EmoSense: Illuminating Emotions with CNN and Transfer Learning

Muskaan Jhunjhunwala
20BEC0168
Vellore Insitutute of Technology
Vellore, India

Ananya Singh
20BEC0476
Vellore Insitutute of Technology
Vellore, India

Reshma Ramakrishnan
20BEC0431
Vellore Insitutute of Technology
Vellore, India

Pranshu Khatri
20BEC0437
Vellore Insitutute of Technology
Vellore, India

Archie Jain
20BEC0732
Vellore Insitutute of Technology
Vellore, India

Gauri Pradeep
20BEC0307
Vellore Insitutute of Technology
Vellore, India

Udayan Borah
20BEC0231
Vellore Insitutute of Technology
Vellore, India

Sabyasachi Mohanty
20BEC0168
Vellore Insitutute of Technology
Vellore, India

Raghav Biyani
20BEC0589
Vellore Insitutute of Technology
Vellore, India

Radhika Toshniwal
20BEC0475
Vellore Insitutute of Technology
Vellore, India

Abstract—**Facial emotion recognition (FER) is an important topic in the fields of computer vision and artificial intelligence owing to its significant academic and commercial potential. This project aims to build a meaningful Machine Learning model that can classify a given image into a fixed range of emotions such as neutral, angry, happy, sad, etc. The Machine Learning algorithm that'll be used here is Convolutional Neural Network or CNN for short. It mimics the human brain (neurons to be specific) and has multiple layers from input to output that can learn different aspects of a given input. The goal is to maximize the accuracy achieved at the end of training. Such a model can be used for various real-life problems such as assisting blind people during social interactions, during therapy sessions, etc.**

Keywords—Machine Learning**,** Convolutional Neural Network, facial detection**,** emotion detection, computer vision.

INTRODUCTION

Facial expressions can be used as a vital identifier for what a person is feeling at the moment. Most of the times, the emotion/expression that people show on their face is a direct non-verbal way of expressing the emotion they feel. There are various algorithms that can be used to arrive at a solution for this problem, like ANNs and its variations, CNN etc.

We're not gonna work with ANN as the computational complexity of ANN is higher than CNN so it won't be feasible to train them easily when the dimension of the dataset is larger. To add more, CNNs are better with pattern recognition within images compared to ANNs and our input is an image in which we have to detect patterns. CNNs also have comparatively less number of parameters to train(like weights) and lesser number of parameters means lesser will be the overfitting. So it is important to take such precautions to make sure our model performs good in a real world scenario where it would be getting never-before-seen inputs.

The first layer would take the input and convolve it using matrix dot product. Such a matrix is called a filter. The next layer pools the output of this layer and so on, till the last layer, which is the output layer, where the loss function associated with each class/category of output is calculated. We usually select the class with the highest probability as the predicted output class

LITERATURE SURVEY

[1] The paper talks about how using the eyebrows and mouth as an anchor point in grey scaled images of faces is helpful for detecting emotions such as Happy, Sad, Surprised, Angry with 80% accuracy. This technique of the extracting intransient facial feature does not require any manual intervention like manual assignment of feature points to detect the various features of face and use that to read emotions

[2] The paper talks about how the perceived waiting time of a customer in a fast food restaurant. The survey was on Malaysian restaurants but it’s generally the same over everywhere. Perceived waiting time depends upon how long the customer waits for his/her food to arrive after ordering, is he alone or has some company, whether he is free or is pre-occupied with other works such as chatting or speaking over a call, etc. The research arrived at a conclusion that surprisingly the perceived waiting time has a positive correlation with customer satisfaction as many people are willing to wait for good food, though it is inconsistent. But waiting time above a point has led to dissatisfaction. We can predict this dissatisfaction and prevent the customer from getting angry or leaving the restaurant without getting the food they ordered by using our proposed model. We can prioritize the orders of such customers so that the identified problem can be addressed using the proposed model

[3] The paper has proposed a gloves, called VibroGlove to help blind or visually impaired people understand the emotion of other people. A blind or visually impaired person cannot judge the emotion or facial expression of another person that is in front of him. This might make it harder for them to have conversation with the other person or to work

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

with them. The gloves can detect the other person’s expression and use haptic interface(by making use of Vibrotactors) to convey that emotion to the blind or visually impaired person. This research paper is relevant to our topic as our proposed model can help detect other’s emotion by seeing their face. We can make use of haptic or use speakers to convey the emotions of other person to the blind or visually impaired person.

[4] This thesis describes in depth how the identification of a real time face can be in order to create a safety alert framework for the workplace. The machine learning algorithm Haarcascade classifier was used to build the given four distinct classes for identification of security equipment and eventually identify the faces in both images and video using python open CV.

It also talks about the Face recognition history + relation with ML, as Face detection is the initial and important of face recognition. In face recognition process, after detection of face or image, the basic question that arises is to “whom this face or image belong”. The facial recognition process solves this question by evaluating through four stages like 1. detection 2.feature extraction 3.tracking and 4.recognition.

They have used the Given Face Recognition process, to successfully recognize faces. Many experiments have been wiped out of the field of face recognition, but the accuracy rate is smaller compared to other person biometric details such as fingerprints, expression, eyes, palm geometry, retina etc.

This Research paper has used and Described the KNN algorithm, stating, one of the basic classification algorithms in machine learning is known to be the k-NN algorithm. In machine learning, the k-NN algorithm is considered a well monitored type of learning. It is commonly used in the sorting of related elements in searching apps. From this we also came to know the importance of this Classification algorithm.

We also came to know about the Pre-Processing techniques used. Elbow for classifier: The Elbow curve is created to train the KNN module for different N-neighbour values and finally pick a value that gives us an error. It also stated how they have used OpenCV with Face recognition, Along with KNN, OpenCV, we also came to know about the Regression, KNN Result, also about How they have used, HaarCascade Classifier, stating: This is a technique focused on machine learning, during which a course work is prepared from a broad measure of positive and negative images. It's normal to identify issues in various images. It's a pretrained facial data model and it's popular to identify faces. Using these Machine Learning techniques, the Research Papers showed how they can detect faces, with great accuracy.

[5] This paper explores the ways, including text and emotion history, to detect Depression with the help of Artificial Intelligence, and Machine Learning. At first, they have given us the symptoms and social problem to be solved, which is Depression and its further impact, including-> Depressive Mood, Loss on interest in activities, Suicidal thoughts, feeling of worthlessness or hopelessness, worsened ability to think and concentrate etc.

The structure of the Research Paper is: Description of the classifiers used in the implementation, methodology of the research paper, with explanation and discussion, which states how well this research paper has been done. They have used dataset of about 10,000 Tweets and they have used the ratio of 80:20 split of training and test dataset.

[6] This Research Paper helps tackle a current social problem, which can be used to tackle many futuristic issues. Images and videos have become omnipresent on the internet, which has encouraged the development of algorithms that can analyze their semantic content for various applications, including search and summarization. They have provided empirical evaluation of CNNs on large-scale video classification using a new dataset of approx. 1 million YouTube videos belonging to 487 classed. They have studied multiple approaches for extending the connectivity of a CNN in time domain to take advantage of local spatio-temporal information and suggested a multiresolution, foveated architecture as a promised way of speeding up the training.

They have used multiple approaches for extending CNNs into video classification on a large-scale dataset of 1 million videos with 487 categories (which we release as Sports-1M dataset) and reported significant gains in performance over strong feature-based baselines, and highlighted it.Unlike Images, video classification is far more complex as videos vary widely in temporal extent and cannot be easily processed with a fixed-sized architecture., and they have successfully done it. Using this below approach

They have also infused Time information using CNNs giving priority to each frames along with Early, Late and slow Fusions which combined to become a Multi-resolution CNNs. The Sports-1M dataset consists of 1 million YouTube videos annotated with 487 classes. And the research paper studied the performance of all of it, with the created CNN model.

[7] This paper deals with emotion detection and sentiment analysis of images. Recognizing that analysing content from social media websites and/or photo-sharing websites like Flickr, Twitter, Tumblr, etc., can give insights into the general sentiment of people about any topic we wish like an upcoming election or a concert or just about anything. Also, it would be useful to understand the emotion an image depicts to automatically predict emotional tags on them - like happiness, fear, etc. They aim to predict the emotional category of an image into 5 distinct categories which are Love, Happiness, Violence, Fear, and Sadness. They try to achieve this by fine-tuning 3 different convolutional neural networks for the tasks of emotion prediction and sentiment analysis. They have collected data from Flickr and have experimented with various techniques and classification methods like SVM on high level features of VGG-ImageNet,

fine-tuning on pretrained models like RESNET, Places205-VGG16 and VGGImageNet.

Here are some of the experiments performed - using a pretrained neural network to get the feature representation of their dataset and then use this representation as an input to SVM and classify the data using one vs all SVM classifiers, fine tuning the SVM classifiers, performing sentiment analysis on their data to understand their results better, fine tuning a model which was pretrained on a scene-based database called Places205, fine tuning with a pretrained ResNet50 etc.

In all of the fine-tuning experiments, it has been observed that standard data augmentation techniques like mirroring and random cropping of images increase accuracy. Having a higher learning rate for the later layers also yields a better accuracy. Finally, their results show that deep learning does provide promising results with a performance comparable to some methods using handcrafted features on emotion classification task, and also a few methods using deep learning for sentiment analysis.

[8] This paper deals with a method for facial recognition and emotion detection using support vector machines. Here, in order to deal with the dimensionality of the images/ frames which play an important role in ML problems, they have used principal component analysis (PCA) in order to reduce the dimensionality of the images (converting high dimensional space to low dimensional space) and linear discriminant analysis (LDA) which computes the group of characteristic features that normalizes the different classes if image data for classification.

The input image is subjected for face detection to detect the face. The detected faces are then extracted from the image and these images are saved as a database. Saved images are used to compare with the input image. The matching of input image is performed to identify the user's identity. The recognition result gives identification of the person. The step-by-step architecture used is also given below:

Facial features such as eyes, nose, lips and face contour are considered as the action units of face and are responsible for creation of expressions on face, are extracted using open-source software called dlib. SVM classifier compares the features of training data and testing data to predict any emotion of the face. Multi-SVM classifier is used for classification of different emotions. Finally, it is also observed that accuracy of both face recognition and emotion detection can be increased by increasing the number of images during training. The detection time is significantly less and hence the system yields less run-time along with high accuracy

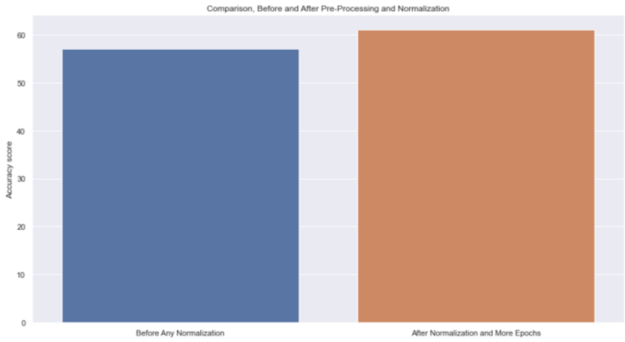
[9] This paper deals with face detection and recognition of natural human emotions using Markov random fields. In the first part, they implement skin detection using Markov random fields models for image segmentation and skin detection where a set of several colored images with human faces have been considered as the training set. It detects the skin color using MRFs estimation in YCbCr space. The second part includes eye and mouth detection and extraction. It uses the HLV color space of the specified eye and mouth region. The third part detects the emotions pictured in the eyes and mouth, using edge detection and measuring the gradient of eyes' and mouth's region figure.

The proposed face detection algorithm contains two major modules: firstly, face localization for finding face candidates and secondly facial feature detection for verifying detected face candidates. Here, the algorithm first detects skin regions that possibly contain a human face. The skin detection algorithm is a segmentation technique that first transforms the image from the RGB to YCbCr color space and the skin detection algorithm is based on statistical image processing model using Bayesian estimation.

The proposed algorithm was evaluated on several image databases, the color images used for assessment have been taken under varying lighting conditions and with complex backgrounds, images contain multiple faces with variations in position, scale, orientation, and facial expression and the average size of each image is 400 X 500 pixels. The final experimental findings s proved that the algorithm can detect multiple faces of different sizes with a wide range of facial variations in an image. The proposed system does not depend on the orientation of the face, and half profile face views can be detected as long as one eye and mouth are visible, or at least part of them and faces can also be detected in the presence of facial hair.

METHODOLOGY

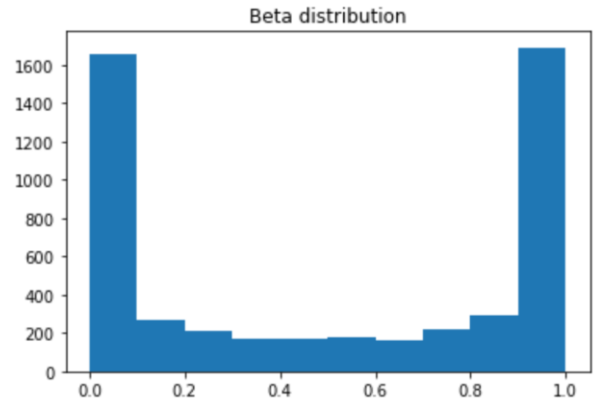
Initially just like the research articles, we also Applied the basic Deep learning techniques and CNN to get an initial comparable model. Then after comparing it with the previous one, it came out to be same.

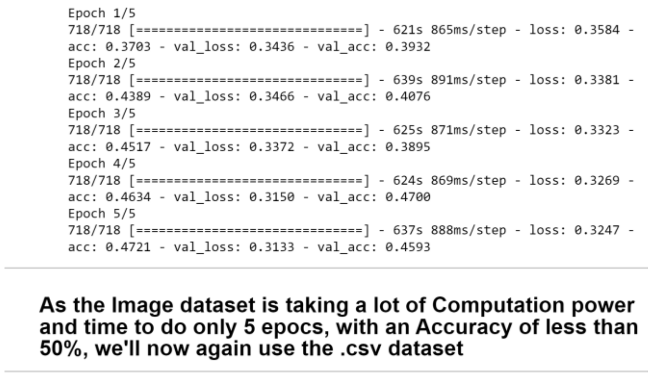


Then we did few Pre-Processing techniques, along with the Batch Normalization. And were able to increase the accuracy upto 5%. i.e. from 57 to approx. 62%.

As the dataset originally is in Image format, So for CNN it required a lot of computational power, so we used the pre-saved csv formal. But since we were instructed to also use it in the original Format, so using MixUp argument we used the original Dataset.

0



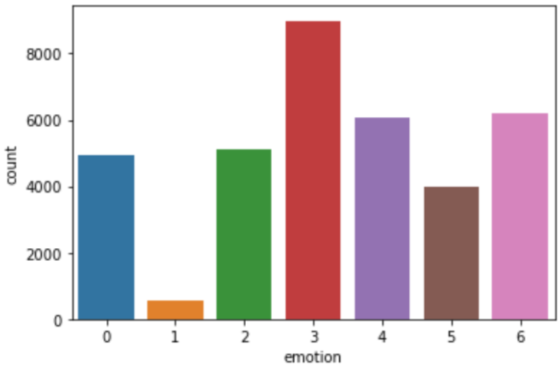


```
Epoch 1/5
718/718 [==============================] - 621s 865ms/step - loss: 0.3584 -
acc: 0.3703 - val_loss: 0.3436 - val_acc: 0.3932
Epoch 2/5
718/718 [==============================] - 639s 891ms/step - loss: 0.3381 -
acc: 0.4389 - val_loss: 0.3466 - val_acc: 0.4076
Epoch 3/5
718/718 [==============================] - 625s 871ms/step - loss: 0.3323 -
acc: 0.4517 - val_loss: 0.3372 - val_acc: 0.3895
Epoch 4/5
718/718 [==============================] - 624s 869ms/step - loss: 0.3269 -
acc: 0.4634 - val_loss: 0.3150 - val_acc: 0.4700
Epoch 5/5
718/718 [==============================] - 637s 888ms/step - loss: 0.3247 -
acc: 0.4721 - val_loss: 0.3133 - val_acc: 0.4593
```

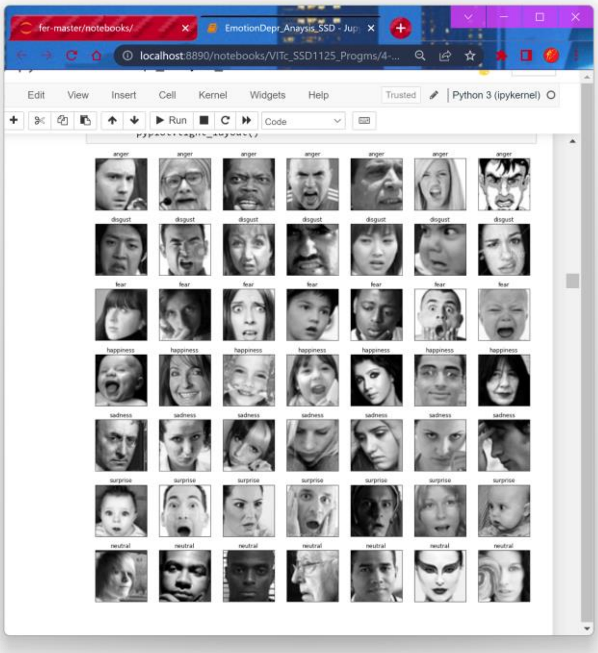
As the Image dataset is taking a lot of Computation power and time to do only 5 epocs, with an Accuracy of less than 50%, we'll now again use the .csv dataset

As the Image's epochs were taking a lot of time, so we again tried the csv file.

To increase the Accuracy more, we even added VGG16 layer, which was the original method with which the FER2013 Competition was won by the team. But still we were only able to increase it for few % more not upto our expectations. So, then we increased the CNN Model's layers from 3 to approx. 5.

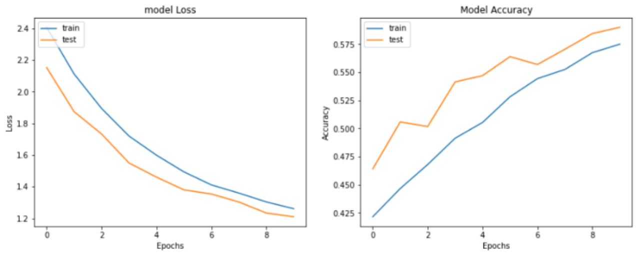


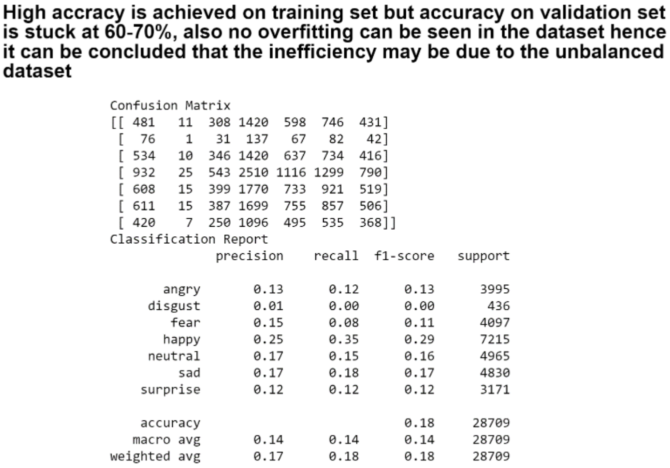
Another thing which we came to know about is that, in the Competition, many have Removed the data which is least available to increase their accuracy to even 80%. But that means we'll have one emotions less, so have not used this loophole to get accuracy to a better value, just by omitting it. So, we have used all the 7 emotions in our Project.



We also used algorithms that avoids Overfitting, so that in the Epoch the train accuracy won't be too much as compared to the real accuracy. Added Deep CNN as well, and using ELU instead of 'RELU' because it avoids dying relu problem.

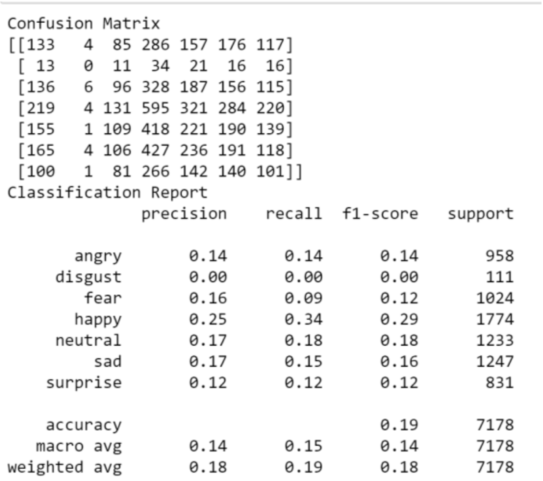
Then to Increase the Accuracy more, we also used the Kernel_Regularizer and to increase the efficiency, we exported it in one of the Most efficient file format for ML, .h5 and were able to get an accuracy of 67%. It can also be increased with more no. of Epochs. And without this, the Accuracy and loss were like this:



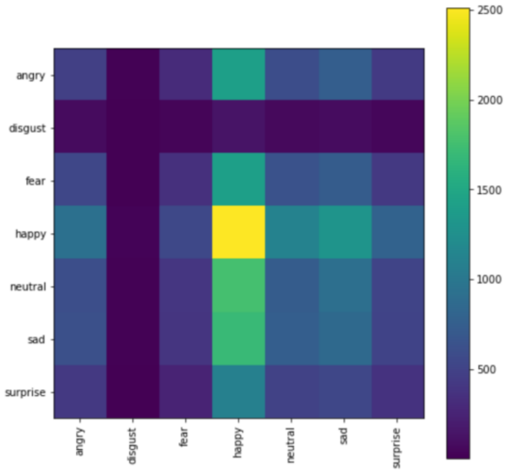


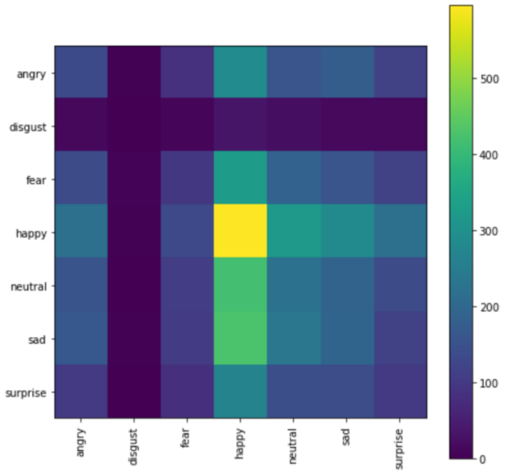
High accracy is achieved on training set but accuracy on validation set is stuck at 60-70%, also no overfitting can be seen in the dataset hence it can be concluded that the inefficiency may be due to the unbalanced dataset

```
Confusion Matrix
[[ 481   11  308 1420  598  746  431]
 [  76    1   31  137   67   82   42]
 [ 534   10  346 1420  637  734  416]
 [ 932   25  543 2510 1116 1299  790]
 [ 608   15  399 1770  733  921  519]
 [ 611   15  387 1699  755  857  506]
 [ 420    7  250 1096  495  535  368]]
Classification Report
              precision    recall  f1-score   support

       angry       0.13      0.12      0.13      3995
     disgust       0.01      0.00      0.00       436
        fear       0.15      0.08      0.11      4097
       happy       0.25      0.35      0.29      7215
     neutral       0.17      0.15      0.16      4965
         sad       0.17      0.18      0.17      4830
    surprise       0.12      0.12      0.12      3171

    accuracy                           0.18     28709
   macro avg       0.14      0.14      0.14     28709
weighted avg       0.17      0.18      0.18     28709
```



```
Confusion Matrix
[[133   4  85 286 157 176 117]
 [ 13   0  11  34  21  16  16]
 [136   6  96 328 187 156 115]
 [219   4 131 595 321 284 220]
 [155   1 109 418 221 190 139]
 [165   4 106 427 236 191 118]
 [100   1  81 266 142 140 101]]
Classification Report
              precision    recall  f1-score   support

       angry       0.14      0.14      0.14       958
     disgust       0.00      0.00      0.00       111
        fear       0.16      0.09      0.12      1024
       happy       0.25      0.34      0.29      1774
     neutral       0.17      0.18      0.18      1233
         sad       0.17      0.15      0.16      1247
    surprise       0.12      0.12      0.12       831

    accuracy                           0.19      7178
   macro avg       0.14      0.15      0.14      7178
weighted avg       0.18      0.19      0.18      7178
```





Future Scope: Using VGGNet 'Very Deep Convolutional Networks for Large-Scale Image Recognition' and implementing Depression detection more efficiently

ACKNOWLEDGEMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.