Emotion Recognition through Gait on Mobile Devices

Mangtik Chiu*, Jiayu Shu*, and Pan Hui*†
*HKUST-DT System and Media Laboratory, Hong Kong University of Science and Technology, Hong Kong
†Department of Computer Science, University of Helsinki, Finland
mtchiu@connect.ust.hk jshuaa@connect.ust.hk panhui@cse.ust.hk

***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 from 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 smartphone cameras.**

I. INTRODUCTION

The topic of emotion recognition has been discussed and researched for a long time. It is reported that state-of-the-art emotion recognition approaches can acheive 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, and the detection of physiological signals requires a delicate detection system in a noiseless environment. As a result, in locations such as streets and shopping malls, where it is infeasible or inconvenient to place feature capturing devices close to the people of interest, the above methods cannot work well. On the contrary, recording human gait does not require a camera to be placed at a very close distance, which provides the opportunity of using motions cues to recognize human emotions.

In general, emotion recognition from gait focuses on analyzing walking patterns. If an individual possesses a certain emotion while walking, it is likely that his or her body and/or body movements will also be affected, such as changes in posture and/or speed. Though there is some research that recognizes emotion from motion, little work has been done to introduce it in real-life situations, where it can be carried out any time without specific and technical setup. Therefore, to solve this practicality problem, we take a visual approach towards emotion recognition from gait, requiring only captured video of the target individuals.

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

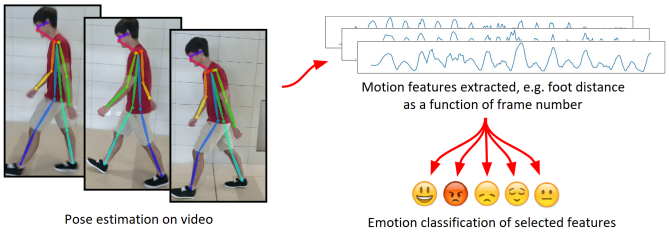

Fig. 1. An overview of our method. We first obtain a pose estimation result from each video frame. Motion features are then extracted and selected for emotion classification.

The contributions of our work are the collection of emotional gait data, the proposal of motion features, and implementation of the mobile emotion recognition system, which provide a reference for future research on this topic. The rest of the paper is organized as follows. In Section II, we summarize the related work. In Section III and Section IV, we present our method of recognizing emotions and describe the system implementation. In Section V, we show the evaluation results. In Section VI, we discuss the limitations and future work. Finally, in Section VII, we conclude our work.

II. RELATED WORK

A. Major Trends in Emotion Recognition

Recognizing emotions from facial expressions is one common approach. Methods to classify facial expressions include analyzing facial features such as mouth positions [2], multiple frames of a subject's facial expressions [3], and the combination of face and speech signals [4]. Recent developments

in deep learning and the availability of large facial expression datasets have enabled even more accurate emotion recognition from faces without the need for manually defining features [1]. Alternatively, physiological signals also reveal emotions. Signals such as heartbeat and respiration rate can be extracted wirelessly to classify emotions [5]. Electroencephalography (EEG) can be fused with other features such as musical features to predict music-induced emotions [6], or eye movements to increase the accuracy in single modality models [7].

While these methods produce satisfactory performance in detecting even implicit emotions, they usually require some special equipment or close proximity of the camera to the subject with a fixed camera orientation. This leads to difficulties in carrying out emotion recognition in public areas, where only mass monitoring is possible. The possibility of detecting emotion from motion, therefore, becomes an alternative emotion recognition technique in some environments, 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]. One study on emotion recognition defines several features from gait that can be obtained from motion-captured data, such as joint angles with 3 to 6 degrees of freedom, speed, and frequency [10]. It has also been suggested by human observers that in addition to straightforward features like speed, a complex and intricate amalgam of body posture and limb movements is also important for correct emotion classification. Therefore, to effectively classify emotion from gait, the combination of various motion features must be taken into consideration.

The advancement of machine learning algorithms has opened up the possibility of emotion classification via supervised learning. In [11], 99% accuracy is achieved by using fixed-length time series data as step features with a Multilayer Perceptron to classify emotion. Their work, however, requires dedicated hardware and a precise environmental setup. Another work on this topic definies abstraction features, such as "Quantity of Motion", to quantify an individual's degree of movement in an image sequence [12]. It also proposes transformations of temporal data into consistent statistical sets of meta-features, such as mean and amplitude. However, the above work gathers specific upper-body movements motion data from a static background, which are rarely observed in real life. In our work, we collect common gait data while individuals are walking, and further define motion features for emotion recognition.

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.

TABLE I
EMOTION TO AROUSAL/VALENCE MAPPING ADOPTED FROM THE CIRCUMPLEX MODEL

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

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 sensations" [15]. An internal sensation element describes the level of pleasure and arousal of an emotion.

While discrete emotions can be simply labeled, continuous emotion models cannot be represented by 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, arousal and valence, which reflect the activeness and the intrinsic positivity perceived by the subject, respectively.

In this work, we set up the emotion classification labels listed in Table I based on the Circumplex Model. In consideration that individuals may not always display expressive emotion, we incorporate an extra "neutrality" category, indicating that no specific emotion is observed.

B. Data Collection

The dataset we collect involves 11 male university undergraduate students. Two or three videos are recorded per emotion per individual for the five emotion categories in a controlled environment. Following previous work done on motion-driven emotion recognition [10], [17], [11], participants are explicitly instructed to walk with an emotion in mind, and they are told to act out the way they think they would walk while experiencing that emotion. Each sample is then directly labeled with the instructed emotion.

All videos are recorded by a mobile phone camera from the left side of the subject, with 30 frames per second (FPS), 1080p resolution, and 16:9 aspect ratio. The first and last one-sixth portions of each video are trimmed to skip the preparation and finishing movements, since they could potentially deviate from normal walking motion.

C. Pose Estimation

Pose estimation is employed to obtain positional information of the subject and calculate motion features. We use the OpenPose model presented for single person pose estimation in each frame obtained from the videos [18]. The model outputs 18 joint positions in pixel xy-coordinate format as listed in Table II. Since the videos are recorded in a tangential angle with respect to the subject, joints features on the

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		

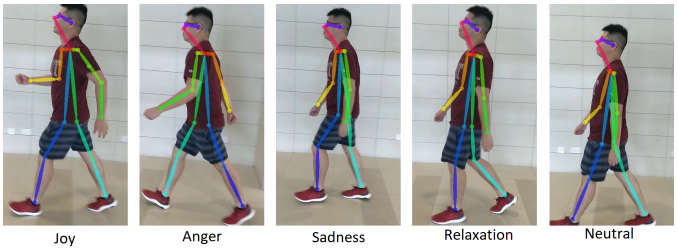

Fig. 2. Sample output of the 16 joint positions (excluding the right ear and the right eye) returned by OpenPose. Selected frames represent the most extended gestures. Note the differences in head inclination and the distances between hands.

TABLE III
EUCLIDEAN DISTANCE AND ANGULAR FEATURES

Euclidean	Description
Hand	Distance between the two hands
Hand/Hip	Distance between the left hand and left hip
Knee	Distance between the two knees
Knee/Hip	Distance between the left knee and left hip
Elbow	Distance between the two elbows
Foot	Distance between the two feet
Area	Minimum bounding box of the subject
Angle	**Description**
Arm	Angle between the two arms
Left Elbow	Left elbow flexion
Right Elbow	Right elbow flexion
Leg	Angle between the two 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 the head and torso

right, specifically the right ear (16) and the right eye (14) are disregarded when we process the data in the next step. Therefore, after carrying out pose estimation on all frames of a video, we take 16 lists of time-series data. Figure 2 shows the results obtained with 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 spike filter is implemented to detect abrupt gradient change within one to two time windows. The lost data point is then reconstructed through linear interpolation.

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 Table 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 below.

TABLE IV
META-FEATURES DERIVED FROM EUCLIDEAN AND ANGULAR MOTION 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 the peak containing the global maximum

1) Euclidean 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 features are normalized by the height of the bounding box surrounding the subject in pixels.

2) Angular Features: Angular features are defined as the angles 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.

A number of meta-parameters are used to convert the above spatial-temporal data into meta-features. These features are time-invariant yet preserve most motion information. Since human gait is periodic, we can extract statistical descriptions of the time-series data for further processing. Table IV lists the meta-feature 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 defined in the Euclidean features.

The average total speed $\boldsymbol{V_{avg}}$ is then calculated by dividing the total walking distance by the number of frames:

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

$\boldsymbol{N}$ 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:

TABLE V
TOP 15 SELECTED META-FEATURES IN TERMS OF F VALUE N 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 ($f_{crit} = 3.074, 72$ features)			
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 ($f_{crit} = 3.074, 23$ features)			
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

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

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

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

E. Model Training

We train and compare six models, including Support Vector Machine (SVM), 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 the Bootstrap Aggregating technique, which improves model stability by training multiple instances of a model and returns the majority of the output. We carry out one-way analysis of variance (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 highest 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 effects, 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 we 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.

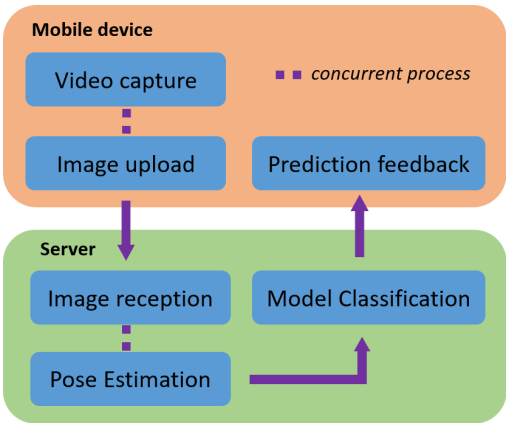

Fig. 3. System architecture diagram

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 3. 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 20 megapixel dual cameras for video recording. The processing server is a desktop computer with an 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 video frame in real time. Upon completion of emotion classification using the recognition model, the server returns the inferred emotion back to the mobile device, where the final emotion inference results will be rendered to the user.

To balance model performance and system response time, video resolution and frame rate can be adjusted. Although system efficiency will be improved if less data is sent for processing, reducing image resolution or video frame rate causes some information to be lost at the same time. The resolution of the video does not affect extracted motion features, yet frame skipping results in inconsistent time-dependent motion features, such as speed. To allieviate the issue, 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.

TABLE VI
AVERAGE ACCURACIES OF EMOTION CLASSIFIERS. "SINGLE" AND "BAGGING" REPRESENT WITHOUT AND WITH BOOTSTRAP AGGREGATING, RESPECTIVELY.

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	

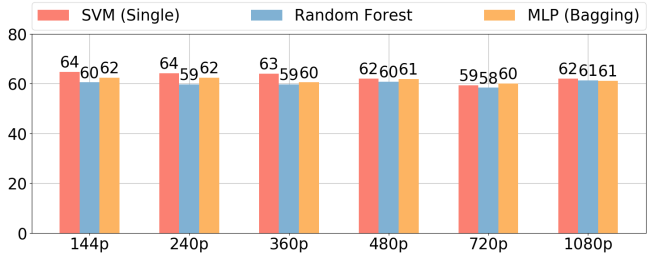

Fig. 4. Accuracies of SVM, Random Forest, and MLP with respect to test time Resolution.

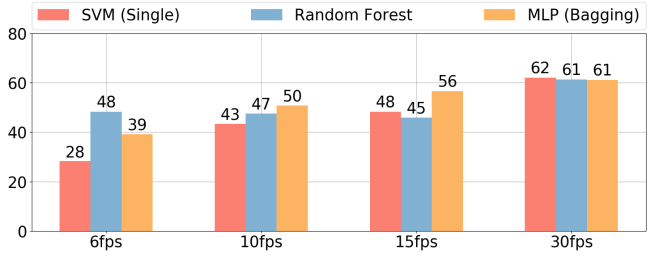

Fig. 5. Accuracies of SVM, Random Forest, and MLP with respect to test time FPS. Interpolation is used to reconstruct lost frames.

A. Model Accuracy

Table VI shows the average accuracies of the six models. 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 :

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

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

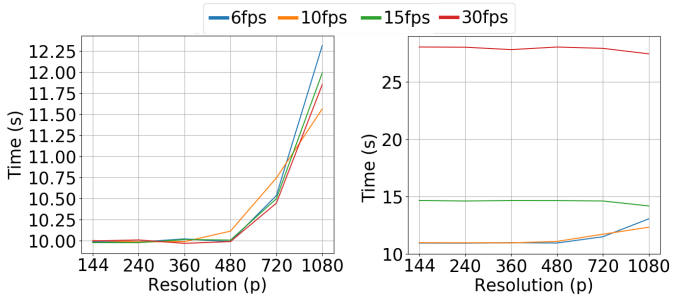

Fig. 6. Left: 10-second video transmission time, from the first frame sent to last frame received. Right: corresponding total system response time, from first frame sent to emotion prediction completed.

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 our emotion recognition model trained with 1080p and 30 FPS image sequences to test different video configurations.

Figure 4 shows the accuracies when testing with various resolutions. We can see that model accuracies remain at around 60% despite significant image quality changes. It can therefore be concluded that resolution does not significantly affect emotion inference accuracy. This may be due to the fact 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 mobile data usage, 144p videos are good enough to maintain classification accuracy.

Figure 5 shows the accuracies when testing with various frame rates. An increase in model accuracies is observed when higher FPS videos are used. SVM and MLP 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 results in a loss of accuracy, which is possibly due to motion feature information being lost. Therefore, to maximize classification accuracy, video FPS should remain at 30 when using current feature extraction methods.

C. Response Time

In order to evaluate the networking performance of the system, we measure the response time using different combinations of image resolution and FPS. The results are the averages over 10 measurements. Since our server was originally built on a lab computer with restricted connection access, to simulate a public connection, we migrate the server to an AWS EC2 t2 Micro instance and perform image transmission testing by using a smartphone with 22MB/s upload speed. We run the remaining system on the lab server by replaying videos according to the reception delay of each frame.

Figure 6 shows the response times of transmitting and processing videos recorded in 10 seconds. The chart on the left shows that video upload time begins to become prolonged at 720p and 1080p, whereas videos with lower resolutions can be uploaded almost in real time. However, the chart on the right shows that there is a significant delay in the overall system response time when 30 FPS videos are used, since pose estimation essentially processes more than twice the number of images at 30 FPS. Therefore, extra time is required to process 30 FPS videos after all the frames are received.

In summary, based on our evaluation results, if fast system response time is desired with little compromise on accuracy, we recommend that with current technology, videos be recorded with 15 FPS and 144p resolution.

VI. Discussion and Future work

Emotion recognition from gait is still a fairly unexplored area of research. As a result, there is a significant deficiency of publicly available data. The use of a small-scale dataset potentially leads to training bias of the models and lowered reliability of the test results. Moreover, a small dataset hinders the incorporation of state-of-the-art deep learning algorithms, which makes over-fitting to be highly likely. Furthermore, data collected in a controlled environment is likely to be different from real-world situations, which also leads to difficulty in developing a well-generalized, gait-based emotion recognition model. Therefore, as an initial attempt, this work is intended to encourage investment of more resources for work on this topic.

In model construction, the data processing step is crucial. 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 coordinates when joints are overlapped. The two situations cause errors and spikes in the extracted time-series data, so a special anomaly detection algorithm needs to be designed to minimize the final signal errors.

In terms of system construction, the time for video uploading and pose estimation present a challenge in minimizing the overall system response time. Unlike video streaming applications, our emotion recognition system does not allow frame skipping without suffering a noticeable loss in accuracy. Therefore, to minimize latency caused by slow frame uploading, other techniques should be considered, such as multi-path TCP.

VII. Conclusion

In this paper, we analyzed human gait from videos to recognize emotions without using complex hardware devices. We introduced a vision pipeline that extracts meta-features for emotion inference. Our performance comparison demonstrated the possibility of classifying emotions by simply recording videos of human gait. We also designed and implemented a system that runs on mobile devices and a cloud server, which enables convenient emotion recognition in public areas using mobile and wearable devices. We envision ample opportunities and applications of the system of 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 to build more advanced systems, with the possibility of incorporating state-of-the-art techniques, such as deep learning.

Acknowledgement

This research has been supported, in part, by projects 26211515 and 16214817 from the Research Grants Council of Hong Kong.

References

[1] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, "Dexpression: Deep convolutional neural network for expression recognition," *arXiv preprint arXiv:1509.05371*, 2015.

[2] L. Luoh, C.-C. Huang, and H.-Y. Liu, "Image processing based emotion recognition," in *System Science and Engineering (ICSSE), 2010 International Conference on*. IEEE, 2010, pp. 491–494.

[3] A. Azcarate, F. Hageloh, K. Van de Sande, and R. Valenti, "Automatic facial emotion recognition," *Universiteit van Amsterdam*, pp. 1–6, 2005.

[4] S. Emerich, E. Lupu, and A. Apatean, "Emotions recognition by speechand facial expressions analysis," in *Signal Processing Conference, 2009 17th European*. IEEE, 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*. ACM, 2016, pp. 95–108.

[6] N. Thammasan, K.-i. Fukui, and M. Numao, "Multimodal fusion of eeg and musical features in music-emotion recognition." in *AAAI*, 2017, pp. 4991–4992.

[7] Y. Lu, W.-L. Zheng, B. Li, and B.-L. Lu, "Combining eye movements and eeg to enhance emotion recognition." in *IJCAI*, 2015, pp. 1170–1176.

[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, 1987.

[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, 1989.

[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, 2014.

[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, 2008.

[12] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 71–82.

[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, pp. 15–15, 2009.

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