Multimodal Activity Recognition

Velpuru Sri Kashyap
University of Stuttgart
Stuttgart, Germany
st159630@stud.uni-stuttgart.de

Deepthi Sreenivasaiah
University of Stuttgart
Stuttgart, Germany
st159142@stud.uni-stuttgart.de

Vivek Ramalingam Kailasam
University of Stuttgart
Stuttgart, Germany
st159350@stud.uni-stuttgart.de

ABSTRACT

Human visual behaviour is a very rich source of information to understand human activities. The combination of this visual behaviour with scene information could significantly boost the performance of recognizing activities. Previous works have focused on using both supervised and unsupervised methods for determination of predefined activities only from short term and long term visual behaviour. In our work, we integrate two different modalities namely visual gaze behaviour and visual scene information modals using late fusion technique, provide a comparison on three different deep learning architectures with a combination of CNN and LSTM for activity recognition by using long term gaze information and egocentric view based visual scene information. Our results show that an architecture with LSTM and pre-trained VGG19 outperformed the other two models.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Computer systems organization** → **Neural networks**;

FIS'19, March 2019, Stuttgart, Germany

KEYWORDS

Multimodal; Gaze; LSTM; CNN; Late fusion

ACM Reference Format:

Velpuru Sri Kashyap, Deepthi Sreenivasaiah, and Vivek Ramalingam Kailasam. 2019. Multimodal Activity Recognition. In *Proceedings of Fachpraktikum Interaktive Systeme (FIS'19).* ACM, New York, NY, USA, 8 pages.

INTRODUCTION & RELATED WORK

Our eye movements are closely related to the activities that we perform during our day. The eye movements such as saccades and fixations, expansion and contraction of pupil, blink rate etc. corroborate as a rich source of information to quantify and distinguish our activities such as at work [1] and concentrated reading [5]. The advances in technology especially in the domain of wearable devices such as smart watches, mobile phones, mobile eye trackers, fitness trackers etc. have predominantly led to the surge of applications like activity recognition and context awareness. Activity recognition has a vast variety of applications such as quantified self, user behaviour analysis, life logging and mental health monitoring.

In [7], the authors focused on unsupervised learning approach for associating long term visual behavior to a set of activity classes. They used a head-mounted video-based eye tracker to build a ground truth annotated natural long-term visual behavior dataset of 10 participants. Later, they used a bag of words representation that encodes different eye movements such as saccades, fixations and blinks with a topic model. We obtained the data set from [7] which contains 80 hours of long term gaze behavior and visual scene information. Our surroundings, that is visual scene information also aid critical information about our activities. The major motivation of our work is to integrate gaze data with visual scene information to obtain a better understanding and recognition of the activities. To realize this, we first integrated the two different modalities: gaze data modal and visual scene information modal using the late fusion technique as in [4]. The authors in [4] have concentrated on building modality specific architectures for each sensor in their system and later on different fusion methods of those modalities. We imbibed the late fusion technique from [4] as it had promising results as compared to the other fusion techniques. Later, we evaluated and compared different neural network architectures: *LSTM model and pre-trained VGG on gaze and scene data, CNN with gaze and scene data, CNN with only gaze data.*

ACTIVITY NUMBER	OUTDOOR	SOCIAL INTERACTION	FOCUSED WORK	TRAVEL	READING	COMPUTER WORK	WATCHING MEDIA	EATING	SPECIAL
ACTIVITY ENCODING	1	0	0	0	0	0	0	1	0

OUTPUT LABEL : Outdoor & Eating

Figure 1: An Example of Label Encoding

One Data Sample	30 Gaze Points	1 Video Frame	1 Label

Figure 2: Structure of a data frame

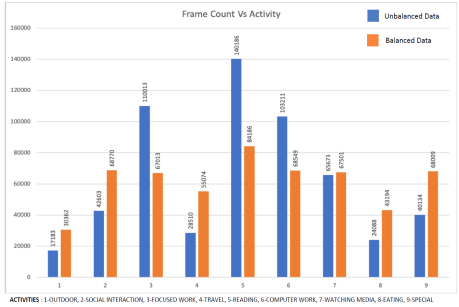

Figure 3: Data Balancing

DATA ANALYSIS AND DATA PREPROCESSING

Dataset

The data used for this project consists of two types of modalities namely gaze data represents the human visual behaviour and scene data represents the egocentric view of a person. The data was collected and annotated by the team members of the research paper [7], and the same data was used for their project too. Ten participants had contributed in the data collection, more than 80 hours of data were collected from them. The apparatus used for the data collection is a PUPIL - head mounted eye tracker. This eye tracker features two cameras, one is eye camera with resolution of 640 x 360 and other is egocentric (scene) camera with resolution of 1280 x 720. The camera's recording frequency is 30 Hz.

Ground truth annotations were carried for full dataset, post recording of data. The annotations table chart consists of the following details – scene frame number, scene frame time, day time, activity label, activity description and activity label encoding for nine different non-mutual activities. Hot coding is used for the activity labelling. i.e. when certain activity is carried out '1' else '0'. The nine types of activities are outdoor, social interaction, focused work, travel (like driving or walking), reading, computer work, watching media, eating, and special (activities like checking mobile phone, packing backpack and tying shoe lace etc). These are non–mutual activities which means a person can carry out more than one activity at a same time. An example of the label encoding is shown in the Figure 1. Thus, this project is identified as multi-label classification. The provided gaze data chart consists of eye frame number, eye frame time, gaze x, gaze y, pupil x, pupil y confidence, ellipse major, ellipse minor and ellipse rotation angle. Among them, only eye frame time, gaze x, gaze y, pupil x and pupil y considered in this project.

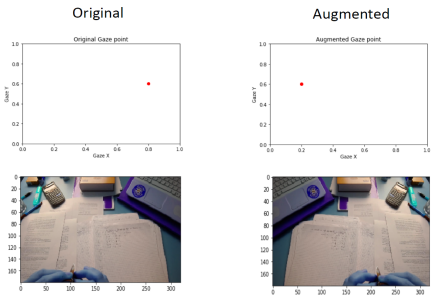

Figure 4: Data Augmentation of Gaze and Image

Data Preprocessing

The eye frame time which is related to eye camera and scene frame time which is related to scene camera correlates with each other (mild deviations between them is ignored). These timestamps are the float values represented in time unit second, around thirty frames of images are captured in one second since frequency is 30 Hz. These float value of time frame is converted to integer by truncating (ignoring the values after decimal point), so that each value of time has around thirty frames of image, corresponding gaze data and ground truth annotations etc. It is decided to consider all thirty gaze points in a second as referred from [3] and to consider only the first scene frame among the thirty frames in a second. That scene frame is resized to resolution 160 x 90 which is one eighth of its original size. Because if all scene frames are considered, it would increase data size tremendously and it is a logical assumption that a scene frame in a second is enough to depict and represent what activities are being carried out in that second. In the gaze data, in some of the timestamps there were more than thirty gaze points found (may be due to some miscellaneous irregularities). For any time stamp value (second), if the gaze points are less than 30, then the last available gaze point is replicated to get exactly 30 gaze points. If they are more than 30, then only the first 30 were taken. For the decision of activity label, among the thirty label encodings, if all were same then that encoding is considered and if more than one label encodings were present then the encoding which has majority of occurrence is considered (voting method). The data frame for one second will have a structure as shown in the Figure 2.

Data Augmentation

Imbalance was found in the provided dataset, i.e. a greater number of the samples belongs to the activities like reading and focused work whereas a lesser number of samples belongs to the activities like eating and outdoor. This imbalance in data may lead the deep learning model to get biased towards activities which have more samples. Thus, data balancing was done by data augmentation and removal of data. Data augmentation to increase the number of samples of activities which have lesser number of samples and data removal to remove some of the samples which have dominating activities. The data with count of each activities before balancing and after balancing is shown in figure 3. Data augmentation for scene image is done by flipping of the image with respect to vertical axis (left becomes right and vice versa). Data augmentation of gaze is done by subtracting gaze x and pupil x value each from value 1 i.e. 1-gaze x ; 1-gaze y. Since the gaze data is normalized to 1, the augmented gaze points is nothing but the points which are flipped with respect to vertical axis. Example of data augmentation is shown in the Figure 4.

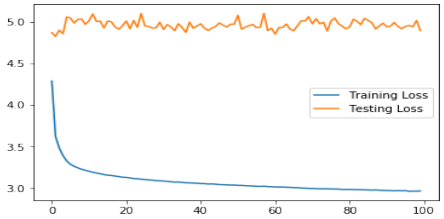

Figure 5: CNN-Gaze data Model Loss

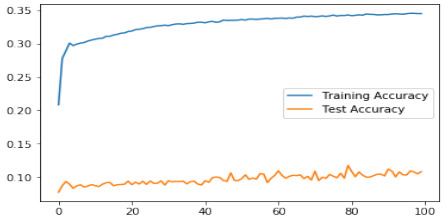

Figure 6: CNN-Gaze data Model Accuracy

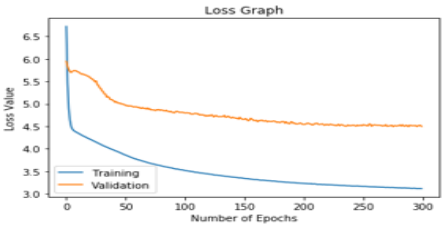

Figure 7: CNN-Gaze+Scene data Model Loss

Table 1: Overview of Dataset

Activity Class	Description	Total Duration(min)
Outdoor	Person is outside	466
Social Interaction	Person is interacting	855
Focussed Work	Person is doing focussed work	1877
Travel	Person is travelling	496
Reading	Person is reading	2371
Computer Work	Person is working on a computer	1720
Watching Media	Person is watching media	1097
Eating	Person is eating	422
Special	Sepcial Events	685

Among the ten persons data, seven persons' data was used for training, two persons' data was used for validation and one person's data was used for testing. This data split is done with respect to persons and not with respect to activities to avoid unwanted resemblance between testing and training samples. if a particular person's data is available in both testing and training, there will be higher resemblance between the test samples and train samples due to environment he/she lives or due to things he/she possess.

DEEP NEURAL NETWORK MODELS

At first we used gaze and scene data sampled every second. It consisted of 227,146 samples in training, 67,648 samples in testing and 29,731 samples in validation. Since the dataset was too huge, we then sampled data every 5 seconds and then every 10 seconds. This reduced the sample count by one-tenth. Intuition behind this was that since many activities like focussed work and reading prolonged for a long time as shown in Table 1, there will be many similar images in the scene data when sampled very frequently.

Loss Function

With binary cross-entropy as loss function, the model did not perform well and the accuracy was not up to the mark. Since the given task is multi-label classification, binary or categorical loss function could not be used. Hence, we used a custom loss function called Multi-task loss function *[*Source:

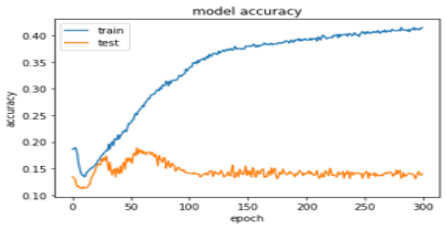

Figure 8: CNN-Gaze+Scene data Model Accuracy

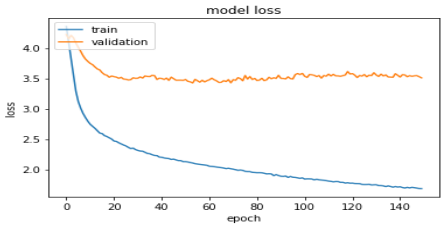

Figure 9: LSTM + pretrained VGG Model Loss

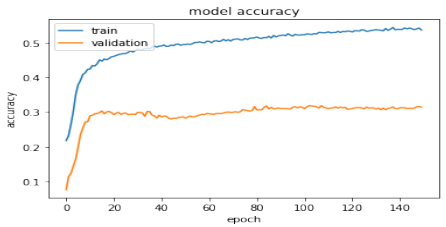

Figure 10: LSTM + pretrained VGG Model Accuracy

https://bit.ly/2C8Cv1D, https://bit.ly/2ExbIfT]. Similar to the binary-cross entropy, the new loss function applies the log loss individually on each output neuron. In addition to that, an extra summation term adds all the losses to give an aggregate loss value.

We implemented neural network for multi-label classification using Keras Framework and tested different network configurations by varying the number of neurons and layers, activation functions, and optimizers provided by Keras. We have used Hamming Loss as an accuracy metric to measure the model's performance. We have also presented the results of individual activity using confusion matrix. We built 3 different neural network structures: one CNN using only gaze data, one CNN using both gaze and scene data and one LSTM model with pretrained VGG network [6] using gaze and scene data. Below are the details of each neural network along with their parameters and training procedure. The results of each model is summarized in the Table 2.

CNN with only gaze data

We used a simple convolution neural network with convolution layers to extract features from gaze data. The extracted features were combined using 3 dense layers. We obtained a good result using 2 CNN blocks with filter size of 3x1. We used ReLU activation function in all of the hidden layers, while in the output we used sigmoid. The optimizer used for training is Adam with learning rate of 0.001. We initialized our weights using Glorots initialization [2] in Keras. A good result was obtained when we used the 4 gaze inputs (gaze x, gaze y, pupil x, pupil y).

CNN with gaze and scene data

As CNN is best suited for image-like data, we used CNN blocks to extract features from both gaze and scene data. We obtained good results with 2 CNN blocks on scene data with filter size of 5x5 and 3x3, and 1 CNN block on gaze data with gaze x and gaze y points using the filter size of 2x2. The features extracted individually were then concatenated using a Concatenate layer. This method of using modality specific network for feature extraction and then combining them together is called late fusion method [4]. The concatenation is followed by CNN and max pooling layers. The optimizer used is Adam with a learning rate of 0.0001. We used ReLU and sigmoid activation functions in all the hidden layers and output layer respectively. We also tried with inception like architecture with different filter sizes. But the observation was that with more number of layers, the model was tending to overfit and perform poorly on the test set.

LSTM model + pretrained VGG on gaze and scene data

We used a pre-trained VGG19 model from Keras to extract features from scene images. This model is not trainable; hence we have used it for transfer learning. The best result was obtained with 2 CNN

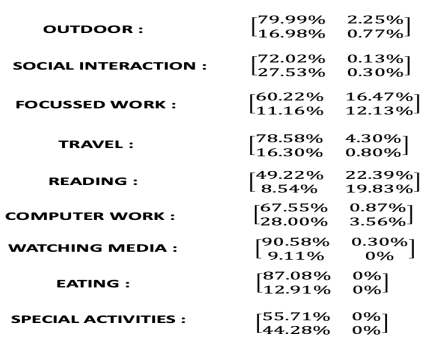

Figure 11: CNN-Gaze data Model Confusion Matrix

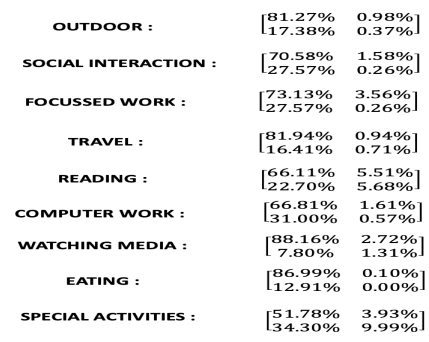

Figure 12: CNN-Gaze+Scene data Model Confusion Matrix

Table 2: Best results of all tested models

Model	Hamming Loss	Multi-Task Loss	Accuracy
LSTM + Pre-trained VGG	0.23	3.47	31.62 %
CNN Gaze + Scene Data	0.24	4.50	21.89 %
CNN Gaze	0.25	5.25	11.43 %

blocks and one LSTM layer. A time-distributed wrapper from Keras was used to load multiple images at a time into the VGG19 model and extract features. The extracted image features were then given to an LSTM layer. The gaze data with gaze x and gaze y is passed to the CNN blocks. The features extracted from individual modalities were concatenated using Concatenate layer in the Keras. This was followed by 3 dense layers. Glorots initialization [2] was used in the CNN blocks. The optimizer used was Adam with learning rate of 0.0001. Timesteps was one of the important hyper parameter in this model. We chose the timesteps as 12, so every time the model takes 12 data samples at a time for processing. We also used the concept of sliding window with the window size of 12 while extracting the data from .h5 file using data generator.

RESULTS

We tested the trained models on person no.6 data. After several iterations by changing different hyper parameters, we obtained the results for 3 neural network approaches as summarized in the Table 2 The results obtained on LSTM + Pretrained VGG model is observed to be better than CNN models. We have calculated individual confusion matrices for each activity since it is a multi-label classification. From the confusion matrices, we can see that the sum of True positive and True negative for each activity is higher in LSTM + Pretrained VGG model than that of the other two models.

CONCLUSION

In this project, multimodal activity recognition is done using two different types of modalities : gaze and scene. The imbalance in the data set is balanced using data augmentation. Three different deep learning architecture models were developed: *CNN with only gaze data*, *CNN with gaze and scene data* and *LSTM model + pretrained VGG on gaze and scene data*.Comparison among these three models were carried out.

It is inferred that integrating scene data with gaze data will boost the performance of model. Hamming loss of *CNN with only gaze data* is 0.25 which is greater than other two multimodal models

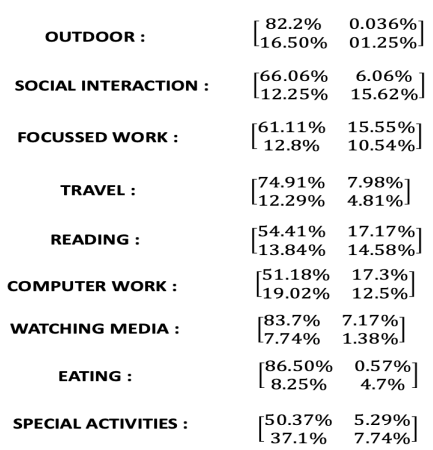

Figure 13: LSTM + pretrained VGG Model Confusion Matrix

hamming losses. For activities like watching media, eating and special activities, no correct prediction was done when only gaze data was used, 0 % true positive.

On comparing the other two multimodal models, *LSTM model + pretrained VGG on gaze and scene data* has shown better results with respect to hamming loss. The hamming loss of *LSTM model + pretrained VGG on gaze and scene data* is 0.23 which lesser than the hamming loss of *CNN with gaze and scene data* (0.24). Upon keen observation of confusion matrices, the true positive prediction for social interaction, focused work, reading, computer work, travel and eating are better in *LSTM model + pretrained VGG on gaze and scene data*. In social interaction activity, true positive prediction is 15.62% in *LSTM model + pretrained VGG on gaze and scene data* whereas it is just 0.26% in *CNN with gaze and scene data*. Likewise, true positive predictions for eating activity is 4.7% when *LSTM model + pretrained VGG on gaze and scene data* is used whereas it is 0% on using *CNN with gaze and scene data*.

REFERENCES

[1] Andreas Bulling, Jamie A. Wardand Hans Gellersen, and Gerhard Troster. 2010. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33 (2010). https://doi.org/10.1109/TPAMI.2010.86

[2] Xavier Glorot and Yoshua Bengio. 2014. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* 9 (2014). http://proceedings.mlr.press/v9/glorot10a.html

[3] Sabrina Hoppe and Andreas Bulling. 2016. *CoRR* abs/1609.02452 (2016). https://arxiv.org/abs/1609.02452

[4] Valentina Radu, Sourav Bhattacharya, Nicholas D. Lan, Mahesh K. Marina, and Fahim Kawsar. 2017. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4 (2017). https://doi.org/10.1145/3161174

[5] Ishimaru S, Weppner J, Kunze K, Kise K, Dengel A, Lukowicz P, and Bulling A. 2014. *Proceedings of the 5th Augmented Human International Conference* 15 (2014). https://doi.org/10.1145/2582051.2582066

[6] Karen Simonyan and Andrew Zisserman. 2010. *CoRR* abs/1409.1556 (2010). https://dblp.org/rec/bib/journals/corr/SimonyanZ14a

[7] Julian Steil and Andreas Builing. 2015. *UbiComp '15* 15 (2015). https://doi.org/10.1145/2750858.2807520