BeautyNet: A Makeup Activity Recognition Framework using Wrist-worn Sensor

Fatimah Albargi[1], Naima Khan[2], Indrajeet Ghosh[2], Ahana Roy[3]
[1]*Umm Al-Qura University, Saudi Arabia*
[2]*Department of Information Systems, University of Maryland, Baltimore County, United States*
[3]*Howard County Public School System, Maryland, United States*
[1]fmoqurmti@uqu.edu.sa, [2](nkhan4, indrajeetghosh)@umbc.edu, [3]ahanarai2020@gmail.com

***Abstract*—The significance of enhancing facial features has grown increasingly in popularity among all groups of people bringing a surge in makeup activities. The makeup market is one of the most profitable and founding sectors in the fashion industry which involves product retailing and demands user training. Makeup activities imply exceptionally delicate hand movements and require much training and practice for perfection. However, the only available choices in learning makeup activities are hands-on workshops by professional instructors or, at most, video-based visual instructions. None of these exhibits immense benefits to beginners, or visually impaired people. One can consistently watch and listen to the best of their abilities, but to precisely practice, perform, and reach makeup satisfaction, recognition from an IoT (Internet-of-Things) device with results and feedback would be the utmost support. In this work, we propose a makeup activity recognition framework, BeautyNet which detects different makeup activities from wrist-worn sensor data collected from ten participants of different age groups in two experimental setups. Our framework employs a LSTM-autoencoder based classifier to extract features from the sensor data and classifies five makeup activities (i.e., applying cream, lipsticks, blusher, eyeshadow, and mascara) in controlled and uncontrolled environment. Empirical results indicate that BeautyNet achieves 95% and 93% accuracy for makeup activity detection in controlled and uncontrolled settings, respectively. In addition, we evaluate BeautyNet with various traditional machine learning algorithms using our in-house dataset and noted an increase in accuracy by $\approx$ 4-7%.**

***Index Terms*—Makeup Activities, Wrist-worn sensors, Internet of Things, Machine learning, Activity recognition**

I. Introduction

The Internet-of-things is an emerging technology that has a high potential to impact various domains. Activity Recognition (AR) is a research area that thrived with the development of IoT wearables and showed great potential to contribute to many fields such as sport [1], [2], healthcare [3]–[6], activity recognition [7]–[9], makeup facial detection [10], [11]. The fusion of various sensors such as motion (IMU), skin conductance (GSR) and heart rate sensors provide ample opportunities to discover the activity execution pattern and enable us to investigate the prospects for future enhancement. *Activity Recognition* is still emerging and growing because of the exponential growth in the number of commercial-off-the-shelf wearable sensors available in the market. Besides, the advancement of deep learning techniques has intrigued researchers to excel in the field. A major characteristic of deep learning architecture is that it learns generic features from the high-dimension dataset. Deep learning techniques prove to aid the activity recognition tasks [1], [9] compared to the traditional machine learning algorithms. However, makeup activity detection is one such field which is still unexplored. So in this work, we work towards filling this gap.

Makeup activity is one of the fields where automated assistive tools can enable self-training and ensure perfection for all levels of users. In a recent massive revolution of content creation, there are over 3M makeup channels available on youtube. There can be found thousands of videos for creating similar makeup looks by multiple users using various sets of makeup tools and products. However, there is no strategic framework for comparing these techniques and styles except visual comparison. A data-driven recognition framework with feedback would be beneficial for beginners and visually impaired people to improve their makeup activity performance. Moreover, although, makeup products and tools are one of the founding sectors of the fashion industry, there has been very little research done in this area from user perspectives. Previously, there has been some research reported that identifies makeup from visual images for facial recognition and security purposes [10], [12]. But there has been no research done yet to the best of our knowledge that analyzes makeup activities from data-driven perspectives.

In this work, we extracted generalized features using LSTM-autoencoder based network and analyzed the patterns of performing makeup activities by different users. The encoded features from LSTM-autoencoder demonstrates the similarity and dissimilarity among different users' style in performing certain makeup activity. We experimented makeup activities in two settings. In one setup, we provided instructions for performing each makeup activity while in another setup, we provided no instructions. From the extracted features in both setups, we presented the makeup activity recognition performance. Our approach presented in this work can further be extended for pointing out the steps where improvements can be done by the users.

Key contributions of the paper are summarized as follows:

- ***BeautyNet***: introduced a novel data-driven approach to recognize makeup activities from wristsensor data by employing LSTM-based auto-encoder network. We also

identified the makeup activities which are most critical for data-driven recognition.

- We demonstrated the imparity in the styles of participating users while performing the makeup activities. We also visualize the similarity and dissimilarities among the users for each activity performed by themselves.
- ***Evaluation***: We evaluate the proposed BeautyNet with our in-house Makeup Activity Recognition (MAR) dataset, comprising 10 participants for 5 makeup activities. We designed two data collection environment settings: controlled (instructions are provided) and uncontrolled (instructions were not provided). We observe that BeautyNet achieves an F1-score of **95%** and **93%** for detecting makeup activities under controlled and uncontrolled settings, respectively.

The paper is organized in the following: section II presents the relevant works, section III discusses the overall framework, section IV presents the experimental results and section VI and V concludes the paper and provide some limitations and potential future work.

II. Related Works

This section reviews and summarizes the related works on Makeup-based activity recognition. We categorize the related work based into:- *human activity recognition using wearable sensors* and *makeup style detection via deep learning*.

- **Activity Recognition (AR) using Wearable Sensors:** Activity Recognition has been a heavily investigated topic in the past decade, especially after deep learning gained popularity due to the ability to learn the representation from the high-dimension sparse dataset. The sensors used in AR research are Inertial Measurement Units (IMU) (Accelerometer, Magnetometer and Gyroscope). Because of privacy concerns, IMU-based sensors have proven to be trustworthy compared to other modalities for activity recognition. IMU-based sensors are used across various fields: sports analytics [13], [14], dance recognition [9], gait analysis [15] and makeup activities detection [11]. In [9], the authors investigate recognizing and assessing indian dance movements from multi-channel sensors and evaluated both shallow learning classifiers as well as deep learning based techniques. This work paves the way to develop a virtual dance trainer that can provide feedback to a dance learner.

  Furthermore, IMU sensors allow us to capture and analyze the significant change in IMU signal while performing the activities. In study [13], the authors proposed a deep learning-based player profiling framework. The work aims to provide a synopsis to evaluate the soccer player's performance and devise attribute profiling techniques to improve the player's competence. They designed Restricted Boltzmann Machine (RBM) to classify the various soccer shots. They also detect the change point detection in statistical analysis to develop attributes of profiling the players. However, the scoring specified in their study refers to the player's ability in the soccer game. Therefore, they constitute attributes of technical judgment, tactical awareness, and physical aspects to develop a personalized soccer profiling module. The work showed promise in extending the use of wrist-worn smartwatches past providing basic analytical information providing more analytical results and how this extended across various domains.

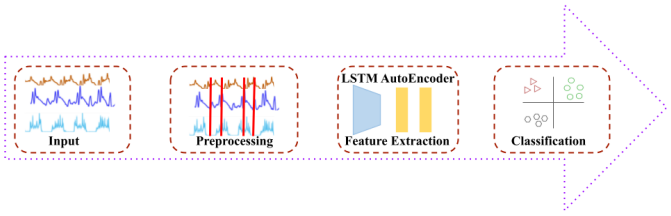


Fig. 1. Overall BeautyNet pipeline

- **Makeup Style detection via Deep learning:** Makeup style recognition has been investigated extensively and explored across facial recognition tasks [16]. State-of-the-art literature focuses mostly on classifying the facial changes occurring when makeup is applied. A cosmetic modification may modify the face's perceived appearance effects like ageing, natural biological change, and plastic surgery, a medically induced change, and other forms of alterations. In study [12], the authors propose a novel-makeup detection approach that assists face recognition systems in achieving high accuracy in detecting makeup images. In their work, makeup features are defined using biologically inspired features such as, average skin tone (AST) and texture of skin extracted from color and histogram oriented gradient (HOG) of the image.

  Similarly, in [17], the authors' demonstrated automated human face recognition algorithms can significantly degrade in recognizing the same image under makeup and disguised appearances and that can be easily scaled and deployed for security and surveillance tasks. First, they segment the image and capture region-of-interest (ROI). Then, they applied two popular image quality enhancement methods: single scale retinex (SSR) and quotient enhancement to eliminate or suppress the influence of external factors variations. Other studies [18], [19] focus on investigating makeup transfer methods where the model can distinguish the makeup/no-makeup facial images. In contrast, our work distinguishes limb pattern movements across various makeup activities. In our work, we utilized a wrist-worn wearable sensor to capture each participant's minute and distinctive hand movements to recognize makeup activity. The *overarching goal* of this work is to capture and study the hand movement patterns when the participants perform in a *controlled setting* (instructions given) and an *uncontrolled setting* (no instruction given).

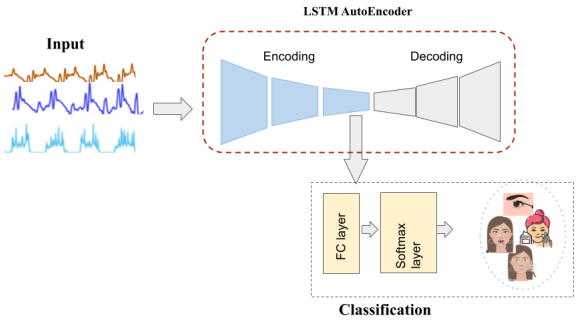


Fig. 2. LSTM-based autoencoder based classifier

III. BeautyNet: Recognizing Makeup Activities

This section enumerates and highlights the overall methodologies adopted for makeup-based activities. Our BeautyNet framework for makeup activity recognition is shown in Fig. 1 and 2. First, we collected the data from ten participants using the Microsoft band smartwatch and the companion smartphone application. Secondly, we considered accelerometer and gyroscope data for detecting the five makeup activities (i.e., applying cream, lipsticks, blusher, eyeshadow, and mascara). Lastly, we designed our data collection in two settings: i) controlled and ii) uncontrolled environment. In a controlled environment, we provided instructions for performing each makeup activity to the participants, while in an uncontrolled environment, participants performed activities without any instructions.

We preprocessed the raw data by removing the missing values and prepared the dataset for deep learning-based classification model. We divided the data into segments of certain window length and passed the segments to the LSTM-based autoencoder model for feature extraction. The LSTM-based autoencoder consists of three encoding layers and three decoding layers. The autoencoder network optimizes the reconstruction loss of multivariate time series and extracts features for each window in the series. Later, autoencoder network is followed by classification network consisting of a dense and a softmax layer for makeup activity recognition. Figure 2 provides the high-level architecture of the model.

We pretrained the autoencoder network in order to utilize the encoded representation in the final classification network. Later, we used the encoded representation to analyze the features for each makeup activity performed by different participants. The dimension of the encoded representation obtained from the autoencoder network is 16. We also applied other shallow learning algorithms (i.e., Naïve Bayes, K-Nearest Neighbors, Random forest, Gradient Boosting) for makeup activity classification.

IV. Experimental Results

In this section, we present the experimental results for makeup activity recognition and discuss the behavior of performing these activities by different participants.

A. Data Acquisition and Experiment setup

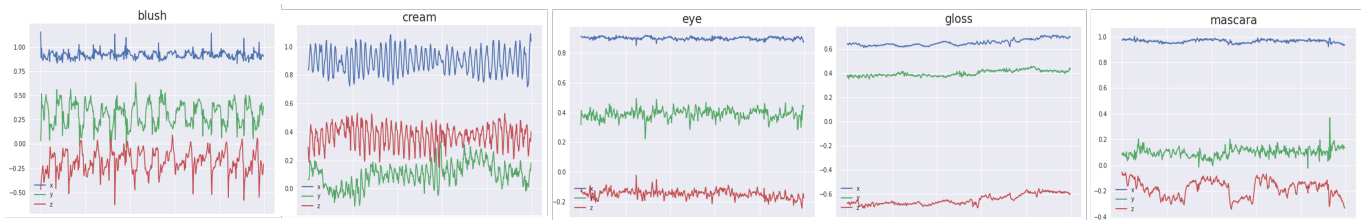
We collected data from ten participants for the five make-up activities in this study. We divided the participants into two categories: participants who perform the experiments in a controlled environment and participants who perform the experiments in an uncontrolled environment. We assigned five different participants for each group. For the controlled environment, we provided the participants with instructions about how each activity should be performed. We ensured the participants follow the instructions correctly. Table I presents the instructions given to the participants in a controlled environment.

TABLE I
Instructions in controlled environment

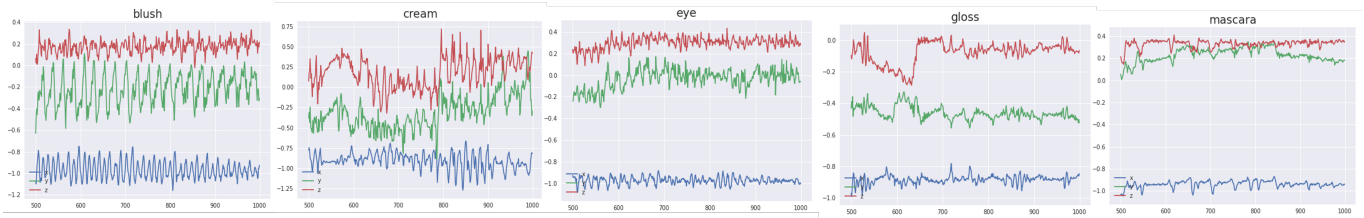
Activity	Instructions
Applying cream	Apply to the whole face in a circular movement that starts from the right cheek and continues to forehead, left cheek, and lastly the chin
Applying gloss	Apply to the lower lip first, then to the upper lip
Applying blush	Apply forward and backward in straight movements
Applying eyeshadow	Apply forward and backward in straight movements starting from right to left.
Applying mascara	Apply only for upper lashes.

For the uncontrolled environment, we allowed the participant to perform the five make-up activities in their own way with no instructions provided. Although, in real life, each make-up activity could take less than a minute to be applied, we wanted to collect data for five minutes duration with each activity. Therefore, we asked the participants to continue each activity for a five-minute period and to follow the sequence of the activities. We further stipulated that the participants perform three of the activities (applying blush, eyeshadow, and mascara) on one cheek and one eye for both controlled and uncontrolled environments. This helps us focus on the exact movement of performing each activity and avoiding any unnecessary movement of the participant's hand for changing the application area. We collected data using the Microsoft-band smartwatch. Figure 4 shows one participant wearing a Microsoft band in her left hand while performing the activities.

We presented a segment of accelerometer signal for participants from both controlled and uncontrolled environments while performing the activities in figure 3. The first row in figure 3 shows the accelerometer signal for a controlled environment, while the second row shows the accelerometer signal for an uncontrolled environment. We can observe the forward and backward hand movements for applying blush activity are similar in a controlled and uncontrolled environment. In



(a) Controlled



(b) Uncontrolled

Fig. 3. Accelerometer signals during makeup activities

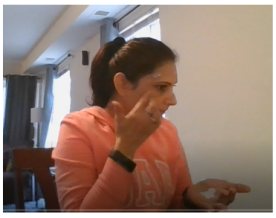


Fig. 4. Example of participant applying cream

a controlled environment, the hand movement for applying cream, eyeshadow, gloss, and mascara is lower than that in an uncontrolled environment.

B. Evaluation

Here we present the performance of different shallow learning algorithms and LSTM-based classifier for recognizing the makeup activities from accelerometer and gyroscope signals. Following evaluation metrics are used in Table II.

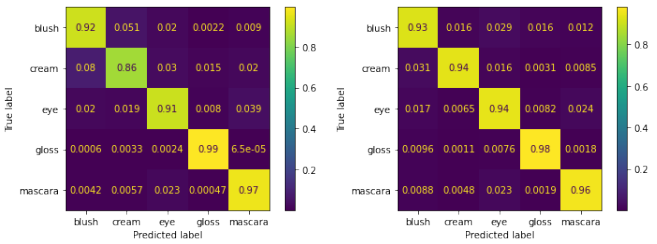
(i) Precision $(P) = TP/(TP + FP)$, (ii) Recall $(R) = TP/(TP + FN)$, (iii) F-1 Score = $2 \times (P \times R)/(P + R)$ and, (iv) Accuracy = $(TP+TN)/(TP+TN+FP+FN)$, where TP, FP, TN, and FN are the number of instances of true positive, false positive, true negative and false negative, respectively. We observe higher evaluation metrics for LSTM-based classifier in the results showed in the table II.

The evaluation metrics in table II suggests LSTM-based classifier outperforms other shallow learning algorithms. Among the shallow learning algorithms, random forest, knn and gradient boosting performs better.

We present the confusion matrix for the LSTM-classifier in figure 5. From figure 5(a), we observe in uncontrolled environment, 8% of instances of applying cream is predicted as applying blush. Besides, 5% instances of blush activity are classified as eyeshadow activity for the similarity of dragging and dabbing between these two activities. Approximately, 4% instances from eyeshadow activity are predicted as mascara activity and 2% instances from mascara activity are predicted as eyeshadow activity due to the closeness of application area. However, the number of true positive instances for applying cream activity is lower than that of other activities. It can

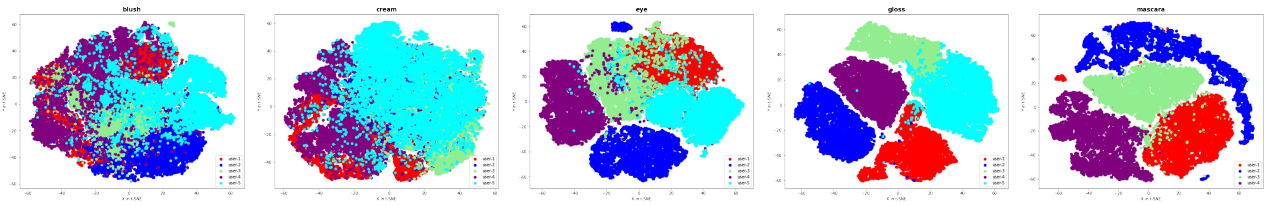
TABLE II
EVALUATION METRICS

Controlled Environment				
Classifier	Accuracy	Precision	Recall	F-measure
Naive Bayes	69.5%	0.72	0.69	0.69
KNN	92.4%	0.92	0.92	0.91
Random Forest	90.8%	0.91	0.91	0.91
Gradient Boosting	92.7%	0.92	0.92	0.92
LSTM-classifier	95.0%	0.94	0.95	0.95
Uncontrolled Environment				
Classifier	Accuracy	Precision	Recall	F-measure
Naive Bayes	64.50%	0.66	0.64	0.64
KNN	83.50%	0.83	0.83	0.83
Random Forest	89.20%	0.89	0.89	0.89
Gradient Boosting	87.80%	0.87	0.87	0.87
LSTM-classifier	93.0%	0.93	0.93	0.93



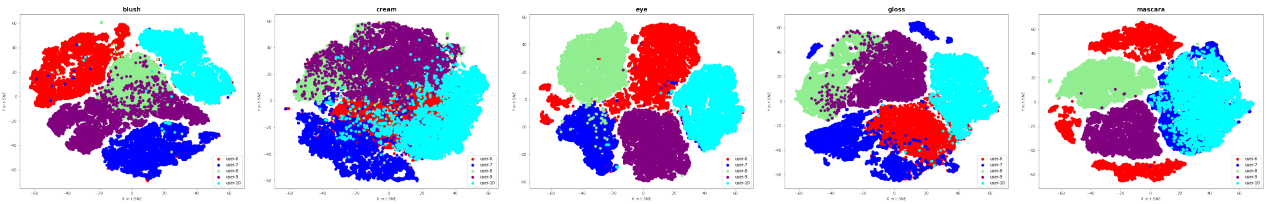
(a) Uncontrolled (b) Controlled

Fig. 5. Confusion metrics for LSTM classifier



(a) blush (b) cream (c) eye (d) gloss (e) mascara

Fig. 6. Features from participants in controlled environment



(a) blush (b) cream (c) eye (d) gloss (e) mascara

Fig. 7. Features from participants in uncontrolled environment

be explained from the fact that, cream activity applies to whole face which more or less overlaps the application area of other makeup activities. The participants applied cream cheeks, forehead, chin area as well as avoided the lip area during applying cream. Most hand movements happened on the cheek area during cream activity. This explains the 8% of cream activity instances falsely predicted as blush activity, as well as 3% predicted as eyeshadow activity.

We observe different scenarios in a controlled environment from figure 5(b). The number of true positive instances for applying cream activity is higher in a controlled environment compared to an uncontrolled environment. In a controlled environment, participants strictly followed the sequence of application areas specified in the instructions. However, even after careful performance, 3% instances of cream activity are predicted as blush activity, and 1% instances of blush activity are predicted as cream activity. The number of false predictions between applying eyeshadow and mascara is also low in a controlled environment.

C. Analysis on participants

In this subsection, we discuss how different participants performed the five makeup activities from the encoded features of each activity by corresponding participants. Figure 6 and 7 presents the T-SNE representations of the encoded features for each makeup activity category in controlled and uncontrolled environment, respectively. From the extracted features for each participants, we can infer the similarity and dissimilarity in performing makeup activities among the participants. For blush and cream activity, features from all corresponding participants are almost inseparable while for other three activities features from different participants are separable. This refers to the fact that all participants almost similarly applied blush and cream while each of them have different styles in performing other three makeup activities i.e., applying eyeshadow, gloss and mascara. In figure 6(a), we observe features from different users overlaps for blush activity. Cream activity features for user-5 are similar to those of user-2 and user-3 while user-1 and user-4 differs. For applying eyeshadow, the features for user-2 is not overlapped with the features from other users while features for user-1, user-3, user-4, user-5 are partially overlapped among each other. Very few overlaps are observed among the users for applying gloss and mascara in controlled environment. In figure 7(b), we observe for blush activity in uncontrolled environment, features of five participants are more separable than that in controlled environment. There is partial overlaps only between the user-8 and user-9 in the blush activity. The features of cream activity for different users is almost inseparable like in controlled environment. It is obvious that in uncontrolled environment each participant has different style in applying eyeshadow as the features of this activity from different participants are almost completely separable from each other. User-7 and user-8 performs mascara activity similarly as their features completely overlaps.

D. Implementation details

We used a training and testing split of 80-20% and a training, testing, and validation split of 60-20-20% for the traditional machine learning and LSTMs-based classifier algorithms, respectively. Furthermore, the validation dataset is use to fine-tune the hyperparameters of the BeautyNet pipeline. The dimensions of intermediate encoding layers in the autoencoer network are 128, 64, 32 while in the decoding layers the order is reverse. The LSTM autoencoder network is pretrained with 25 epochs and the classifier network is trained with 50 epoch to reach to optimized loss functions i.e., mse and categorical cross-entropy, respectively. The experiment is conducted on a Linux server integrated with Intel i7-6850K CPU, 4x NVIDIA GeForce GTX 1080Ti GPUs and 64GB RAM. All the codes of data preprocessing, traditional machine

learning and deep learning algorithms are implemented with Python. Especially for deep learning, PyTorch libraries are used.

V. Limitation and Future Work

Employing our proposed framework, we can learn data-driven patterns of performing different makeup activities for each participants. However, in this work, we focused on detecting different makeup activities from wrist-worn sensor data and ignored various delicate hand movements (i.e., dabbing, dragging) during makeup activities. For further analysis on these activities, tracking hand movements plays an important role as it actually defines the unique style of each participants for applying a certain makeup item. Besides, although participants are given same instructions in controlled environment, few participants were not able to continue following the instructions properly for several reasons, such as, cramps, muscle spasms in fingers etc. for repeating the same activity. Nevertheless, through this work, we successfully learn the fundamental nature of different makeup activities for different participants. As, future work, we would enhance the proposed framework to track delicate hand movements during makeup activities in more detail and provide useful personalized feedback to improve the activities which could be beneficial for visually impaired users as well. Additionally, we would like to investigate sophisticated activity learning techniques, such as, self-taught learning, self-supervised learning etc. techniques. Lastly, we would like to incorporate dense captioning techniques of the actions and movements performed by the subjects to comprehend the associated actions and perform makeup activities efficiently.

VI. Conclusion

We introduce BeautyNet in this work which provides a machine learning-based data-driven approach to analyze makeup activities from user perspectives. We collected a novel dataset consisting of wrist-worn sensor data from ten different participants for five makeup activities. Through our extensive experiment results indicate that *BeautyNet* achieves **95%** and **93%** accuracy for makeup activity detection in controlled and uncontrolled settings, respectively. Our analysis of the extracted features reveals some interesting facts on the considered makeup activities as well as the pattern of performing these activities by participating users. This provides the motivation of investigating further in this regard. In future, we plan to extend this work to develop a feedback-providing system for novice users to improve their performance in various makeup activities.

VII. Acknowledgement

This research is supported by the NSF Research Experience for Undergraduates (REU) grant # CNS-2050999, NSF CAREER Award # 1750936 and U.S. Army Grant # W911NF2120076.

References

[1] A. Chakma, A. Z. M. Faridee, N. Roy, and H. S. Hossain, "Shoot like ronaldo: Predict soccer penalty outcome with wearables," in *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pp. 1–6, IEEE, 2020.

[2] R. Goodrich, "Accelerometer vs. gyroscope: what's the difference?," *Internet: http://www. imf. org/external/datamapper/NGDP_RPCH@ WEO/OEMDC/ADVE C/WEOWORLD*, vol. 3, 2018.

[3] S. R. Ramamurthy, I. Ghosh, A. Gangopadhyay, E. Galik, and N. Roy, "Star: A scalable self-taught learning framework for older adults' activity recognition," in *2021 IEEE International Conference on Smart Computing (SMARTCOMP)*, pp. 121–128, IEEE, 2021.

[4] J. Parkka, M. Ermes, P. Korpipaa, J. Mantyjarvi, J. Peltola, and I. Korhonen, "Activity classification using realistic data from wearable sensors," *IEEE Transactions on information technology in biomedicine*, vol. 10, no. 1, pp. 119–128, 2006.

[5] H. Li, A. Shrestha, F. Fioranelli, J. Le Kernec, and H. Heidari, "Hierarchical classification on multimodal sensing for human activity recogintion and fall detection," in *2018 IEEE SENSORS*, pp. 1–4, IEEE, 2018.

[6] L. Malott, P. Bharti, N. Hilbert, G. Gopalakrishna, and S. Chellappan, "Detecting self-harming activities with wearable devices," in *2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, pp. 597–602, IEEE, 2015.

[7] T. Rahman, M. Czerwinski, R. Gilad-Bachrach, and P. Johns, "Predicting" about-to-eat" moments for just-in-time eating intervention," in *Proceedings of the 6th International Conference on Digital Health Conference*, pp. 141–150, 2016.

[8] D. T. Nguyen, E. Cohen, M. Pourhomayoun, and N. Alshurafa, "Swallownet: Recurrent neural network detects and characterizes eating patterns," in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pp. 401–406, IEEE, 2017.

[9] A. Z. M. Faridee, S. R. Ramamurthy, H. S. Hossain, and N. Roy, "Happyfeet: Recognizing and assessing dance on the floor," in *Proceedings of the 19th International Workshop on Mobile Computing Systems & Applications*, pp. 49–54, 2018.

[10] T. Li, R. Qian, C. Dong, S. Liu, Q. Yan, W. Zhu, and L. Lin, "Beautygan: Instance-level facial makeup transfer with deep generative adversarial network," in *Proceedings of the 26th ACM international conference on Multimedia*, pp. 645–653, 2018.

[11] T. Alzahrani, B. Al-Bander, and W. Al-Nuaimy, "Deep learning models for automatic makeup detection," *AI*, vol. 2, no. 4, pp. 497–511, 2021.

[12] S. Rasti, M. Yazdi, and M. A. Masnadi-Shirazi, "Biologically inspired makeup detection system with application in face recognition," *IET Biometrics*, vol. 7, no. 6, pp. 530–535, 2018.

[13] H. S. Hossain, M. A. A. H. Khan, and N. Roy, "Soccermate: A personal soccer attribute profiler using wearables," in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pp. 164–169, IEEE, 2017.

[14] I. Ghosh, S. R. Ramamurthy, and N. Roy, "Stancescorer: a data driven approach to score badminton player," in *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pp. 1–6, IEEE, 2020.

[15] N. Y. Hammerla, S. Halloran, and T. Plötz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," *arXiv preprint arXiv:1604.08880*, 2016.

[16] X. Ma, F. Zhang, H. Wei, and L. Xu, "Deep learning method for makeup style transfer: A survey," *Cognitive Robotics*, vol. 1, pp. 182–187, 2021.

[17] T. Y. Wang and A. Kumar, "Recognizing human faces under disguise and makeup," in *2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, pp. 1–7, IEEE, 2016.

[18] H.-J. Chen, K.-M. Hui, S.-Y. Wang, L.-W. Tsao, H.-H. Shuai, and W.-H. Cheng, "Beautyglow: On-demand makeup transfer framework with reversible generative network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10042–10050, 2019.

[19] W. Jiang, S. Liu, C. Gao, J. Cao, R. He, J. Feng, and S. Yan, "Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5194–5202, 2020.