Happiness Intensity Estimation for a Group of People in Images using Convolutional Neural Networks

Guanming Lu, Wenjing Zhang
College of Telecommunications & Information Engineering
Nanjing University of Posts and Telecommunications
Nanjing, China
lugm@njupt.edu.cn

***Abstract*—In this paper, we present a group-level happiness intensity estimation approach based on convolutional neural networks. First, Multi-task Cascaded Convolutional Networks (MTCNN) are applied to detect all faces in the image. Then, the face-level happiness intensity of each detected individual face is respectively predicted by using a deep convolutional neural network with transfer learning from pre-trained VGGFace model. Finally, the weighted average of face-level happiness intensities for all detected faces in the group is used to estimate the happiness intensity of a group of people in the image. Experimental results on validation dataset of HAPPEI database demonstrate that the proposed method has promising performance, which is 0.58 in terms of Root Mean Square Error (RMSE), and outperforms the baseline approach that uses the CENsus TRansform hISTogram (CENTRIST) descriptor and Support Vector Regression with a non-linear Chi-square kernel, which gives an RMSE of 0.78.**

Keywords-facial expression recognition; group-level emotion recognition; convolutional neural networks; group happiness intensity

I. INTRODUCTION

Facial expressions are the facial changes in response to a person's internal emotional states, intentions, or social communications. Facial expression analysis has been an active research topic for behavioral scientists. Automatic facial expression recognition has attracted increasingly attention in the field of computer vision due to its importance in practice for a wide range of applications, including intelligent human-computer interaction and affect computing [1]. However, most of the existing affect analysis and recognition methods focus on analyzing the expression and emotion of an individual, and relatively little attention has been given to the estimation of the overall 'Group' emotion of a group of people in an image.

With the popularity of modern social networking sites such as YouTube and Flickr, the number of images and videos being uploaded on websites every day is growing exponentially. These images and videos from social events and gatherings usually contain multiple people. Interestingly, as people tend to present themselves in a favourable way, most of the uploaded and shared images on websites are positive, which were recorded in social activities such as birthday parties, weddings, graduation ceremonies and so on. It is of real-world application to analyse the theme emotion conveyed by an image of a group of people, such as emotion ranking, candid photo shot selection in mobile phone, image search and retrieval, event summarization, etc. In this paper, we address the happiness intensity estimation for a group of people in images.

The rest part of this paper is organized as follows. In Section II, some previous related works are introduced. In Section III, we describe in detail the proposed approach which includes face detection, face-level happiness intensity estimation, and group-level happiness intensity prediction based on the weighted average of all face-level happiness intensities. The experimental results are presented in Section IV. Finally, concluding remarks are provided in Section V.

II. RELATED WORKS

A. Dhall et al. [2, 3] addressed the problem of expression analysis for a group of people, and proposed a bottom-up approach based on analyzing the contribution of each person of the group towards the overall group expression and a framework for computing the group expression, i.e. Group Expression Model (GEM) based on topic modeling and manually defined attributes such as individual person's happiness intensity, occlusion intensity, relative distance and relative size of a face. The hypothesis is that there are certain attributes, which affect the perception of happiness of a group of people in an image. The manually defined attributes are applied as weights to the Group Expression Model. They also created the HAPPEI (HAPpy PEople Images) database for inferring the group-level happiness intensity of a group of people in an image.

Since 2013, as part of the ACM International Conference on Multimodal Interaction (ICMI), the EmotiW (Emotion recognition in the Wild) challenge series has been run as a grand challenge. The aim is to provide a platform for researchers to benchmark the performance of their methods on 'in the wild' data. The fourth EmotiW challenge in 2016 contains two sub-challenges. One is VReco (Video based emotion Recognition) sub-challenge, and the other is GReco (Group-level emotion Recognition) sub-challenge [4]. The task of GReco sub-challenge is to infer the overall happiness intensity of a group of people in a given image.

978-1-7281-3584-7/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on June 23,2020 at 13:39:55 UTC from IEEE Xplore. Restrictions apply.

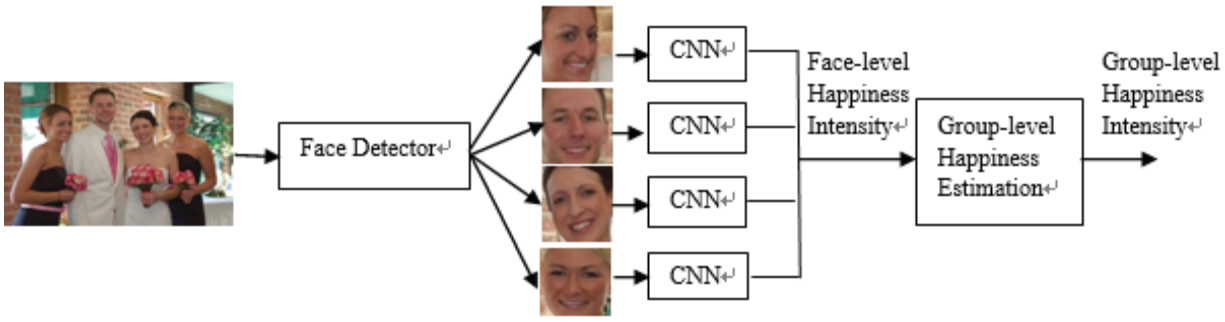

Figure 1. The pipeline of the proposed approach

III. PROPOSED APPROACH

The pipeline of the proposed approach is shown in Fig. 1. There are three main stages in the proposed approach. Firstly, the faces of the individuals in the image are detected by using Multi-task Cascaded Convolutional Networks (MTCNN) face detector. Then, we use a convolutional neural network to extract discriminative facial features and to estimate the face-level happiness intensity of each detected individual face in the image respectively. Finally, the overall group-level happiness intensity is estimated by the weighted average of face-level happiness intensities for all detected faces in the group. For face detection and face-level happiness intensity estimation we employ Convolutional Neural Networks (CNNs), which have shown good capability to learn discriminative facial features without the need for hand-crafting in advance.

A. Face Detection based on MTCNN

Recently, Convolutional Neural Networks (CNNs) have been widely used and achieved the state-of-the-art performance in many pattern recognition tasks such as face recognition. Some of the CNN-based face detectors greatly outperform the previous Viola-Jones cascade classifiers in accuracy. The face detector based on cascade CNNs [5] can be considered as a representative of the combination of traditional boosted cascade framework and CNNs. However, this face detection algorithm requires bounding box calibration with extra computational expense and ignores the inherent correlation between facial landmarks localization and bounding box regression. Zhang et al. [6] proposed a new face detection algorithm named as Multi-task Cascaded Convolutional Networks (MTCNN), which integrates both face detection and face alignment tasks using unified cascaded CNNs by multi-task learning. In this paper, we applied MTCNN for face detection.

B. Face-level Happiness Intensity Estimation using Deep CNN

Due to the current lack of sufficient annotated training samples, deep CNNs have not been thoroughly evaluated for the estimation of happiness intensity. It is well known that direct training of deep CNNs with limited training samples may even result in poor performance due to over-fitting. To address this problem, several researchers have applied transfer learning strategies to find solutions [7,8]. Compared with training a deep CNN model on the small-scale training sample dataset, the transfer learning method can avoid over-fitting and improve the performance of deep CNN. Transfer learning has provided a reasonable compromise that enhances the accuracy of CNN-based classification systems without requiring very large datasets.

In this section, the VGGFace [9] model which was pre-trained for face recognition using the large VGGFace dataset (2.6M images of 2622 identities), is directly exploited to construct the backbone architecture of target CNN model for the estimation of face-level happiness intensity. Pre-training the model on large-scale database helps in achieving generalized weight configuration, suitable for extracting general image features. The weights of pre-trained VGGFace model are used as the initial weights of target CNN model for further training. Namely, we transfer the knowledge from the pre-trained VGGFace model to target CNN model and fine tune it on the datasets of interest.

C. Group-level Happiness Intensity Estimation

Having obtained face-level happiness intensity for each individual face in an image, we now combine them in order to get the overall group-level happiness intensity of a group of people. It has been shown that the perception of the group-level happiness intensity is affected by both top-down and bottom-up components [3]. The top-down global context attributes include factors such as the relative position and face size of the person in the group. The bottom-up local context attributes are relevant to each individual person's occlusion intensity and happiness intensity.

In this paper, we investigate group-level happiness intensity predominantly from a bottom-up perspective. Given an image I containing n faces, a simplest approach is to use the mean of face-level happiness intensities for all faces in the image as a prediction of Group-level Happiness Intensity (GHI) [3], which can be formulated as

$$I_{\mathrm{G}} = \mathrm{round}\left[\frac{\sum_{i=1}^{n} I_{\mathrm{f}}^{i}}{n}\right] \tag{1}$$

where, n denotes the total number of detected faces in the image, I_{f}^{i} is the face-level happiness intensity for i-th face, and $\mathrm{round}[\bullet]$ is a function that rounds its argument to the nearest integer. In this simple formulation, the social context information based on the global structure on where people are located in a given scene are being ignored.

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on June 23,2020 at 13:39:55 UTC from IEEE Xplore. Restrictions apply.

Generally speaking, in the image of group photos, people standing closer to the camera have relatively larger faces. Here, we assume that the happiness intensity of the persons closer to the camera has a greater impact on the overall group-level happiness intensity as compared to the persons standing in the back. In this paper, we takes into consideration the relative size of a face S_{f}^{i} as weight factor to the process of determining the group-level happiness intensity in the image. This weight is applied to the I_{f}^{i} of each face in the group, and the weighted GHI can be defined as

$$I_{\mathrm{G}}^{\mathrm{W}} = \mathrm{round}\left[\frac{\sum_{i=1}^{n} S_{\mathrm{f}}^{i} I_{\mathrm{f}}^{i}}{\sum_{i=1}^{n} S_{\mathrm{f}}^{i}}\right] \tag{2}$$

where, S_{f}^{i} is the size of the bounding box of the *i*-th detected face (in pixels).

This formulation takes into consideration the global structure of a group of people. The contribution of individual person's happiness intensity towards the overall happiness intensity of the group is weighted. It is evident that larger faces get a higher weight.

IV. Experiments and Evaluations

A. HAPPEI Databases

HAPPEI database [3] was collected from Flickr which is a photo sharing platform by using keyword search. Keywords related to social events, such as 'convocation', 'marriage', 'party', etc, were used to fetch images from difference social activities. This dataset is divided into three data partitions i.e. training dataset (1500 images), validation dataset (1138 images) and test dataset (496 images). Some sample images from HAPPEI database are shown in Fig 2.

These images were further labelled by human annotators. Besides the group-level happiness intensity were labelled, 8500 faces were also manually annotated for each individual face-level happiness intensity, the location of the face, and occlusion intensity for each person in images. The group-level or face-level happiness intensity in the database is denoted with a label between 0 to 5, which corresponds to six classes of happiness intensity: Neutral, Small Smile, Large Smile, Small Laugh, Large Laugh and Thrilled, where 0 corresponds to 'Neutral' and 5 being the highest score corresponding to 'Thrilled'. Some face images with face level happiness intensity label are shown in Fig 3. These annotations serve as ground truth, which contesting models attempt to infer.

Figure 2. Some sample images from HAPPEI database.

Figure 3. Some face images with face-level happiness intensity label from HAPPEI database.

B. Experimental Setup

Firstly, we detect the face images in each group image using MTCNN-based face detector. The MTCNN model is trained on the WIDER FACE [10] training dataset. Positive training samples are taken from the ground-truth bounding boxes and other bounding boxes having a sufficiently large overlap with the ground-truth face. In this paper, we take an Intersection over Union (IoU) larger than 0.7 as the criterion for selecting positive training samples. Negative training samples are the square image patches taken from the background of those images, or those square patches having small overlap with any ground-truth faces (IoU < 0.2). Part faces are the square image patches whose IoU is between 0.4 and 0.7 to a ground-truth face. Positive and negative training samples are used for face classification tasks, positive training samples and part faces are used for bounding box regression.

Then, we adopt the CNN model based on pre-trained VGGFace to estimate the happiness intensity of the individual corresponding to detected face. In order to compatible with the input data size of VGGFace model, each of the detected face images is resized to 224×224 pixels with 3 channels. For the specific task of face-level happiness intensity estimation, the last fully-connected layer and classification layer of the pre-trained VGGFace model are removed, and a fully-connected layer with 1024 neurons and softmax layer for 6-class happiness intensity classification task are added in the target CNN model. While fine-tuning the target CNN model, we could retain the most pre-trained weights for the initialization of target CNN model and then fine-tune the target CNN model with training samples from HAPPEI database.

After detecting all the faces from the training dataset and validation dataset in HAPPEI database, we only keep the annotated faces with a label. This resulted in approximately 8500 annotated face images. These annotated face images are future divided into two subsets: training dataset (7500 face images), validation dataset (1000 face images). In order to obtain sufficient amount of training samples for fine-tuning the target CNN model, we applied data augmentation technique to expand the training dataset by rotation, color jittering, and horizontal flipping. The final training dataset consisted of 30000 annotated face images.

In fine-tuning phase, we initialize the target CNN model with pre-trained weights except that the weights of task-oriented fully connected layers and softmax layer are initialized by using the Xavier method. The weights of all layers in target CNN model are then fine-tuned with the mini-batch gradient decent method by Adam optimizer. The batch size is set to 32, and the initial learning rate is set to 1e-4. The target CNN model is trained 30 epochs to get optimized results.

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on June 23,2020 at 13:39:55 UTC from IEEE Xplore. Restrictions apply.

Finally, the overall group-level happiness intensity is obtained from the weighted average of face-level happiness intensities for all detected faces in the image. The performance is evaluated by the Root Mean Square Error (RMSE), which is defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{k}\left(I_{\mathrm{G}}^{\mathrm{W}} - I_{label}\right)^2}{k}} \quad (3)$$

where, k is number of test image samples, $I_{\mathrm{G}}^{\mathrm{W}}$ denotes the estimated group-level happiness intensity and I_{label} stands for the ground truth intensity label of test image sample.

C. Experimental Results

We conducted experiments on validation dataset of HAPPEI database to evaluate the proposed group-level happiness intensity estimation algorithm using the weighted average of face-level happiness intensities for all detected faces in the group, compared with the mean of face-level happiness intensities for all faces. On the other hand, we also investigated the performance of proposed approach based on transfer learning from pre-trained VGGFace model, and compared with other well-known pre-trained deep CNN models such as VGG16, InceptionV3, ResNet50. Table 1 shows the comparison of RMSE for group-level happiness intensity estimation with different methods. It is noteworthy that the smaller RMSE values the better estimation performance.

Experimental results demonstrate that, compared to other pre-trained deep CNN models such as VGG16, InceptionV3, ResNet50, adopting pre-trained VGGFace model as the backbone architecture of target CNN model for the estimation of face-level happiness intensity will improve group-level happiness intensity estimation accuracy significantly. Experiment results also show that group-level happiness intensity estimation based on weighted average of face-level happiness intensities have better performances over simple mean Group Expression Model (GEM) proposed in [3]. These results are as we expected. The proposed method has promising performance, which gives an RMSE of 0.58, and outperforms the baseline approach [4] that uses the CENsus TRansform hISTogram (CENTRIST) descriptor and Support Vector Regression with a non-linear Chi-square kernel, which gives an RMSE of 0.78.

TABLE I. PERFORMANCE COMPARISON OF DIFFERENT METHODS IN TERMS OF RMSE.

	Pre-trained deep CNN Model			
	VGGFace	***VGG16***	***Inception-V3***	***ResNet50***
Mean GEM[3]	0.61	0.68	0.65	0.66
Weighted average (ours)	0.58	0.64	0.63	0.63

V. CONCLUSIONS

In this paper, a group-level happiness intensity estimation approach based on the weighted average of face-level happiness intensities is proposed. The relative size of face is considered to be a weight factor for determining the group-level happiness intensity in the image. Experiment results show that assigning relative weights to happiness intensities helps in better estimation of the group-level happiness intensity. In the future, we will further investigate how to make full use of other social context factors such as age, gender to improve estimation performance. Further extensions of the work can benefit from robust face detection and alignment and the use of faster and more discriminative expression classifiers.

ACKNOWLEDGMENT

This work was supported by the Key Research and Development Program of Jiangsu Province (Grant No. BE2016775) and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. KYCX17_0787, No. KYCX19_0899, No. KYCX19_0954).

REFERENCES

[1] Y. Miao, H. Dong, J. M. Al Jaam, A. El Saddik, "A deep learning system for recognizing facial expression in real-time," ACM Trans. Multimedia Comput. Commun. Appl., vol. 15, no. 2, pp. 1–20, 2019.

[2] A. Dhall, J. Joshi, I. Radwan, and R. Goecke, "Finding happiest moments in a social context," Lect. Notes Comput. Sci., vol. 7725 LNCS, pp. 613–626, 2013.

[3] A. Dhall, R. Goecke, and T. Gedeon, "Automatic group happiness intensity analysis," IEEE Trans. Affective Comput. , vol. 6, no. 1, pp. 13–26, 2015.

[4] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon, "Emotiw 2016: Video and group-level emotion recognition challenge," Proc. ACM Int. Conf. Multimodal Interact., Tokyo, pp. 427–432, 2016.

[5] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit, Boston, pp. 5325–5334, 2015

[6] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE Signal Process. Lett. Vol. 23, no. 10, pp. 1499–1503, 2016.

[7] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order," Pattern Recogn., vol. 61, pp. 610–628, 2017.

[8] H. Kaya, F. Gürpınar, and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," Image Vision Comput, vol. 65, pp. 66–75, 2017.

[9] O. M. Parkhi, A. Vedaldi, A. Zisserman, "Deep face recognition," British Machine Vision Conference, 2015.

[10] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit, Las Vegas, pp. 5525–5533, 2016.

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on June 23,2020 at 13:39:55 UTC from IEEE Xplore. Restrictions apply.