CNN as a feature extractor in gaze recognition

ARUN GOPAL GOVINDASWAMY, DePaul University, USA
ENID MONTAGUE, DePaul University, USA
DANIELA STAN RAICU, DePaul University, USA
JACOB FURST, DePaul University, USA

In this paper, we employ a Convolutional Neural Network (CNN) in predicting physician gaze. This paper focuses on two aspects – one comparison between hand-crafted features and CNN-based learned features, and two in investigating the impact of fully-connected layers in an end-to-end CNN model. The pre-trained CNN model based on VGG16 through transfer learning is used as a feature extractor and a K-Nearest Neighbor and a Random Forest (RF) algorithm were used as the classifier of physician gaze. The CNN-RF and CNN–K-NN models were compared with the traditional end-to-end CNN model and through a series of experiments and statistical tests of significance, we show that the power of CNN comes from the features extraction part and that the fully connected layers of the CNN have comparable performance to the random forest and the k-NN classifiers. We also show that the CNN-based learned features provide substantial distinguishable power in classifying physician gaze.

CCS Concepts: • **Computing methodologies** → **Activity recognition and understanding**; **Supervised learning by classification**; *Classification and regression trees*; *Neural networks*.

Additional Key Words and Phrases: neural networks, gaze recognition, random forest

ACM Reference Format:
Arun Gopal Govindaswamy, Enid Montague, Daniela Stan Raicu, and Jacob Furst. 2020. CNN as a feature extractor in gaze recognition. In *2020 3rd Artificial Intelligence and Cloud Computing Conference (AICCC 2020), December 18–20, 2020, Kyoto, Japan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3442536.3442542

1 INTRODUCTION

Recent advances in healthcare and technology has given rise to many e-health applications. One such application is the Electronic Health Record (EHR) system facilitating smooth flow of accurate information, better medication management and better documentation of health care records helping the healthcare provider making informed decisions. The usage of EHR inside clinical settings has increased and research shows both positive and negative impacts of the EHR. Patterns of EHR usage by the physician is imperative in understanding the patient outcomes and physician burnout [20] - [22]. Physician gaze has been one of the important non-verbal feature and needs accurate prediction in the understanding of patient-physician interaction [23]. Physician gaze recognition in clinical settings has been a challenging task because of the varied nature of the clinics, light settings, camera angles and constant movement of the physician.

In gaze recognition, traditional methods of designing features using audio and video data have been previously employed in training machine learning models. Although the combination of hand crafted features and machine learning models achieved high performance in recognizing gaze, these models had low generalizing ability and the

Manuscript submitted to ACM

model performance lowered with increasing data set. In negotiating with these limitations, a CNN model based on VGG16 model through transfer learning is employed in gaze recognition. A CNN model known to have achieved superior performance in various computer vision tasks is composed of two parts – one being the feature extraction part where the input image is reduced to a set of feature maps through a series of strategically arranged convolution and pooling layers and two being the classifier where the features are passed into a series of fully connected or hidden layers and a output layer. The combination of these feature extractor and fully connected layers called as the end-to-end CNN model is used as a baseline model and is compared with traditional image classification technique of hand crafted features with a random forest classifier and a novel approach of a CNN-RF model.

The contribution of this work is two-fold – one investigation and direct comparison between hand-crafted and CNN based learned features, two – analyzing the impact of two different parts of the CNN model (feature extraction part with convolutional and pooling layers and classifier part involving fully connected layers) in gaze recognition task. This paper highlights the downsides of using hand-crafted features involving extensive human labor in the feature extraction phase and points the efficacy of the CNN model in automatically extracting deep high-level features. This work shows that the high-level feature extracted from the pre-trained CNN model has substantial distinguishable power in classifying physician gaze and shows that the choice of classifier is not significant in this application. This work provides statistical and experimental evidence that the end-to-end CNN need not always be the go-to mechanism for image recognition tasks and that the fully connected layers can be replaced by other choice of classifiers depending on the application.

2 RELATED WORK

In gaze recognition, Gutstein et al. [1] [2] used hand-crafted features to train AdaBoost [3] models in predicting physician gaze. Three separate doctor-specific models were built using extracted optical flow [4] features and Mel-Frequency Cepstral Coefficient (MFCC) features [5]. Although these models showed high performance, these models did not generalize well on new interactions. Similarly, hand-crafted features had poor generalizability in other applications as well [25] - [27].

Gaowei et al. [6] shows that the combination of CNN features with random forest classifiers perform better than traditional end-to-end CNN model. Features from multiple convolutional layers were extracted and fed into three independent random forest classifiers. The author proposes to use multi-level features in classification task and showed that the combination of multi-level features with random forest classifiers perform better than the traditional CNN with only high-level features. Gaowei et al. used a CNN model based on LeNet-5 in extracting features for the images in the data set. The features from three different layers were extracted and used in training three independent random forest models. The classification results from the 3 models were then combined using winner-takes-all ensemble strategy. The results suggest that multi-level features provide better generalizability of the model than only high-level features and that the CNN features with random forest works better than the end-to-end CNN model.

Niu and Suen [7] recognized handwritten digits using a novel method. In this approach, a traditional CNN model was trained and then the output from the hidden layer was extracted from the pre-trained CNN model and were used in training an SVM classifier. Niu and Suen used the CNN as a feature extractor and the SVM as a classifier. The results show that the error rate of the hybrid model to be lower than the CNN model itself. The paper recommends a hybrid model for image recognition tasks as the hybrid model combines the advantages of both CNN and SVM – where CNN can be used to extract high level features and SVM can be used as classifier. Reference [7] also supports the use of

learned features in image recognition tasks as opposed to hand-crafted features which are tedious and time-consuming to generate.

Basly et al. [8] combined the deep learning-based method and a traditional classifier based hand-crafted feature extractors in order to replace the artisanal feature extraction method with a new one. In this approach, the CNN based learned features were extracted from a pre-trained CNN model based on ResNet and the features were then used to train an SVM model in recognizing human activity. In this approach, the CNN model was used as a feature extractor and the SVM model was used as the recognizer or the classifier. The results show that the CNN-SVM model produced 99.92% accuracy and outperformed traditional CNN model and other fusion algorithms.

Liu et al. [9] performed a combination of CNN and SVM in recognition of Gender based on gait. The VGGNet-16 model was used through transfer learning for the gender recognition task. The authors employed different methods in tuning the VGGNet-16 model and extracted features from three different fully-connected layers. The softmax layer was replaced by an SVM classifier and the results shows that the CNN-SVM model performs better than the traditional CNN model.

Cao et al. [10] used a hybrid approach of combining a CNN with a random forest algorithm for segmenting electron microscopy images. In this approach, a CNN model consisting of convolutional layers, pooling layers, fully connected layers and a softmax was trained with input images. The trained CNN model was then used to extract features for the images. The output from the last convolutional layer of the CNN model was extracted and fed into a random forest classifier. The results showed that the hybrid method was successful than a traditional CNN model in segmenting electron microscopy images.

In this paper, we first train an end-to-end CNN model based on VGG-16 and then use the same model in extraction of features to the images in the dataset. The features extracted were then used to train a random forest model and a K-NN model separately. We perform 4 different experiments in training the CNN model and through thorough experimentation show that the power of prediction lies in the features extracted and not in the type of classifier used.

3 METHODOLOGY

3.1 Data

The current data base consists of 101 interactions between the patient and the physician. The study involves 10 doctors and 101 patients which was performed through the University of Wisconsin-Madison at five primary care clinics in 2011 [11]. Every patient in the study agreed to be videotaped and to participate in the study and signed a consent form. The 101 interactions were highly dynamic, as the lighting, camera placement, and number of people fluctuated between each interaction. These 101 interactions were captured using 3 different cameras (Figure 1) – each placed at different positions and angles in the clinic. Patient-Centered camera – focuses on the patient's chair, Doctor-Centered camera – focuses on the doctor's face and Wide-Angle camera – captures both the patient and the doctor from a wide angle. All these cameras recorded the clinical interactions at 30 frames per second (fps). The Multi-Channel view is a collection of the Patient-Centered, Doctor-Centered and the Wide-Angle frames capturing at a given time. Only the doctor-centered videos were used in the study to predict physician gaze. The doctor-centered camera focuses on the doctor capturing subtle optical flow changes. Further, human encoders annotated the entire duration of the video for each interaction.

The manual annotations encoded physician communication, physician gaze, and patient gaze through the Noldus Observer XT software [12]. The start and end time as well as duration were recorded for each of the patient and physician behaviors. There were different annotations determining where the physician gazes at a given time. This

Fig. 1. Interaction video data: example of Patient-Centered, Doctor-Centered, Wide-Angle, and Multi-Channel videos from a particular time [1] [2]

study simplifies the physician's gaze to two levels. If the physician was deemed to be looking at the patient, then it was labeled as Patient. And, if the physician was not deemed to be looking at the patient, then it was labeled as Other. Since the analysis was performed on a frame level basis, all the original annotations were mapped to each frame. Of the 101 interactions, 15 interactions from 3 doctors were used in the previous works. To maintain consistency across studies and to have direct comparison of the methodologies, we chose to have the same interactions in this work as well. To have a consistent number of frames across each interaction, only 6 minutes of the entire duration of each interaction were used.

From the 6 minutes video sequence of each interaction, the first two minutes of video sequence were used as a training set, the next 1-minute of video sequence was used as the testing set, and the last 3 minutes were used as the validation set (Figure 2).

3.2 Designed feature extraction and random forest classification

We follow the approach used by Gutstein [1] [2] to extract the optical flow measurements [4]. Optical flow measurements are used to estimate the motion of the physician between successive frames. For each optical flow computation, 15

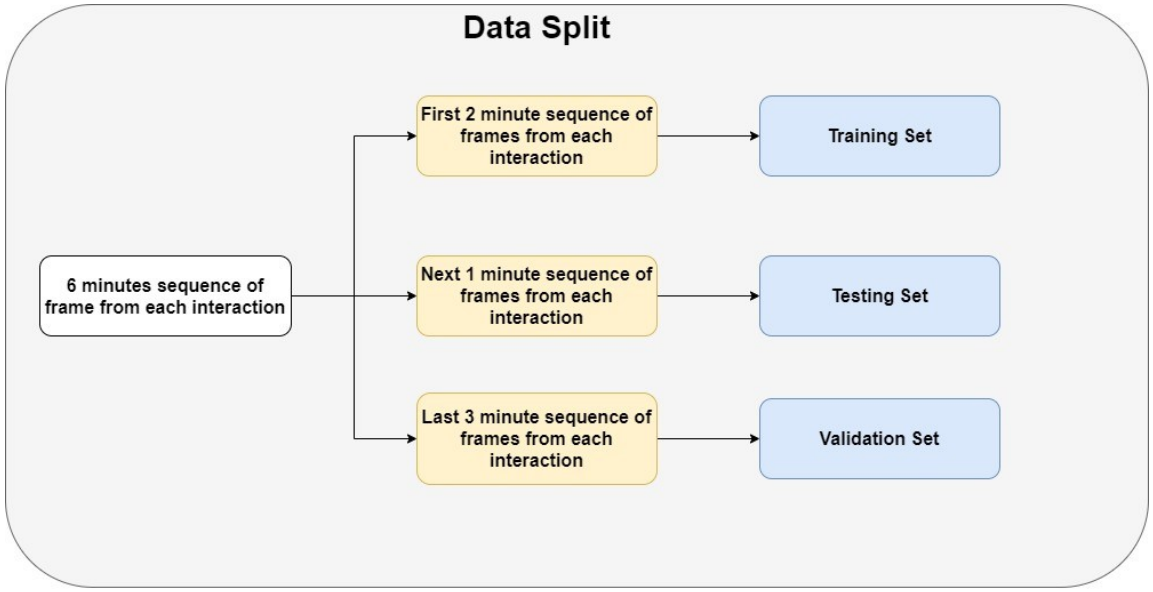

Fig. 2. Data Preparation - Split of data into training, testing and validation data

summary statistic variables were calculated regarding each of the following features – velocityU (x component of velocity), velocityV (y component of velocity), orientation and magnitude. The 15 summary statistics are as follows- maximum, minimum, 25th percentile, 50th percentile, 75th percentile, sum, sum squared, skewness, kurtosis, range, mean, variance, standard deviation, covariance, and non-zero values. The statistic non-zero Values refers to the number of non- zero values for the designated feature in the region of interest (Patient-Centered Physician, Patient-Centered Patient, or Physician-Centered frame) for optical flow measurement. Due to the large number of null optical flow values regarding velocityU, velocityV, orientation, and magnitude, the variables for velocityU, velocityV, orientations and magnitude - other than Non-Zero Values were calculated for the top 25th percentile of feature values with respect to the regions of interest. Since the doctor was exclusively present in the doctor-centered video sequence, the optical flow estimates were computed from the entire frame for the doctor-centered physician. In total, 60 optical flow features for the Doctor-Centered Physician were computed. Further, audio features were extracted from the Doctor-Centered Video. The 14 Mel Frequency Cepstral Coefficients, along with 14 delta (change in coefficients), coefficients and 14 deltaDelta (change in delta) coefficients were calculated using MATLAB's Audio Toolbox were extracted [5][18][19]. In total, 54 audio features were extracted for each frame of the video interaction. Three different random forest [16] models were trained. One model was trained using only the audio features. Second model was trained using only the video features and the third model combined the audio and video features in training the model. The models were tuned for hyper-parameters and the optimal results are shown in Table 1.

3.3 Transfer learning and CNN network architecture

In this study, we also use convolutional neural networks on frame-level images to predict physician gaze. We use transfer learning [13] to build our CNN model. We employ the VGG16 [14] model also called as the OxfordNet named after the Visual Geometry Group from Oxford as our base model. Any CNN model will have two parts – feature learning

part (convolutional and pooling layers) and the classification part (fully connected layers). In our approach, we borrow the architecture of the feature learning part of the VGG16 model and add a GlobalMaxPooling Layer, 5 fully connected layers along with a dropout layer. As seen from Figure 3, the VGG16 model has 13 convolutional layers, and 5 MaxPool layers.

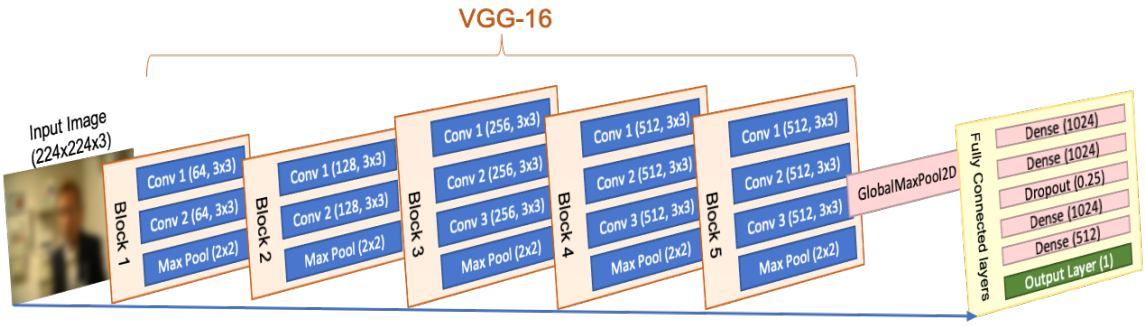

Fig. 3. The architecture of the CNN model based on VGG16

3.4 End-to-end convolutional neural network in predicting physician gaze

The VGG16 model was pretrained using the ImageNet [24] dataset. While employing transfer learning techniques, the original weights learned can be kept alike or few layers can be retrained to tweak the model for our application. In our approach, we have borrowed only the convolutional and pooling layers from the VGG16 model. Usually in transfer learning, only the last few convolutional layers will be retrained to make the features extracted application specific. In our approach, we conduct four experiments – one in which no convolutional layer was retrained, two in which last convolutional layer was retrained, three in which last 2 convolutional layers were retrained and four in which last 3 convolutional layers were retrained. We experiment only with the convolutional layers from block 5 shown in Figure 3. Usually retraining the last layer of convolutional layer is enough to gain application specific features, but we wanted to experiment retraining more convolutional layers and hence the choice of 4 experiments.

Hence in this study, 4 experiments were performed in training the CNN model. In each of the four experiments, different number of convolutional layers were retrained. In the experiment named Experiment#0, none of the convolutional layers were retrained meaning that the original weights of the VGG16 model were used during the training of the end-to-end CNN model. In another experiment named Experiment#1, the last convolutional layer (which is Block 5 – Conv 3 layer) was retrained. By retraining the convolutional layers with images from our study, the CNN model captures application specific information during the feature extraction part which further improves performance during the classification part of the CNN model. In furthering experiments named Experiment#2, the last 2 convolutional layers (Block 5- Conv 2 and Conv 3) were retrained and in Experiment#3, the last 3 convolutional layers (Block 5 – Conv 1, Conv 2, and Conv 3) were retrained. While the number of convolutional layers retrained varied across experiments, the network architecture remained the same.

The network weights were optimized using the Adam algorithm [15] which is a stochastic gradient descent method with adaptive estimator of lower-order moments with an adaptive learning rate for Experiment#3 and with a learning rate of 0.001 for all other experiments and the batch size for Experiment#1 and Experiment#2 were 64 and Experiment #3 and Experiment#4 were 32.

Table 1. Performance of random forest classifier in predicting physician gaze using hand-crafted features

Experiment	Training Accuracy	Testing Accuracy	Validation Accuracy
Audio Features	91.58%	67.51%	57.75%
Video Features	98.11%	67.87%	58.84%
Audio + Video Features	97.45%	68.01%	59.46%

3.5 Learned features extraction from the trained CNN models

After the 4 experiments were conducted, each of the 4 model were used in extracting features for the input dataset. Since each model has different weights for the last few convolutional layers, the features extracted from each of the models were different. The output of the GlobalMaxPooling layer were 512 in dimension meaning each image had 512 features that were automatically learned by the CNN model. The features were extracted from Experiment#0, Experiment#1, Experiment#2, and Experiment#3 and were named as Learned_CL#0, Learned_CL#1, Learned_CL#2, and Learned_CL#3 respectively. The Learned_CL#0 for example means that these features were learned through retraining of last 0 layers of the CNN model. Similarly, Learned_CL#1 means that the features were learned through retraining of last 1 layer of the CNN model and so on for Learned_Cl#2 and Learned_CL#3.

3.6 Learned feature with random forest and k-nearest neighbor algorithms in predicting physician gaze

The 512 features learned from the trained CNN models were further used in training a Random Forest model and a K-Nearest Neighbor model. Four different RF [16] and K-NN models [17] were trained using the four different learned features (Learned_CL#0, Learned_CL#1, Learned_CL#2, and Learned_CL#3).

4 RESULTS AND DISCUSSIONS

4.1 Designed features and random forest in predicting physician gaze

The optical flow features extracted from the frame level images of the doctor-centered videos were used in training the random forest model. The random forest model was trained using the training set, tuned for hyper-parameters using the testing set, and validated using the validation set. The performance of the model on training set was 98%, testing set was 67% and validation set was 58%. The results (Table 1) showed evidence of high over fitting and the performance on the validation set was just above random guess and the results suggest that the designed optical flow features does not work in predicting physician gaze.

4.2 End-to-end convolutional neural network in predicting physician gaze

An end-to-end convolutional neural network (CNN) model was adopted in predicting physician gaze. The network architecture was held constant as shown in the previous section and the number of convolutional layers retrained was varied across experiments. While the Adam optimizer was used in learning the weights of the neurons, an adaptive learning rate was used for Experiment#3 whereas a learning rate of 0.001 was used for the other experiments. The performance of the models on training, testing and validation set is shown in the following table.

The results from Table. 2 show significant increase in performance of the models especially on the testing and validation set. Clearly the end-to-end CNN model outperformed the traditional approach of using designed features and a machine model like random forest in predicting physician gaze. Moreover, the performance of the model increased by

Table 2. Performance of the end-to-end CNN model in predicting physician gaze

Experiment	Training Accuracy	Testing Accuracy	Validation Accuracy
Experiment#0	96.72%	89.38%	83.95%
Experiment#1	97.71%	92.22%	89.11%
Experiment#2	98.85%	92.29%	89.56%
Experiment#3	96.15%	92.29%	89.25%

Table 3. Performance of different classifiers in predicting physician gaze

Learner Used	Feature Used	Training Accuracy	Testing Accuracy	Validation Accuracy
End_to_end CNN	Learned_CL#0	96.72%	89.38%	83.95%
End_to_end CNN	Learned_CL#1	97.71%	92.22%	89.11%
End_to_end CNN	Learned_CL#2	98.85%	92.29%	89.56%
End_to_end CNN	Learned_CL#3	96.15%	92.29%	89.25%
Random Forest	Learned_CL#0	99.27%	89.62%	83.45%
Random Forest	Learned_CL#1	98.48%	94.47%	89.51%
Random Forest	Learned_CL#2	98.36%	93.07%	90.04%
Random Forest	Learned_CL#3	99.59%	93.69%	89.33%
K-Nearest Neighbor	Learned_CL#0	98.51%	88.60%	83.03%
K-Nearest Neighbor	Learned_CL#1	97.85%	93.89%	88.50%
K-Nearest Neighbor	Learned_CL#2	98.50%	92.05%	88.75%
K-Nearest Neighbor	Learned_CL#3	98.43%	93.07%	88.55%

each addition of retrained convolutional layers. The results suggest that retraining the last convolutional layer was enough to achieve an accuracy of 89% in predicting physician gaze.

4.3 Learned feature with random forest and k-nearest neighbor algorithms in predicting physician gaze

A typical end-to-end convolutional neural network (CNN) model consists of two parts – feature extraction part and the classification part. The feature extraction part usually consists of convolutional layers and pooling layers and the classification part consist of fully connected layers and dropout layers. The high performance of the end-to-end CNN model lead to further investigation in understanding the importance of either parts of the CNN model. The features from all the four trained CNN models were extracted and were used in training a random forest model and a k-nearest neighbor model. The 4 different learned features were used in training, testing and validating the 8 different models and Table. 3 shows the performance of the optimized models.

Three paired t-test were conducted between each pair of validation accuracy. The smaller the p-value, the stronger the evidence to reject null hypothesis. The null hypothesis that the two samples are similar can be accepted with a p-value of less than 0.05. A paired t-test between the validation accuracy of end-to-end CNN model and random forest provided a p-value of 0.296 suggesting that there is no evidence in rejecting null hypothesis. This means that validation accuracy of end-to-end CNN and random forest are similar. The paired t-test between validation accuracy of end-to-end CNN and k-nearest neighbor algorithm provided a p-value of 0.876 suggesting that there is no evidence in rejecting null hypothesis. The paired t-test between validation accuracy of random forest model and k-nearest neighbor algorithm

provided a p-value of 0.295 suggesting that there is no evidence in rejecting null hypothesis. From all the 3 paired t-test, the results suggest that the validation accuracy of all the three models are similar.

5 CONCLUSION AND FUTURE WORK

In this paper, we investigated the use of hand-crafted features in predicting physician gaze. Optical flow features and MFCC features of the patient-physician interaction were extracted and fed into the random forest classifier. The results showed high evidence of overfitting. Although previous works of using hand-crafted features showed promise, the designed features were found to not have the power of generalizing and the performance of the models provided evidence to the hypothesis. On the other hand, the CNN based learned features extracted from the pre-trained CNN model showed significant improvement over traditional methods and provided more reliable features in predicting physician gaze. The VGG16 based CNN model was also fine-tuned to different convolutional layers and the results showed that retraining the last convolutional layer was enough to capture additional information from the features. This paper also investigated the two important tasks of a CNN – feature extraction and classification. The end-to-end CNN model was kept the baseline model and the model was used to extract features from the input images. The extracted features, then used to train a random forest and a K-NN classifier, produced similarly performing gaze recognition models. Through different experiments and statistical tests for significance, the classifiers were found to have similar performance and in this paper, we conclude that the power of CNN has been in the convolution and pooling layers than the fully connected layers. It could be safely concluded that the CNN does not always need to have fully connected layers for optimal performance and that the different choice of classifiers can be experimented depending upon the application.

Although our results show that the fully connected can be replaced by any other classifier depending on the application, the fully connected layers have anyways contributed to the feature extraction during the training of the CNN model. In other words, the feature was extracted from a pre-trained CNN model and the fully connected layers contributed in training the network through back and forward propagation methods. In order to completely replace the fully connected layers, we propose a novel method of replacing the fully connected and softmax output layer with a random forest algorithm. We set a loss function and based on output from the random forest classifier, we propose to update the weights of the neurons in the convolutional layers. This way we replace the fully connected layers with a random forest classifier and the proposed idea would be a novel hybrid end-to-end CNN with random forest classifier.

ACKNOWLEDGMENTS

This research was supported by NSF Division of Information & Intelligent Systems Award - “CHS: Small: Extracting affect and interaction information from primary care visits to support patient-provider interactions” (Grant No: 1816010).

REFERENCES

[1] Gutstein, D., Montague, E., Furst, J.D., and Raicu, D.S.: ‘Hand-Eye Coordination: Automating the Annotation of Physician-Patient Interac- tions’, 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), 2019, pp. 657-662

[2] Gutstein, D., Montague, E., Furst, J.D., and Raicu, D.S.: ‘Optical Flow, Positioning, and Eye Coordination: Automating the Annotation of Physician-Patient Interactions’, 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019, pp. 943-947

[3] Y. Freund and R.E Schapire, "Experiments with a New Boosting Algorithm Machine Learning", Proc. of the Thirteenth Int. Conf., pp. 148-156, 1996.

[4] B. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision", Proc. of Imaging Understanding Workshop, pp. 121-130, Apr. 1981.

[5] H. Fayek. “Speech Procesing for Machine Learning: Filter banks, Mel-Frfequency Cepstral Coefficents (MFCCs) and What’s In-Between,” April 2016, https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html.

[6] Xu, Gaowei; Liu, Min; Jiang, Zhuofu; Söffker, Dirk; Shen, Weiming. 2019. "Bearing Fault Diagnosis Method Based on Deep Convolutional Neural Network and Random Forest Ensemble Learning." Sensors 19, no. 5: 1088.

[7] Xiao-Xiao Niu, Ching Y. Suen, A novel hybrid CNN–SVM classifier for recognizing handwritten digits, Pattern Recognition, Volume 45, Issue 4, 2012, Pages 1318-1325, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2011.09.021.

[8] Basly, H., Ouarda, W., Sayadi, F.E., Ouni, B., and Alimi, A.M.: 'CNN-SVM Learning Approach Based Human Activity Recognition', in Editor (Ed.) 'Book CNN-SVM Learning Approach Based Human Activity Recognition' (Springer International Publishing, 2020, edn.), pp. 271-281

[9] T. Liu, X. Ye and B. Sun, "Combining Convolutional Neural Network and Support Vector Machine for Gait-based Gender Recognition," 2018 Chinese Automation Congress (CAC), Xi'an, China, 2018, pp. 3477-3481, doi: 10.1109/CAC.2018.8623118.

[10] Cao G, Wang S, Wei B, Yin Y, Yang G (2013) A Hybrid CNN-Rf Method for Electron Microscopy Images Segmentation. J Biomim Biomater Tissue Eng 18:114. doi:10.4172/1662-100X.1000114

[11] Haskard, K.B., Williams, S.L., DiMatteo, M.R., Heritage, J., and Rosen- thal, R.: 'The Provider's Voice: Patient Satisfaction and the Content- filtered Speech of Nurses and Physicians in Primary Medical Care', Journal of Nonverbal Behavior, 2008, 32, (1), pp. 1-20

[12] Zimmerman, P.H., Bolhuis, J.E., Willemsen, A., Meyer, E.S., and Noldus, L.P.: 'The Observer XT: a tool for the integration and syn- chronization of multimodal signals', Behav Res Methods, 2009, 41, (3), pp. 731-735

[13] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345–1359, 2009.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv: 1409.1556, 2014.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization,"arXiv preprint arXiv: 1412.6980, 2014.

[16] Breiman, "Random Forests", Machine Learning, 45(1), 5-32, 2001

[17] N. S. Altman (1992) An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression, The American Statistician, 46:3, 175-185, DOI: 10.1080/00031305.1992.10475879

[18] Md. Sahidulla and G. Saha, "Design, Analysis, and Experimental Evaluation of Black Based Transformation in MFCC Computation for Speaker Recognition," Journal of Speech Communication, Volume 54, pp. 543–565, May 2012, www.sciencedirect.com/science/article/pii/S0167639311001622?via%3Dihub.

[19] "MFCC, Extract mfcc, log energy, delta, and delta-delta of audio signal," Mathworks, [Online]. [Accessed: October 3, 2019].

[20] Friedberg, M.W., Chen, P.G., Van Busum, K.R., Aunon, F., Pham, C., Caloyeras, J., Mattke, S., Pitchforth, E., Quigley, D.D., Brook, R.H., Crosson, F.J., and Tutty, M.: 'Factors Affecting Physician Professional Satisfaction and Their Implications for Patient Care, Health Systems, and Health Policy', Rand Health Q, 2014, 3, (4), pp. 1-1

[21] Sinsky, C.A., Dyrbye, L.N., West, C.P., Satele, D., Tutty, M., and Shanafelt, T.D.: 'Professional Satisfaction and the Career Plans of US Physicians', Mayo Clinic Proceedings, 2017, 92, (11), pp. 1625-1635

[22] Babbott, S., Manwell, L.B., Brown, R., Montague, E., Williams, E., Schwartz, M., Hess, E., and Linzer, M.: 'Electronic medical records and physician stress in primary care: results from the MEMO Study', Journal of the American Medical Informatics Association, 2013, 21, (e1), pp. e100-e106

[23] Cousin, M.S.M.a.G.: 'The Role of Nonverbal Communication in Medical Interactions: Empirical Results Theoretical Bases and Methodological Issues' (2013. 2013)

[24] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[25] Y. Chherawala, P. P. Roy and M. Cheriet, "Feature Design for Offline Arabic Handwriting Recognition: Handcrafted vs Automated?," 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, 2013, pp. 290-294, doi: 10.1109/ICDAR.2013.65.

[26] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. 2014. Speech Emotion Recognition Using CNN. In Proceedings of the 22nd ACM international conference on Multimedia. Association for Computing Machinery, New York, NY, USA, 801–804. DOI:https://doi.org/10.1145/2647868.2654984

[27] R. Sa et al., "Intervertebral disc detection in X-ray images using faster R-CNN," 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Seogwipo, 2017, pp. 564-567, doi: 10.1109/EMBC.2017.8036887.