Emotion Recognition through Gait on Mobile Devices

Mangtik Chiu
mtchiu@connect.ust.hk

Jiayu Shu
jshuaa@cse.ust.hk

Pan Hui
panhui@cse.ust.hk

HKUST-DT System and Media Laboratory
The Hong Kong University of Science and Technology, Hong Kong

Abstract—**Building systems that have the ability to recognize human emotions has attracted much interest in recent years. Common approaches toward machine emotion recognition focus on detection of facial expressions and analysis of physiological signals. However, in situations where these features cannot be easily obtained, emotion recognition becomes a challenging problem. In this paper, we explore the possibility of emotion recognition through gait, which is one of the most common human behaviors. We first identify various motion features based on pose estimation of captured video frames. We then train several supervised learning models including SVM, Multilayer Perceptron, Naïve Bayes, Decision Tree, Random Forest and Logistic Regression using selected features and compare their performances. The best model trained to classify five emotion labels has an accuracy of 64%. Finally, we implement a proof of concept mobile-server system for emotion recognition in real life scenarios using mobile phone cameras.**

I. INTRODUCTION

The topic of emotion recognition has been introduced and researched on for a long time. It is reported that the state-of-the-art emotion recognition approach can reach up to 99% accuracy using Convolution Neural Networks [1]. These methods or systems mainly recognize emotions by analyzing facial features or physiological signals. However, the analysis of facial expression depends on clear detection of frontal faces, while detection of physiological signals requires delicate detection system in a noiseless environment. For example, in locations such as streets and shopping malls, where near placement of feature capturing device is inconvenient or infeasible, features required for the above approaches become difficult to obtain. On the other hand, recording human gait does not require a camera from a very close distance, which provides the opportunity for using motions cues in recognizing human emotions.

In general, emotion recognition from gait focuses on analyzing movement patterns when an individual is walking. When the individual possesses a certain emotion while walking, it is likely that the body motion would also be affected, such as changes in speed and posture. While there has been research on emotion recognition from motion, little work has been done to introduce it to practical situations where can be carried out any time without specific and technical setup. As a result, the progress of practical deployment is hindered by the difficulty of capturing human motion in public environments. To solve the practicality problem, we take a visual approach towards emotion recognition from gait, which only requires captured video of the target.

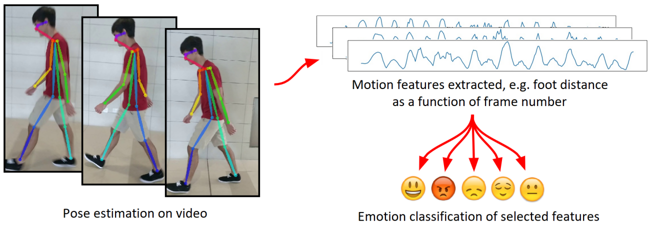
Motion features extracted, e.g. foot distance
as a function of frame number
Pose estimation on video
Emotion classification of selected features

Fig. 1. An overview of our method. We first obtain pose estimation results on each video frame. Motion features are then extracted and selected for emotion classification.

The workflow of our method is visualized in Figure 1. In order to train the model, we collect data by recording individuals walking under various emotions (i.e., joy, anger, sadness, relaxation and neutral) through a phone camera for about 10 seconds, where subjects are explicitly told the emotion to perform while walking. We then process the video by first feeding each frame into a pose estimation model, which returns 17 joint interest points per frame of the subject. The time-series motion data are then transformed into motion features by a set of manually defined parameters such as head inclination and stride distance. These features are further abstracted into statistical meta-features as the input of the classification model. We train 6 machine learning models, SVM, Multilayer Perceptron, Decision Tree, Naïve Bayes, Random Forest, and Logistic Regression using meta-features. The results show a best validation accuracy on collected dataset of 64%, while classification by human observers achieved 72% on average.

Our work has several contributions. First, we collect a dataset of 120 samples for model training. Second, we propose and extract systematic motion features from sequences of video frames. Third, we design and implement a mobile-server system that can be used in practical scenarios. The rest of the paper is organized as follows. In Section II, we summarize the related work. In Section III, we present our method to process data and recognize emotion. In Section IV, we describe the system implementation. In Section V and VI, we show the evaluation results and analysis. Finally, in Section VII we conclude our work.

II. Related work

A. Major Trends in Emotion Recognition

Learning facial key points on face images is a common approach towards emotion recognition. In [2], the system extracts facial features such as eyebrow and mouth. Positional information of the components are then fed into the classification model for emotion prediction. Another method analyzes the facial movements on multiple frames of a person's facial expression [3], which improves the performance compared to separate models in [4].

Furthermore, the availability of large dataset and the rapid development of deep learning enabled the implementation of even more advanced techniques in emotion recognition. Using Convolution Neural Networks, researchers are able to carry out end to end emotion classification from raw facial images, instead of manually defining features [1]. In addition to facial expressions, physiological signals have also been studied. A group uses radio signal reflected from an individual's body to extract heartbeat and respiration rate [5]. Moreover, fusion of electroencephalography (EEG) and musical features in music induced emotion recognition [6] shows significant improvement in accuracy, and fusion of EEG and eye movements [7] improves the accuracy of around 10% compared to single modality models.

While these methods show significant performances in detecting even implicit emotions, they require a close proximity to the subject and a fixed orientation of the features to be detected. This leads to the difficulties of carrying out emotion recognition in public areas where only mass monitoring is possible. The possibility of detecting emotion from motion therefore becomes one way of assisting emotion recognition in some situations such as streets and shopping malls.

B. Emotion Recognition from Gait

Researches have successfully shown that emotions can be differentiated and labeled by human observers via monitoring body movements of subjects [8], [9]. A study on emotion recognition defines several features from gait over motion-captured data, such as joint angles with 3 to 6 degrees of freedom, speed, and frequency [10]. It is also suggested by human observers that in addition to straight-forward features like speed, complex and intricate amalgam of body postures throughout are also keys to correct emotion classification. Therefore, to effectively classify emotion from gait, combinations of motion features must be taken into consideration.

The advancement of machine learning algorithms opened up the possibility of emotion classification via supervised learning. In [11], 99% accuracy is achieved by using fixed-length time series data for step features with a Multilayer Perceptron. Their work, however, requires dedicated hardware and precise environmental setup. Another work done on the topic defines data abstraction features such as "Quantity of Motion", which attempts to quantify an individual's degree of movement in an image sequence [12]. This work also proposes further transformations of temporal data into consistent statistical sets of meta-features such as average and amplitude. This approach effectively maintains the fixed feature length and generally lower dimensionality of input data, hence allows for simpler model implementation and more efficient training. Our work refers to this work when defining motion features for emotion classification. However, instead of gathering motion data from static background without lower body movements, we collect both gesture and gait data of individuals' walking, which contain more information and is more applicable to realistic environments.

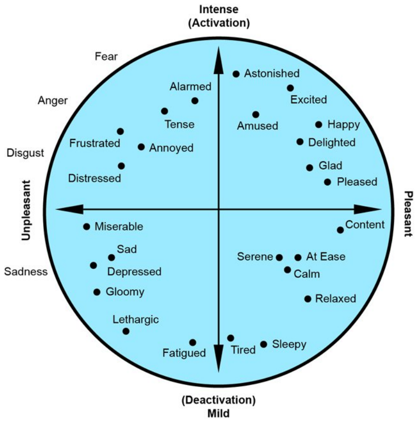

Fig. 2. Circumplex model describing the unidimensional scaling of emotion with respect to valence(horizontal) and arousal(vertical).

TABLE I

EMOTION TO AROUSAL/VALENCE MAPPING ADOPTED FROM THE CIRCUMPLEX MODEL

Emotion	Joy	Anger	Sadness	Relaxation	Neutral
Arousal	High	High	Low	Low	Medium
Valence	High	Low	Low	High	Medium

III. METHODOLOGY

In this section, we first introduce the emotion model used in our work. We then describe the data collection process and feature extraction method. We finally present feature selection and various machine learning models used in the training step.

A. Emotion model

Researchers have been discussing whether human emotion can be characterized discretely or continuously.The discrete emotion theory suggests that distinct features, including facial expressions, physiological features, and behaviors indeed correspond to a set of emotions [13], [14]. However, some argue that discrete affections are constructed from more fundamental elements, such as "internal sensation". [15]. The internal sensation element describes the level of pleasure and arousal of an emotion.

While discrete emotions can be simply labeled, continuous emotion models is not a single dimensional feature. Effective modeling of emotions based on degree of arousal and pleasure can be achieved by employing the Circumplex Model [16]. This 2-dimensional matrix model classifies emotions according to 2 parameters, valence and arousal, which reflect the intrinsic positivity and the activeness perceived by the subject respectively.

In this work, we setup emotion classification labels listed in Table I based on the circumplex model as shown in Figure 2. In consideration that individuals may not always

TABLE II

POSITION INDEX AND THEIR RESPECTIVE JOINTS

Index	Description	Index	Description
0	Nose	9, 12	Knees
1	Neck	10,13	Ankles
2, 5	Shoulders	14,15	Eyes
3, 6	Elbows	16,17	Ears
4, 7	Wrists		

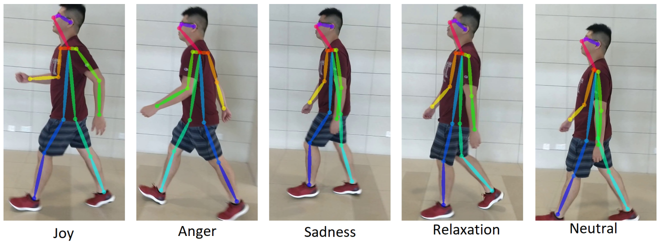

Fig. 3. Sample output of the 16 feature points(excluding right ear and right eye) returned by OpenPose. Selected frames are at the most extended gesture. Note the difference in hand distance and head inclination among emotions.

possess expressive emotion, we incorporate an extra "Neutral" model indicating that no specific emotion is observed.

B. Data collection

The dataset we collect involves 11 participants who are university students. Two to three videos are recorded per emotion per individual for the aforementioned five emotion categories. Following previous work done on motion-driven emotion recognition [10], [17], [11], participants are explicitly instructed to perform gait with an emotion. Subject feedback is therefore unnecessary due to a more strictly controlled experimental setup. The main purpose of this dataset is to directly sample explicit motion traits and compare machine emotion recognition performance from gait to that of a human observer. All videos are recorded by a mobile phone camera from the left side of the subject. Videos are recorded in 30FPS with 1080p resolution and 16:9 aspect ratio. The first and last one-sixth portion of each video is trimmed to skip preparation and finishing gesture that may deviate from normal gait motion.

C. Pose estimation

Pose estimation is employed to obtain positional information of the subject to calculate motion features. We use the OpenPose model [18] for single person pose estimation on each frame obtained from the videos. The model outputs 18 joint positions in pixel xy-coordinate format as listed in Table II. As the videos are recorded in a tangential angle with respect to the subject, joints features on the right, specifically right ear(16) and right eye(14) are disregarded when we process the data in the next step. After carrying out pose estimation on all frames of a video, we obtain 17 lists of time-series data. Figure 3 shows the results obtained by the OpenPose model. We observe that joint occlusions and errors in pose estimation can result in zero-value joint coordinates. To recover lost data, a spiker filter is implemented to detect abrupt gradient change within one to two time windows. The lost data point is then reconstructed through linear interpolation.

TABLE III

EUCLIDEAN DISTANCE AND ANGULAR FEATURES

Euclidean	Description
Hand	Distance between two hands
Hand/Hip	Distance between left hand and hip
Knee	Distance between two knees
Knee/Hip	Distance between left knee and left hip
Elbow	Distance between two elbows
Foot	Distance between two feet
Area	Minimum bounding box of the subject
Angle	Description
Arm	Angle between 2 arms
Left Elbow	Left elbow flexion
Right Elbow	Right elbow flexion
Leg	Angle between 2 legs
Left Knee	Left knee flexion
Right Knee	Right knee flexion
Abs Knee	Vertical angle of the front thigh
Rel Knee	Angle between front thigh and torso
Abs Head	Vertical angle of head inclination
Rel head	Angle between head and torso

D. Feature extraction

We define 18 motion features by using spatial temporal data obtained from the pose estimation output, including Euclidean features, Angular features and Speed. We list the euclidean and angular features in III. These features are then transformed into meta-features as final inputs to the classification model. In total, we obtain 158 meta-features. The details are described as below.

1) Euclidean and Angular Features: Euclidean features are defined as the L2 distances between joint positions. To transform our data into record-distance-invariant and scale-uniform features, all euclidean distances are normalized by the height of the bounding box surrounding the subject in pixels.

Angular features are defined as the angle of joint positions or vectors. Absolute angles are measured with respect to the horizontal ground. Relative angles are the acute angles between three joint positions.

2) Meta-features: A number of meta-parameters are used to convert the above spatial-temporal data into meta-features. These parameters find the general trend of the time-series data by extracting statistical measures of the time-series data. Table IV lists the meta-features details.

3) Speed: The walking speed of the subject is approximated by the sum of step distances of the subject divided by the number of frames. The total walking distance S can be calculated as:

$$S = \sum d_i \tag{1}$$

$\boldsymbol{d_i}$ denotes the i-th local maximum foot distance. The average total speed $\boldsymbol{V_{avg}}$ is then calculated by dividing the

TABLE IV

META FEATURES

Meta-feature	Description
Max	Global maximum value
Min	Global minimum value
Avg	Mean value
Std	Standard deviation
Avg Max	Average of local maxima
Std Max	Standard deviation of local maxima
Pre-slope	Slope of the line from global maximum to the preceding minimum
Post-slope	Slope of the line from global maximum to the following minimum
Peak duration	Number of timestamps from peak containing global maximum

TABLE V

TOP 15 SELECTED META-FEATURES IN TERMS OF F VALUE ON THE DATASET FOR DISCRETE AND AROUSAL/VALENCE MODEL. THE PLUS SIGNS(+) INDICATES THE NUMBER OF OTHER META-FEATURES FROM THE SAME MOTION FEATURE RANKED IN THE TOP 15.

Feature	Meta-feature	f value	p value
Emotion ($f_{crit} = 2.451$, 71 features)			
Relative head angle	Avg, +2	33.484	1.656e-18
Absolute head angle	Avg, +3	27.708	4.104e-16
Hand distance	Local Avg, +1	21.849	1.975e-13
Leg angle	Avg	18.286	1.189e-11
Knee distance	Local Avg, +1	15.606	3.154e-10
Left elbow angle	Std	14.859	8.127e-10
Foot distance	Avg, +1	14.428	1.412e-9
Arousal ($f_{crit} = 3.074, 72features$)			
Leg angle	Avg, +3	32.978	4.382e-12
Knee distance	Local Avg, +2	29.990	3.055e-11
Left elbow angle	Std, +1	29.597	3.967e-11
Foot distance	Avg, +1	28.882	6.391e-11
Hand distance	Local Avg, +1	27.594	1.523e-10
Hand-Hip distance	Pre-slope	18.488	1.054e-7
Speed	Min	16.404	5.247e-7
Valence ($f_{crit} = 3.074, 23features$)			
Absolute head angle	Avg, +4	53.864	2.611e-17
Relative head angle	Avg, +5	53.308	3.491e-17
Hand distance	Avg, +2	11.814	2.122e-5
Right elbow angle	Local Avg	5.1414	7.242e-3

total walking distance by the number of frames:

$$V_{avg} = \frac{S}{N} \tag{2}$$

N denotes the total number of frames. The resulting speed has a unit of "subject height per frame". Per step speed information is also calculated as a comparison to the average total speed. Specifically, each step speed $\boldsymbol{V_i}$ is calculated as:

$$V_i = \frac{d_i}{t_i} \tag{3}$$

$\boldsymbol{t_i}$ denotes the number of frames from the i-th foot distance local maximum to the (i+1)-th foot distance local maximum.

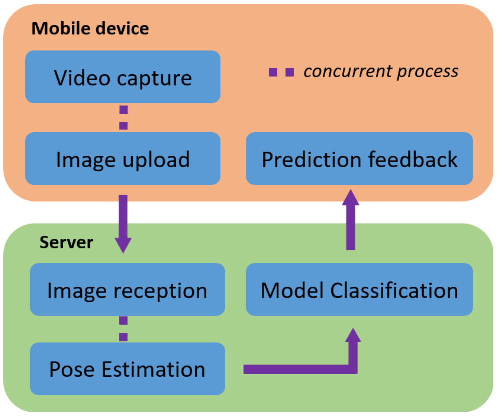

Fig. 4. System architecture diagram.

Since $\boldsymbol{V_i}$ is now a type time-series data, it is then transformed into meta-features. As a result, 5 speed meta-features, $\boldsymbol{V_{avg}}$, $\boldsymbol{avg(V_i)}$, $\boldsymbol{std(V_i)}$, $\boldsymbol{min(V_i)}$, $\boldsymbol{max(V_i)}$ are returned as speed-meta features.

E. Model training

We train and compare six models, including Support Vector Machines (SVMs), Multi-layer Perceptron (MLP), Decision Tree, Naïve Bayes, Random Forest, and Logistic Regression. To avoid possible overfitting caused by imbalanced data, we perform data under/over-sampling. We also use Bootstrap Aggregating technique, which improves model stability by training multiple instances of a model and returns the majority output. We carry out one-way ANOVA using critical f-score thresholding to lower the dimensionality of the obtained 158 features, which reduces the probability of over-fitting. Table V lists the top five meta-features that have the higher f-scores.

In practice, classification accuracy may be affected by image resolution and video frame rate when recording videos using mobile devices. To investigate the effect, we carry out validation data modification and feed them into models trained with the original training data. Specifically, we train models using 1080p 30FPS image sequences, then evaluate the performance of the model on image sequences with lower resolutions and frame rates. This allows us to find the optimal recording configuration to minimize network load while maintaining classification accuracy.

IV. SYSTEM IMPLEMENTATION

We design and implement an emotion recognition system composed of a client and a server. The system architecture is shown in Figure 4. The client side runs on a mobile phone with camera and network functionality. In our implementation, we use a OnePlus 5 mobile phone with the 16 and 20MP dual camera for video recording. The processing server is a desktop computer with Intel i7-7700 CPU and one Nvidia GTX 1080Ti GPU. When a user wishes to infer others' emotions through their gaits, the device connects to the server and begins uploading each captured image in real time. Upon completion of emotion classification using the recognition model, the server returns the predicted emotion

TABLE VI

ACCURACIES OF EMOTION CLASSIFIERS. "SINGLE" STANDS FOR WITHOUT BOOTSTRAP AGGREGATING.

Estimator	Accuracy (Single)	Accuracy (Bagging)
Human observer	0.72	
SVM	**0.621**	0.573
MLP	0.598	0.612
Naïve Bayes	0.553	0.523
Decision Tree	0.533	0.536
Random Forest	0.613	
Logistic Regression	0.569	

back to the mobile device, where the final emotion prediction results will be rendered to the user.

To balance model performance and system response time, video resolution and frame rate can be adjusted. Though system efficiency will be improved if less data is sent for processing, by reducing image resolution or video frame rate, the information is lost at the same time. Resolution of the video does not affect extracted motion features, yet frame skipping results in inconsistent time-dependent motion features such as speed. Therefore, linear interpolation is carried out on obtained joint position data as an attempt to reconstruct original motion data. The detailed emotion recognition performances with regard to different image resolutions and video frame rates are presented in the next section.

V. EVALUATION

In this section, we first present the model performances. We then evaluate the implemented system in terms of recognition accuracy and response time under various image configurations, in order to discover a video configuration which minimizes network load but still maintains accuracy.

A. Model Accuracy

Table VI shows the average accuracies of the six models respectively. We use 12-fold cross validation as the samples are limited. To obtain stable accuracy values, we take the average accuracy over 20 iterations for each model. The average accuracy is therefore defined as :

$$Accuracy = \frac{1}{20}\sum_{i}^{20}\frac{Correct\ prediction\ at\ iteration\ i}{Total\ test\ samples\ at\ iteration\ i} \tag{4}$$

The model with the best performance achieves 62.1% accuracy, which is a single SVM with radial basis function kernel. MLP with Bagging and Random Forest achieve second and third performances, with 61.2% and 61.3% accuracies respectively. We therefore use these three best performing models in the following system evaluations.

B. Resolution and FPS

While feature extraction and emotion prediction on the server side are extremely efficient, image transmission and pose estimation are relatively time-consuming, and are significantly affected by frame rate and resolution. To discover the best video configuration without giving up too much performance, we carry out various tests with regards to different video FPS and frame resolutions. We use the emotion recognition model trained on 1080p and 30fps image sequences to test different video configurations.

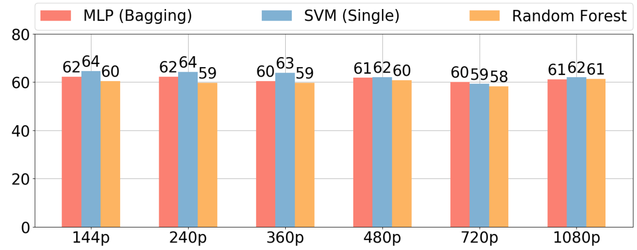

Fig. 5. Accuracies of MLP, SVM and Random Forest with respect to test time Resolution.

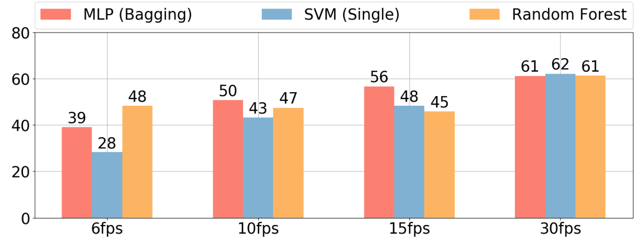

Fig. 6. Accuracies of MLP, SVM and Random Forest with respect to test time FPS. Interpolation is used to reconstruct lost frames.

Figure 5 shows the accuracies when tested against various resolutions. We find that model accuracies remain at around 60% despite significant image quality changes. It can therefore be concluded that resolution does not significantly affect prediction accuracy. This may result from the reason that pose estimation only requires a rough silhouette of the subject, which can still be clearly seen in low resolution images. As a result, to minimize network traffic and reduce data usage, 144p videos can be taken while maintaining classification accuracy.

Figure 6 shows the accuracies when tested against various frame rates. Decrease in model accuracies is observed when lower FPS videos are used. MLP and SVM perform better as the frame rate increases, while Random Forest shows the opposite trend except when FPS reaches 30. In general, using videos with FPS lower than 30 result in accuracy loss, which possibly results from information lost in motion features. Therefore, to maximize classification accuracy, video FPS should remain at 30 using current feature extraction methods.

C. Response Time

We measure the response time using different combinations of image resolution and FPS, in order to provide a reference to the networking performance of the system. The results are the averages over 10 measurements. Since our server is built on a lab computer with restricted connection access, to simulate public connection, we migrate and

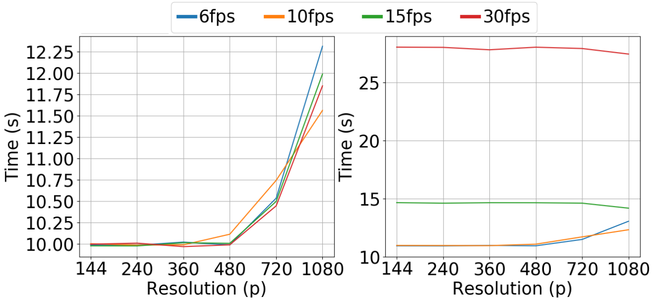

Fig. 7. Left: 10-second video transmission time, from first frame sent to last frame received. Right: corresponding total system response time, from first frame sent to emotion prediction completed.

perform image transmission test by connecting the smart phone with 22MB/s upload speed to an AWS EC2 t2 Micro instance. We run the remaining system on the lab server by replaying videos according to reception delay of each frame.

Figure 7 shows the response times of transmitting and processing videos recorded in 10 seconds. It is found that video upload time begins to prolong at 720p and 1080p, whereas videos with lower resolutions can be uploaded almost in real-time. However, significant delay in the overall system response time is observed when using 30FPS videos, since pose estimation essentially processes more than twice the amount of images at 30FPS. Therefore, more time is required to process the video after all frames are received.

Based on our evaluation results, if fast system response time is desired with little compromise on accuracy, we recommend to record videos with 15FPS and 144p resolution.

VI. Discussion and Future work

In model construction, one major challenge is data processing. Current state-of-the-art pose estimation algorithms do not include tracking functionality, which results in inconsistency of joint positions, such as swapping and fluctuations across subsequent frames. The employed pose estimation model is also prone to joint occlusion, which returns invalid joint coordinates when joints are overlapped. The two situations cause error and spikes in the time-series data extracted, and a special anomaly detection algorithm must be designed to minimize the final signal errors.

In terms of system construction, the time for video upload and pose estimation pose a challenge in minimizing the overall system response time. Contrary to video streaming, our current method does not allow frame skipping in order to maintain model accuracy. Therefore, to minimize latency caused by slow frame upload, techniques such as multi-path TCP should be considered.

VII. Conclusion

In this paper, we analyzed human gait through videos to recognize emotions without using complex hardware devices. We introduced a vision pipeline that extracts meta-features for emotion inferring. Performance comparison demonstrated the possibility of classifying emotions by simply recording videos of human gait. We also designed and implemented the system on mobile devices and a cloud server, which introduces convenient emotion recognition in public areas using mobile and wearable devices. We envision ample opportunities and applications of the system via mass emotion monitoring, including emotion-based product recommendation in shopping centers, smart public area atmosphere adjustment, and emotional anomaly detection. Future research on the topic can refer to this work as a baseline for building more advanced systems.

References

[1] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, “Dexpression: Deep convolutional neural network for expression recognition,” CoRR, vol. abs/1509.05371, 2015. [Online]. Available: http://arxiv.org/abs/1509.05371

[2] L. Luoh, C.-C. Huang, and H.-Y. Liu, “Image processing based emotion recognition,” in 2010 International Conference on System Science and Engineering, July 2010, pp. 491–494.

[3] A. Azcarate, F. Hageloh, K. V. D. S, and R. Valenti, “Automatic facial emotion recognition,” 2005.

[4] S. Emerich, E. Lupu, and A. Apatean, “Emotions recognition by speechand facial expressions analysis,” in 2009 17th European Signal Processing Conference, Aug 2009, pp. 1617–1621.

[5] M. Zhao, F. Adib, and D. Katabi, “Emotion recognition using wireless signals,” in Proceedings of the 22Nd Annual International Conference on Mobile Computing and Networking, ser. MobiCom ’16. New York, NY, USA: ACM, 2016, pp. 95–108. [Online]. Available: http://doi.acm.org/10.1145/2973750.2973762

[6] N. Thammasan, K. ichi Fukui, and M. Numao, “Multimodal fusion of eeg and musical features in music-emotion recognition,” 2017. [Online]. Available: https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14831

[7] Y. Lu, W.-L. Zheng, B. Li, and B.-L. Lu, “Combining eye movements and eeg to enhance emotion recognition,” in Proceedings of the 24th International Conference on Artificial Intelligence, ser. IJCAI’15. AAAI Press, 2015, pp. 1170–1176. [Online]. Available: http://dl.acm.org/citation.cfm?id=2832249.2832411

[8] J. M. Montepare, S. B. Goldstein, and A. Clausen, “The identification of emotions from gait information,” Journal of Nonverbal Behavior, vol. 11, no. 1, pp. 33–42, Mar 1987. [Online]. Available: https://doi.org/10.1007/BF00999605

[9] M. de Meijer, “The contribution of general features of body movement to the attribution of emotions,” Journal of Nonverbal Behavior, vol. 13, no. 4, pp. 247–268, Dec 1989. [Online]. Available: https://doi.org/10.1007/BF00990296

[10] G. Venture, H. Kadone, T. Zhang, J. Grèzes, A. Berthoz, and H. Hicheur, “Recognizing emotions conveyed by human gait,” International Journal of Social Robotics, vol. 6, no. 4, pp. 621–632, Nov 2014. [Online]. Available: https://doi.org/10.1007/s12369-014-0243-1

[11] D. Janssen, W. I. Schöllhorn, J. Lubienetzki, K. Fölling, H. Kokenge, and K. Davids, “Recognition of emotions in gait patterns by means of artificial neural nets,” Journal of Nonverbal Behavior, vol. 32, no. 2, pp. 79–92, Jun 2008. [Online]. Available: https://doi.org/10.1007/s10919-007-0045-3

[12] G. Castellano, S. D. Villalba, and A. Camurri, Recognising Human Emotions from Body Movement and Gesture Dynamics. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 71–82. [Online]. Available: https://doi.org/10.1007/978-3-540-74889-2_7

[13] S. Tomkins, Affect imagery consciousness: Volume I: The positive affects. Springer publishing company, 1962.

[14] ——, Affect imagery consciousness: volume II: the negative affects. Springer Publishing Company, 1963.

[15] L. Barrett, M. Gendron, and Y.-M. Huang, “Do discrete emotions exist?” vol. 22, pp. 427–437, 08 2009.

[16] J. Ressel, “A circumplex model of affect,” J. Personality and Social Psychology, vol. 39, pp. 1161–78, 1980.

[17] C. L. Roether, L. Omlor, A. Christensen, and M. A. Giese, “Critical features for the perception of emotion from gait,” Journal of Vision, vol. 9, no. 6, p. 15, 2009. [Online]. Available: + http://dx.doi.org/10.1167/9.6.15

[18] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in CVPR, 2017.