways of optimizing time is knowing what task a worker is performing. In one such work [8], it is outlined how data from phone can be used to predict human activity with the help of traditional machine learning. They have collected data from the accelerometer of a phone and has been used to map the task a person is performing. A different approach has been undertaken by other authors [9] where they have used RGB camera and Kinet to detect human activity for surveillance. The model has taken frames of images with Kinect and used them to detect human activity. They have applied SVM to classify and detect the work although it has a limitation as it could only detect the activity of a short time period. On the other hand in the reference [10] they have, instead of only using one medium of data, using two sources of data and by implementing CapsNet and LSTM, the have extracted features from the datasets originating from camera and sensors to accurately classify nine activity in a short time span. However, in this work [11], instead of using any image they have solely used wearable to determine hand gestures, input modalities, motor-powered object detection and so on and this combined system is called ViBand. A. Piergiovanni [12] and co. have followed a different strategy where instead of using supervised methods, they have used unsupervised learning to extract features from raw data collected from suing sensors. Later, they have used RBF-SVM to classify activity. Additionally, to classify human activity temporal attention filters have been used. Any high-level activity can be sub-divided into multiple small events also known as sub-events or temporal events and they can vary in terms of duration. From a video feed, with the help of temporal filters coincided with segment-based CNN and recurrent LSTM, the system learns these small events which correspond to a certain activity. This approach has removed the need for preemptive classification of the work rather the network defines the work based on an image. L. Chen et al. [13] have used attention mechanism to denote images with a caption which describes the image in context to its actual representation. These researchers have shown how images can be represented in language through the use of deep learning without the mapping of features to classification.

Although all the methodology mentioned above provide a compact system of detecting and monitoring human activity, these recognition systems require the user to put on wearable devices, such as Kinect, all the time. These types of classification limit us to segment the work process into categories which hinders the ability to generalize the detection process. We have used deep learning to find a less invasive approach of detection of complex work as well as to detect the work to produce a time frame through which we can apply the already established traditional approach of increasing productivity and safety for the industry and workers.

3. Proposed Methods

To carry out the task few machine learning models have been considered. The main three models that have been implemented on the datasets are VGG19, ResNet and Inception-V3. They all use the mechanism of transfer learning and have convolutional layers in their core system. Additionally, within and outside these models, from end to end, two neural networks i.e. Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) along with Long Short-term Memory (LSTM) and Show-Tell-Attend methods have been used and images will be analyzed by these suggested methodologies to produce comprehensive and meaningful results.

3.1. CNN

One of the benefits of using this network is, unlike other algorithms, less pre-processing is required which makes it much more suitable for image analysis. The other convenient factor is it can capture spatial and temporal dependencies while reducing the number of parameters and reusing weights to understand a sophisticated image. The dimension of the input image is denoted as such, $height \times breadth \times channels(RGB)$ and after performing

convolution operation with a kernel or a filter a new matrix having convoluted features is generated. Whether there is one channel or are multiple channels the kernel strides in such a way which creates squashed one-depth channel convoluted feature output. The first layer extracts low-level features and the following layers extract high-level features which creates a network to understand an image like a human would do. However, the dimensional reduction process is carried out by valid padding and in order to preserve the same dimension or increase its, the same padding is incorporated. Pooling is of two types i.e. max and average and this technique is brought to extract dominant feature and decrements the spatial size which allows using low computational power. Max pooling is preferable since it is noise suppressant. After going through all these processes, the obtained values are flattened into a column vector which is then fed to a conventional feed-forward neural network along with back-propagation techniques.

3.2. RNN

Both CNN and RNN have fundamental similarities which are sharing parameters. RNN has the ability to generate future information based on its past. A general NN remembers things during training and while RNN does the same, additionally, it remembers stuff from previous inputs during producing outputs. Also, unlike NN, RNN can tackle an unlimited number of inputs (not fixed initially) and these input vectors are manipulated by the weights of the inputs and hidden state vectors. Thus, this can give rise to one or more output vectors. Since no fixed input is fed into this model there cannot be any fixed weight for individual input. Thus weights are being shared by each input and to maintain versatility and depth hidden state vectors come into action creating a link between two inputs. This parameter sharing strategy makes it different than conventional NN. Furthermore, to have multi-level abstraction and representation any of the four following methods can be tried; (a) have more hidden states, (b) have more non-linear hidden layers and lay them between input and hidden state, (c) have more depth within hidden states and (d) have more depth in between hidden states and output layer. These techniques can also be found in Bidirectional RNN, Recursive neural network, Encoder-Decoder Sequence to Sequence RNN and last but not least in LSTM with slight variation.

3.3. Transfer Learning

Transfer learning is a concept in deep learning where a model can leverage knowledge i.e. features, weights to another model. Instead of building models from the scratch for similar but new tasks, pre-trained models can be used. For instance, if a model is built to tackle NLP (Natural Language Processing) related task in the English language, the same model might be used for the German language. One such model is VGG16 which, in its core, has convolution neural net (CNN) architecture. It has convolution layers of 3x3 filter with a stride 1 and padding and maxpool layer of 2x2 filter of stride 2. It is used to handle computer vision-based problems such as image classification, image captioning, feature extraction etc.

3.4. Show, Attend and Tell

This methodology is proposed by Kelvin Xu et. al. which is able to automatically detect distinct objects and generate caption from an image. This model contains a CNN-LSTM network. This types of captioning not only require to understand the objects in an image but also the relation between them. previously, fully connected layer system has been used to detect features. But, in this paper a lower level CNN is used to extract features. One the other hand, the LSTM has been used and trained in sequence to sequence manner to generate words. The show, Attend and Tell uses not only CNN but also soft and hard attention techniques. This paper uses VGG architecture which is a pre-trained model to generate feature maps. This later gets converted into vectors. These features are then sent to LSTM with attention model. Attention is a feature of the system that causes the system to find and extract the most obvious features of an image. However, LSTM with attention model has multiple gates. The inputs are modified before going through the next gate. The generated vectors are called context vectors. The formula for generating vectors is: $e_{ti} = f_{att}(a_i, h_{t-1})$, $\alpha_{ti} = \frac{exp(e_{ti})}{\sum_{K=1}^{L} exp(e_{tk})}$ $\hat{z} = \phi(a_i, \alpha_i)$. Predicting next word in the caption is done by this formula, $p(y_t|a, y_1^{t-1}) \propto exp(L_0(Ey_{t-1}+L_h h_t+L_z\hat{z}_t))$. Going back to attention, hard attention is a complex technique where the location with the feature, is sampled from a multi-nomial distribution. As opposed to this soft attention is much more trivial.

3.5. VGG19

VGG19 consists of 19 layers which are 16 convolution layers, 3 fully connected layers, 5 maxpool layers and 1 SoftMax layer. It is a variant of VGG model and the other variants are VGG11, VGG16 etc. A fixed-size RGB image (224x224) is given as input which generates a matrix of size (224,224,3). The core Kernel is of size 3*3 with a stride size of 1 pixel. Usually, the max-pooling is conducted over a 2x2 pixel windows with stride 2. This model is well suited for facial recognition.

3.6. ResNet

Residual Network (ResNet) not only has deep convolutional layers but also it introduces the mechanism of "identity shortcut connection". The model can deal with hundreds and thousands of layers without degrading the network's performance. The identity mapping allows the system to look over certain inactive layers which resolve the notorious vanishing gradient problem. Some of the variants of ResNet are ResNeXt and DenseNet and this model works well in object detection and face recognition.

3.7. Inception-V3

Inception-V3 is a CNN based network that uses an auxiliary classifier to propagate label information below the network. It also uses batch normalization for layers in the sidehead. Some of the features of these models are label-smoothing, convolutional factorization (7x7), dense connection and so on. The main usage of this model consists of image classification, quantization, adversarial attack etc.

3.8. BLEU Score

The model is evaluated using the BLEU(Bilingual Evaluation Understudy) score. Initially, BLEU has been developed to measure machine translation. But, in the paper, it is used to compare predicted captions and reference captions. The result excelled the result obtained from the previous model.

4. Methodology

Initially, we have visited one of the industries in Bangladesh to get a better understanding of the working environment. But we have not been able to collect any data from there due to internal industry policy. Therefore, of the three datasets we have used, one of them has been recorded by us. The methodology consists of image extraction, feature extraction, captioning, graph generation and model formulation.

4.1. Industry Visit

Our visit to a garments manufacturing plant has led us to get a better grasp of the issues faced by the labour. We have seen various machinery being manipulated by a worker for long hours. The factory is huge and has a lot of areas to cover in the respect of supervision and monitoring. Therefore, in this huge area, it becomes a hassle for top management to properly care for workers to check for fatigue or manage delay in production as it happens. A good monitoring system that can inform supervisors quickly of the delay in production occurring in real-time, worker working overtime or breaking standard protocol is bound to boost the work environment.

Figure 1. Factory Visit

4.2. Dataset

We have used three data-sets around which we have built and implemented our model. All three videos demonstrate a process and image extraction rate from those videos were 25 frames per second (fps).

4.2.1. Dataset-1. The first data-set has been extracted from a video which has been recorded in a generic household kitchen. The video consists of a trivial illustration of poaching an egg on the pan. Within the frames, the following are shown: a hand, a pan on the stove, egg in different phases and ingredients. The process involves putting oil on the pan and heating it, breaking the egg and placing it on the pan, pouring salt and usage of apparatus. The recorded video length is three minutes and we have extracted 5400 frames from it. We basically, recorded it to build a test run for the neural network. We used it to pre-formulate the idea of the model.

4.2.2. Dataset-2. The second video has been collected from YouTube streaming service. It is a 5-minute video and we have extracted 9000 frames from it. In the video, it is shown

that a person performing a simple task in an industrial space. He is operating various tools from moving one place to another. The process contains diverse mechanical tools and the person is picking and placing them throughout the video.

4.2.3. Dataset-3. The third and final video is another YouTube feed whose length is 10 minutes. This video feed contains the whole procedure of garments manufacturing [14]. A total number of 18000 frames have been extracted from the video. The frames then subdivided into three equal portions, each corresponds two unique tasks. The first task is sewing, the second one is cutting and the last one is dyeing.

4.3. Pre-Model Formulation

Videos are basically images changed frame by frame [15]. We have fragmented the video stream into video frames. Figure 2 represents the generalization of the video frame procedure. The objective here is to generate a

Figure 2. Video feed to video frames

description of a complex work that has been retrieved from a video feed. To accomplish that, initially, we have captured a video of a poached egg to determine the capability of the network. The process of poaching an egg consists of eight steps where two particular steps are identical to another two. The duration of each step is 10 to 12 seconds long so that we could avoid over-fitting. At first, the image has been extracted from these video feeds. Then we have used transfer learning methodologies to extract features from images [16] [17]. After extracting the features from images we have captioned the images based on the work they represented.

4.4. Image extraction

To [1] extract the images, we have used one of the most popular libraries known as OpenCv [18]. As such, we have generated frames from video using this library. A total number of 650 frames has been extracted from the video. Figure 3 represents one of the frames that has been extracted.

4.5. Feature Extraction

The images have been passed through a CNN to extract features that would represent an important part of the images. In addition, we have visualized the feature map to fully understand what part of the images CNN is learning. In Figure 4 we showed the feature maps and what features were collected from the image.

Figure 3. A frame of the complex work showing egg is being hold by a person

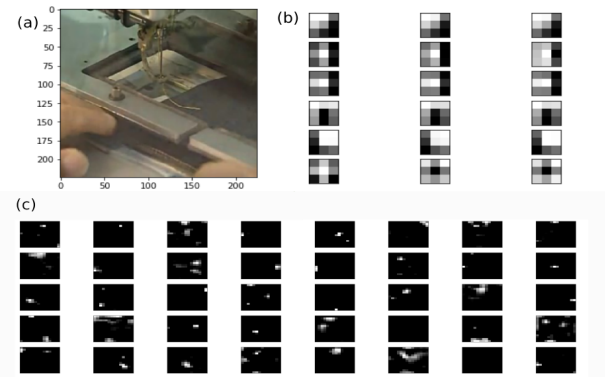

Figure 4. Feature map and feature extraction

In Figure 4(a), it is shown that the picture of sowing and figure 4(b) represents the feature maps. The feature maps have extracted information like edges, shape and corner represented in the image. In Figure 4(c), we have shown the parts of the images that have been used as important features.

4.6. Image captioning

In the beginning, unique ID has been created for each image and against each ID two captions were assigned. Since a bunch of images represents a particular phase captions generated for those images are same. Now, to extract the features from images transfer learning models are called upon which are pre-trained models. They cause faster computation and less memory consumption. The focus here revolves around returning the dictionary which contains image features of internal layers and saves it to a file.

The captions that have been generated earlier are loaded from a file and at the same time unique frame identifier is returned based on the description. To alleviate the difficulty

working with the description, they have been tokenized and cleaned. Tokenization [19] is a procedure where description turns into individual words or vocabulary and cleaning process involved making the description in lowercase and remove redundant words, numbers, punctuation marks and articles. This newly created dictionary is then saved to a new file to be loaded later on.

At the beginning of our training, photo features are loaded from the file that has been previously created. Along with that, the description has been loaded as well. In order to map description to a unique integer, value tokenization has been performed on the training set before feeding it to the model. Two strings, 'startseq' and 'endseq', have been used to mark the start and end of a caption. The sequence of words has been generated based on the parameters i.e. highest length of the sequence, tokenizer and both descriptions (image and text).

As for the model, Long Short-Term Memory, Recurrent Neural Network (LSTM-RNN) with attention mechanism has been implemented. The text, which has been encoded as integers, is fed to one part of the model and the image is fed to another part. This has created a probabilistic distribution of the caption to be matched with an image [13].

After all that, a test dataset has been formed containing new photos. Then, the model has been called recursively to generate captions against the test dataset. A mapping between the caption and image has been observed to see if the model can successfully generate the description. Figure 6 and Figure 7 illustrates how well captions have been generated for a new dataset.

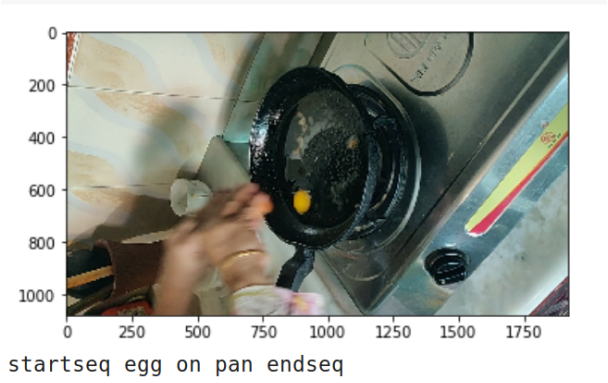

Figure 5. Successful prediction of an image from test dataset

4.7. Graph generation

Before generating the graph, the descriptions have been split into noun and preposition. We have a list of preposition and a dictionary containing pairs of nouns to be connected by any of the preposition from the list. Each node of the graph represents a noun and each directed edges the preposition. The number associated with preposition indicates the position of the word in the sentence. For all the similar captions there is the only graph mapped to it, illustrating a certain step. Figure 8 demonstrate the outlook of the graph for a certain caption.

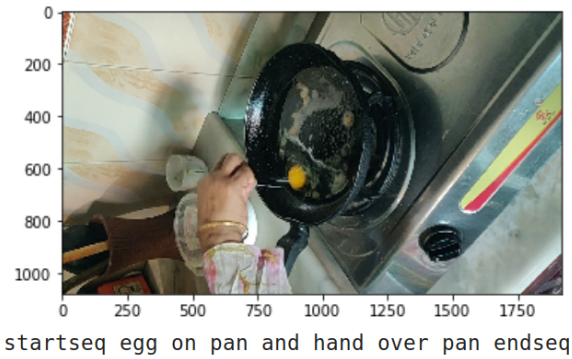

Figure 6. Successful prediction of another image from test dataset

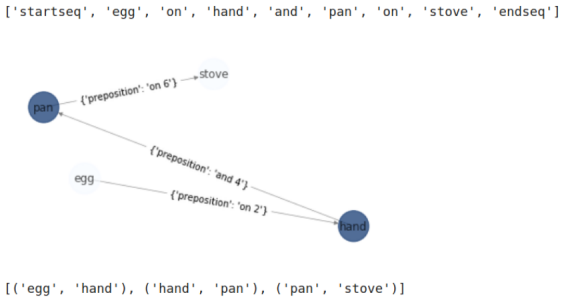

Figure 7. Directed graph of a caption

Our objective is to match the generation of the graph in a continues loop to determine the work process and how long it was performing. Even though, the model tends to give near similar result in some instance it is necessary for us to generate a graph and match the two instances to determine a similarity of the caption per frame.

4.8. Model Finalization

In the pre-model formulation, we have tried to find the possibility and applicability of the image captioning technique. We have been successfully able to achieve a good decisive result. We have trained and applied our network on a small video from YouTube to test the comprehensiveness of the deep learning model. We have converted the video feed into video frames. The dataset has been divided into train, validation and test set. After training, we have decided to test our model to see the caption and determine its capability.

Our main goal in this stage is to observe if the work of placing an object in a different place could be captioned. As such, we have successfully able to caption the states of activity in the video. In Figure 9 we have shown the caption generated and it is successfully able to caption the activities the person has performed.

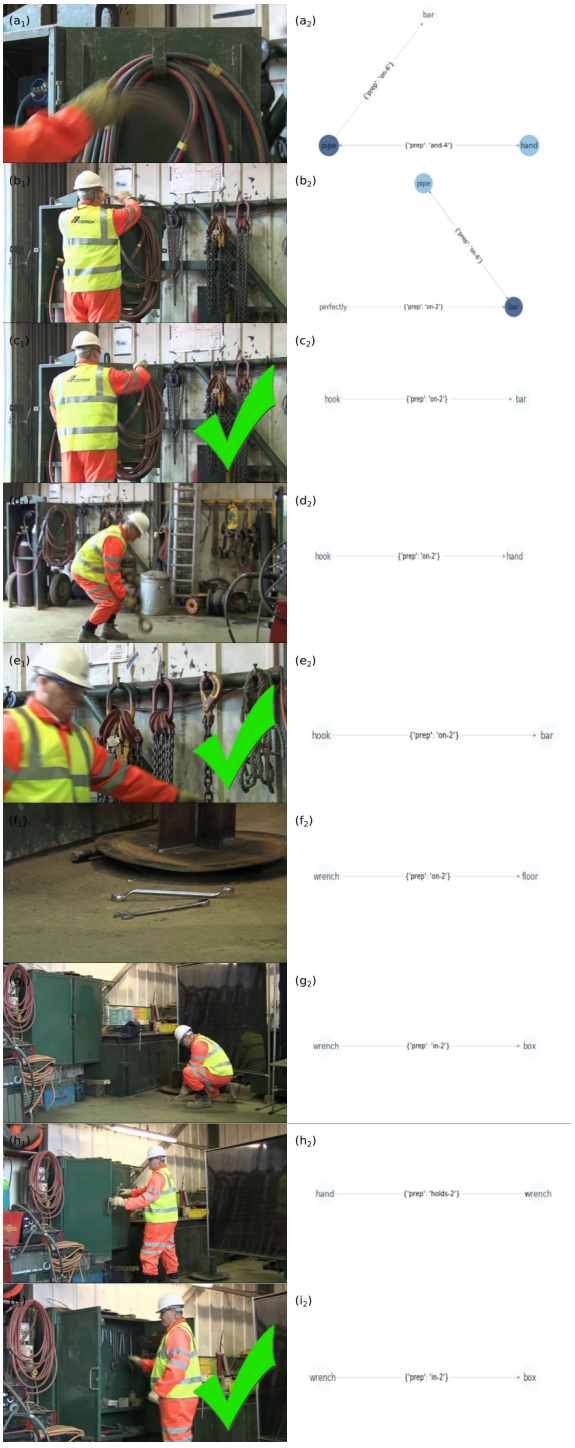

Figure 8. Generated graphs using the frames of a video

4.9. Derivation of the Model

Initial trials have produced good prediction on the complex work. As such, we have derived a model that has generated caption to determine the activity, which involves how a worker is performing and how it integrates itself with the working environment to answer the question in relevance with industry-established methods. In addition, *TimeDelay* helped us to flag key parts of the work process to automate the garments evaluation process. Figure 9 is the complete illustration of the model.

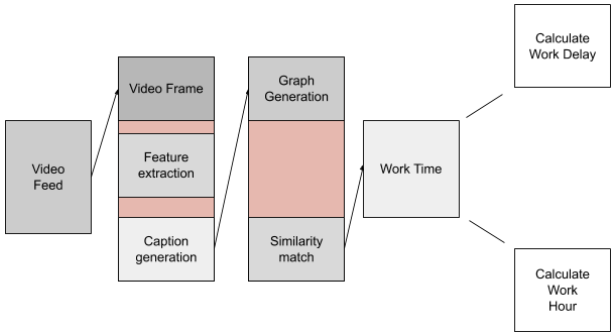

Figure 9. Automation Model

The *WorkTime* is calculated while generating a work description by the means of graph generation. For, consecutive image frames the descriptions are expected to be similar since a certain portion of the video denotes a certain subtask. As long as image frames and descriptions are the same we consider that portion a single sub-task and calculate how long does it take to complete that task. Whenever a new frame is inserted, the description changes accordingly and thus we start from T=0 to calculate the time for the new sub-task. In this manner, we calculate the time for every smaller task.

$$TimeDelay = OptimumTime - WorkTime \quad (1)$$

In equation(1) *TimeDelay* indicates the lag in performing an activity while *OptimumTime* is the known value for performing a work on time. In addition to that, *WorkTime* is the time a worker has been actually involved in doing the activity. *TimeDelay* is a good indicator of determining whether a person is doing overtime or not. As such, doing overtime creates fatigue and distress among workers, making them sluggish and less responsive to the environment. Moreover, workers handle complex and dangerous machines to perform their tasks, which if not properly managed in full awareness would create serious life-threatening accidents.

5. Result

To evaluate the model we have considered our final dataset. In this dataset the task is segmented into three activities which are sewing, cutting and dyeing. Nevertheless, we only focused on the activities where accidents and delays are frequent due to poor supervision [20].

5.1. Sew

Sewing involves piecing the components together which is demonstrated in Figure 10. We implemented our designed

Figure 10. Frame of a video in sewing

model which follows the same procedure for each task. Figure-11 shows the efficiency of the overall detection of sewing task. Additionally, Figure-11 shows the outlook of

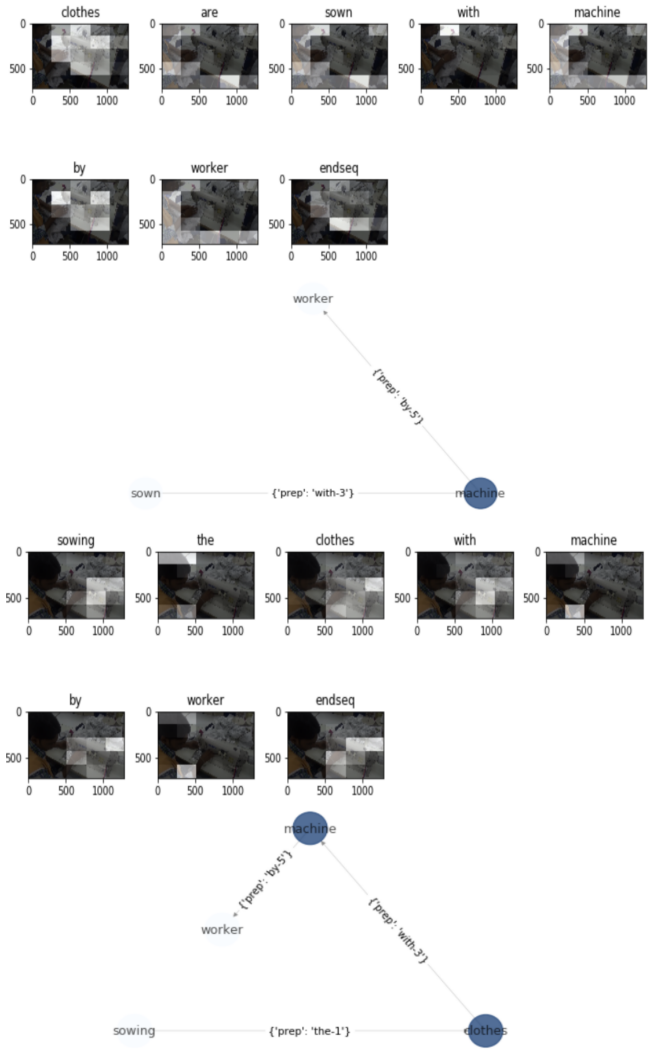

Figure 11. Caption for sewing

the successful captions and description generation. Each image is equipped with one word only. For example, the caption ”clothes are sown with machine by worker” contrasts well with the graph description which is ”sown with machine by worker”. We calculated the time by multiplying the video frame number with its FPS. We knew the optimum time for this activity, we calculated the lag in sewing using the equation 1.

5.2. Cut

Cutting is another essential procedure in the industry as revealed in Figure 13. We have evaluated this activity using our model as Figure 12 represents the overall detection success ratio of cutting. To analyze the parallel contrast of

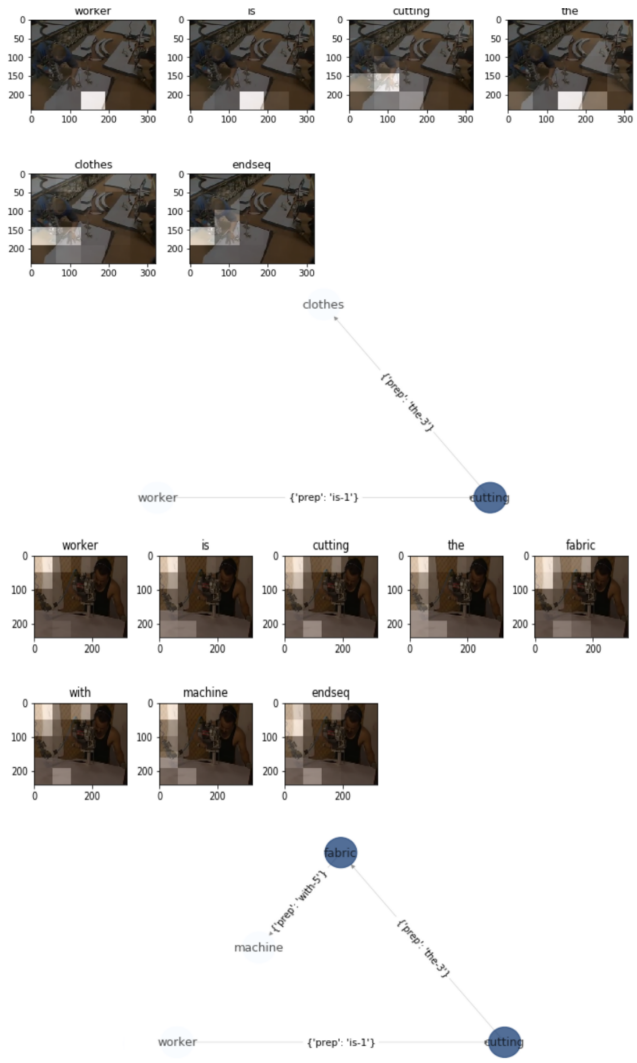

Figure 12. Caption for cutting

the captions and descriptions we can consider the following example wherein one side we have this caption ”worker is cutting the clothes” and in another side, the description is ”worker is cutting the clothes” which is an exact match.

Figure 13. Frame of a video in cutting

Like before, we have been able to programmatically determine how long worker is performing the task by multiplying the graphs generated with FPS and determined the delay in his activity using equation 1.

5.3. Dye

The final task we have considered is dyeing which is necessary to give clothes its vibrant colour as illustrated in Figure 14. Similarly to cutting and sewing, we have detected this work using our model in the same manner. Figure-15 represents how well the system can detect dyeing. As opposed to caption "spot the clothes for coloring" we have exact same description generated by the graph.

Figure 14. Frame of a video in dyeing

One thing is noticeable here is the same caption or description can be described in a various manner. For example, in the cases of cutting active and passive expressions have been used and the corresponding graphs have been generated accordingly. It is also worth mentioning that the nodes can contain nouns or verbs or any other words as well which emphasizes on the efficiency of our model to detect an action. This type of result would act as a supplement for the solution to this problem of management in the garment industry.

5.4. Score

The evaluation of the model is done with the help of the BLEU score. This scoring system generates a score from 0

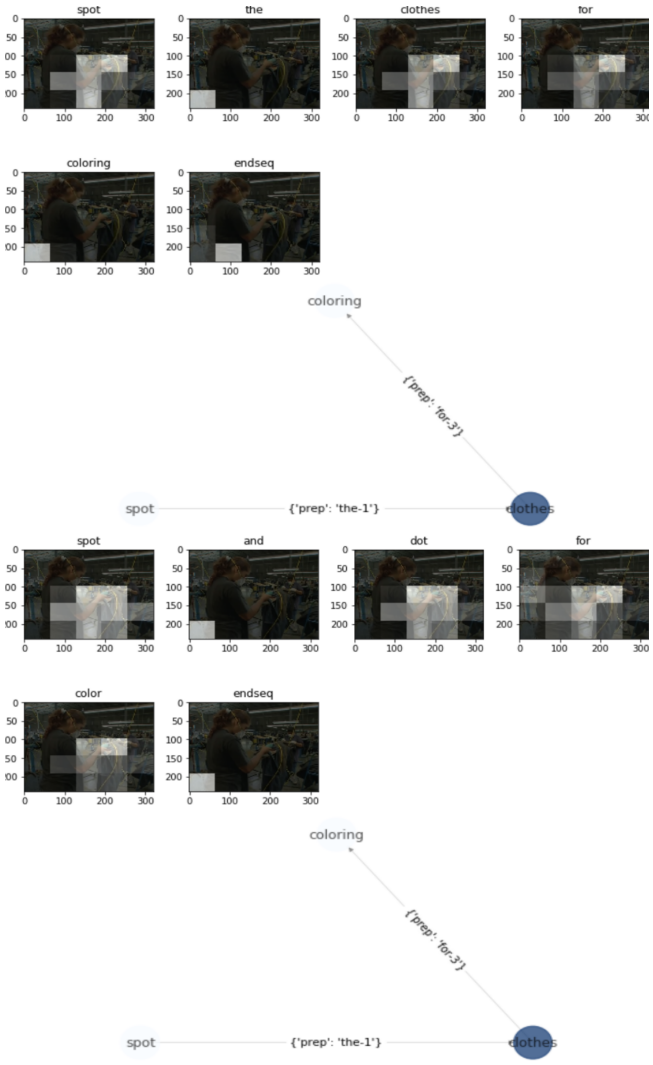

Figure 15. Caption for Dyeing

to 1 by contrasting two sentences describing the same task. Therefore we have made three sets of such sentence pairs. In set-1, we are comparing 1 vs 1, in set-2 we are doing it for 2 vs 2 and in the final set we have made a 3 vs 3 pair to do the contrasting. The first two pairings have produced better results than the last one. The score of the first two pairings, in most cases, above 0.35 where the average score for the last pairing is roughly 0.2.

As mentioned earlier, we have used ResNet, VGG19 and Inception-V3 to extract images and among all of them, Inception-V3 showed the most promising results. The following Table 1 illustrates the results of caption generation. BLEU score 1, 2 and 3 stands for scores generated from all the three pairings respectively.

Network Evaluation			
CNN Name	BLEU Score 1	BLEU Score 2	BLEU Score 3
VGG19	0.360800	0.349552	0.187343
ResNet	0.375260	0.360864	0.201157
Inception-V3	0.386725	0.379816	0.215263

Table 1: A comparison of different models operative on same set of data

From Table 1 it is evident that Inception-V3 produced a better score for all three cases. The score for the first two cases is above 0.37. A notion can be derived from this that its architecture [21] played a crucial role for such performance. ResNet also produced good results where the score is pretty close to those of Inception-V3. Lastly, VGG-19 performed poorly among all three having a score of around 0.35.

However, while working with the second dataset we have encountered with visual noises. For instance, some of the images in Figure 6 i.e. c_1, e_1 and i_1 noise is present (graphical tick mark). This is because we used a video that is not recorded by us. In real-life cases, this type of noises can easily be avoided and performance can be improved even further. Also, we only generated graph captions for those which are relevant to the event. For instance, if the man putting a hook on the bar then the captions will be like {hook, on, bar} with arrows in between. Redundant information is ignored.

6. Conclusion

Our model has produced expected results without the intervention of any type of wearable devices. This non-evasive approach of detecting complex actions in the sphere of industry shows promises for a better work environment. Transfer learning methods like VGG19, ResNet and Inception-V3 have played a crucial role in these types of activity recognition along with the show, attend and tell method. This proposed method is able to provide safety and security to the factory worker as well as improve productivity. The method highlights how it is possible to use deep learning in the industry to work as a safety net to handle imposing accidents caused due to overwork or fatigue and production hindrance that could arise for not following proper rules and regulation of the industry. Even so, the model could improve further if the data set was more rich and diverse. Image recognition is cheap and less invasive, giving it more room for application in industry 4.0. The method has been tested on the real-world scenario to show the possible usage of the technique. We plan to improve our technique and further extend its functionality by adding an additional source of features and data. We hope such addition might increase the reliability of the model to be used in an even larger scale.

Acknowledgments

References

[1] S. Kader and M. M. K. Akter, "Analysis of the factors affecting the lead time for export of readymade apparels from bangladesh; proposals for strategic reduction of lead time," *European Scientific Journal*, vol. 10, no. 33, 2014.

[2] M. M. Khatun, "Application of industrial engineering technique for better productivity in garments production," *International Journal of Science, Environment and Technology, II*, pp. 1361–1369, 2013.

[3] M. M. Khatun, "Effect of time and motion study on productivity in garment sector," *International Journal of Scientific & Engineering Research*, vol. 5, no. 5, pp. 825–833, 2014.

[4] S. Jadhav, G. Sharma, A. Daberao, and S. Gulhane, "Improving productivity of garment industry with time study," *International Journal on Textile Engineering and Processes*, vol. 3, no. 3, pp. 1–6, 2017.

[5] J. C. Hiba, *Improving working conditions and productivity in the garment industry: An action manual*. Intl Labour Organisation, 1998.

[6] M. Ahmed, T. Islam, and G. Kibria, "Study on 6s method and improving working environments in the garments industry," *International Journal of Scientific & Engineering Research*, vol. 9, no. 3, pp. 737–754, 2018.

[7] E. Manavalan and K. Jayakrishna, "A review of internet of things (iot) embedded sustainable supply chain for industry 4.0 requirements," *Computers & Industrial Engineering*, vol. 127, pp. 925–953, 2019.

[8] A. Taha, H. H. Zayed, M. Khalifa, and E.-S. M. El-Horbaty, "Human activity recognition for surveillance applications," in *Proceedings of the 7th International Conference on Information Technology*, pp. 577–586, 2015.

[9] K. V. Bhaltilak, H. Kaur, and C. Khosla, "Human motion analysis with the help of video surveillance: a review," *International Journal of Science, Engineering and Computer Technology*, vol. 4, no. 9, p. 245, 2014.

[10] Y. Lu and S. Velipasalar, "Autonomous human activity classification from ego-vision camera and accelerometer data," *arXiv preprint arXiv:1905.13533*, 2019.

[11] G. Laput, R. Xiao, and C. Harrison, "Viband: High-fidelity bio-acoustic sensing using commodity smartwatch accelerometers," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pp. 321–333, 2016.

[12] A. Piergiovanni, C. Fan, and M. S. Ryoo, "Learning latent sub-events in activity videos using temporal attention filters," *arXiv preprint arXiv:1605.08140*, 2016.

[13] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5659–5667, 2017.

[14] T. Kütüphane, "Garment construction- jean construction," 01 2016.

[15] A. M. Tekalp, *Digital video processing*. Prentice Hall Press, 2015.

[16] Y. Wu, X. Qin, Y. Pan, and C. Yuan, "Convolution neural network based transfer learning for classification of flowers," in *2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP)*, pp. 562–566, IEEE, 2018.

[17] M. Leonardo, T. Carvalho, E. Rezende, R. Zucchi, and F. Faria, "Deep feature-based classifiers for fruit fly identification (diptera: Tephritidae)," pp. 41–47, 10 2018.

[18] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.", 2008.

[19] P. Mcnamee and J. Mayfield, "Character n-gram tokenization for european language text retrieval," *Information retrieval*, vol. 7, no. 1-2, pp. 73–97, 2004.

[20] S. Akhter, A. Salahuddin, M. Iqbal, A. Malek, and N. Jahan, "Health and occupational safety for female workforce of garment industries in bangladesh," *Journal of Mechanical Engineering*, vol. 41, no. 1, pp. 65–70, 2010.

[21] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019.