MultiEmotiCon: Multimodal Emotion Recognition in the Context

Haoyu Wang
u6783702

Zhifeng Wang
u6068466

Shan Wang
u7094434

Xu Han
u6821692

Abstract

Emotion plays an important role in humans' daily life. Throughout different emotions, humans build different relationships with others and take various actions to respond to their emotions. Currently, scientists propose accurate facial expressions models to predict emotions. However, we can also refer to surroundings to improve the accuracy of emotion recognition. In this paper, we propose an end to end network to extract the features of face, body, background, depth and semantic segmentation, and use these features to predict emotions. Finally, we test our method on EMOTIC and the result has 1% improvement when compared with the Kosti's result.

key words: emotion recognition, end-to-end network, face, body, background, semantic segmentation

1 Introduction

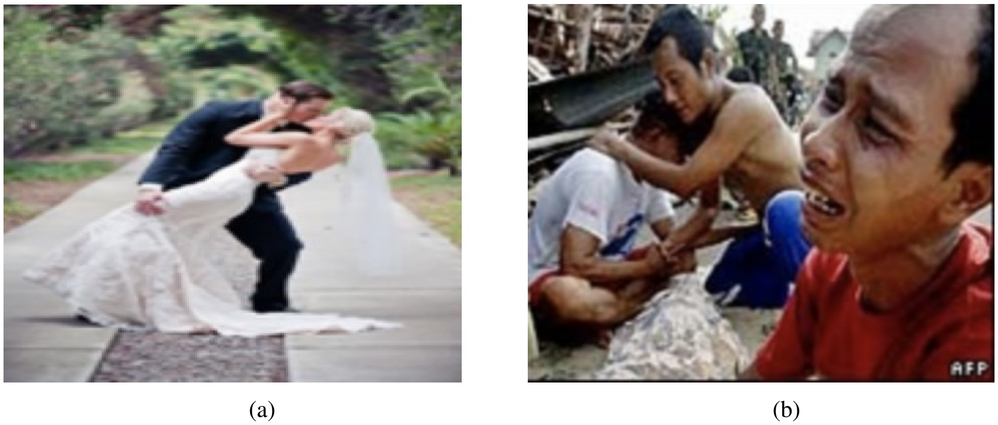
In real human society, emotion plays an important role. In daily communication, people express their emotions and get the corresponding responses from others. Through the kind of emotional interaction, various interpersonal relationships are created between people, such as friends, enemies, and partners. Generally, humans usually take different behaviors to others by their perceived emotions.

Hundreds of years ago, psychologists began to study human emotion recognition. In 1917, Ekman and Feiesen [1] systematically established a facial emotion image dataset with the 6 basic human emotions, which includes fear, joy, surprise, anger, disgust and sadness. Thus, early emotion recognition mostly refers to facial expression recognition [2–5], which use traditional computer vision algorithms to detect face and categorize emotions. For example, Viola et al. [5] uses SVM (Support Vector Machine) with Radial Basis Function kernels to classified facial expressions. Subsequently, [6–8] are not limited on facial expression recognition, and begin to use multiple modalities, such as gestures and body postures, to improve the accuracy of emotion recognition.

With the invention of running neural networks on GPU in 2012, deep learning, especially convolution neural networks, shows the great potential in computer vision. For instance, Hasani et al. [9] use an enhanced deep 3d Convolution neural network to extract facial landmarks, which can improve the facial expression recognition of subtle changes. And FLM-CNN [10] can define the difference between facial shapes by adopting human vision inspired pooling and multi scale encoding strategy.

However, the above researches focus on the human features, such as facial expression and body postures to analyze human emotions. Some psychology research points out that the impact of context on human emotions is significant [10, 12]. In the Figure 1, we can clearly recognize people's emotions.

In particular, though the Figure 1.a cannot capture the frontal faces, the emotion of this couple will be happiness and affection. From Figure 1.b, this is a scene of ruins, we can feel that the child in the red frame should be sad and fearful. Therefore, adding context information can be seen as an effective way to increase the accuracy of emotion recognition. In mostly recent years, some papers start to



(a)

(b)

Figure 1: Human emotion under different contexts

combine the context with human features to estimate emotions. More details will be introduced in section “Related work”.

In this report, we integrated the idea from [13–15] and based on Kosti et al. [14]’s open code sources to develop our works from the

Three aspects:

1. Decomposing the whole human features as the two parts, including face and body.
2. Transferring the input from RGB image to RGB-D image. The D means the depth map, which makes input three-dimensional. It is beneficial to weight pixels in the context.
3. Using semantic segmentation to understand the meaning of objects in the background context.

For the structure of this report, the Section.2. will focus on related work. Then, the Section.3. will explain some details about our work. Next, we will introduce the experimental dataset (EMOTIC) in the Section.4. And the Section.5. shows the result of our work and other emotion recognition methods. Finally, the Section.6. is the conclusion.

2 Related Work

In this section, we will introduce some related works.

Group-level Methods: In recent years, some researchers consider group-level methods for emotion recognition. For example, Kosti et al. [14] not only extract the target human’s features but only use the whole image as an input to analyse the target’s emotion, which is the global features of images to support emotion recognition. Lee et al. [13] do similar works as Kosti et al. [14]. One difference between [13] and [14], Lee et al. only extract facial features when Kosti et al. make body and face as a whole to extract human’s features. Furthermore, Lee et al. hide the facial regions when extracting context information. Compared to [14], Lee’s context regions are more discriminative that are more effective to improve the accuracy of emotion recognition. However, sometimes the camera cannot always capture the facial area of a person. [13] only use facial expressions to recognize emotions that might cause the large error in the real environment. Therefore, we decided to use both face and body but the different streams in our networks. Except face and body stream, this report will improve the context stream. In Mittal et al. [15]’s paper, the authors refine the context features. The Context 2 represents the background context, which uses semantic understanding to obtain insights of the target human’s surrounding environments and activities. Due to one specific agent’s emotion will be influenced by its surrounding agents, the Context 3 use depth map and GCN-based to obtain the social-dynamic inter-agent interactions context. However, we think that the network architecture is too complex, and the understanding of Context 2 has a little bit of overlap with Context3. Thus, we only use one stream for context features in our network. Based on Mittal et al. 's [15] idea, the depth map and semantic understanding are useful to extract the most important context features. Zhang et al. [21] states adding depth information into an RGB image can improve the performance of semantic segmentation. Thus, we will propose an RGB-D image as the input to extract context features. Additionally, semantic segmentation can be seen as another method to extract context features.

Places365-CNNS: In Places365 [23], there are two versions of the dataset. Places365-standard provides approximately 1.8 million images, which belong to 365 scene categories. Furthermore, Places365-Challenge further adds 8 million extra images. Under the large data, Places365 releases various pre-trained CNNs models, such as Alex-places365, VGG16-places365 and ResNet-places365 etc. Based on places365's contribution, we can put different models into our network architecture to find the best one for us.

In the next section, we will give more details about our approaches.

3 MultiEmotiCon: Our Approach

3.1 Network architecture

We propose an end-to-end network to estimate emotion recognition from discrete categories and continuous dimensions, which is shown in Figure 2 **MultiEmotiCon**.

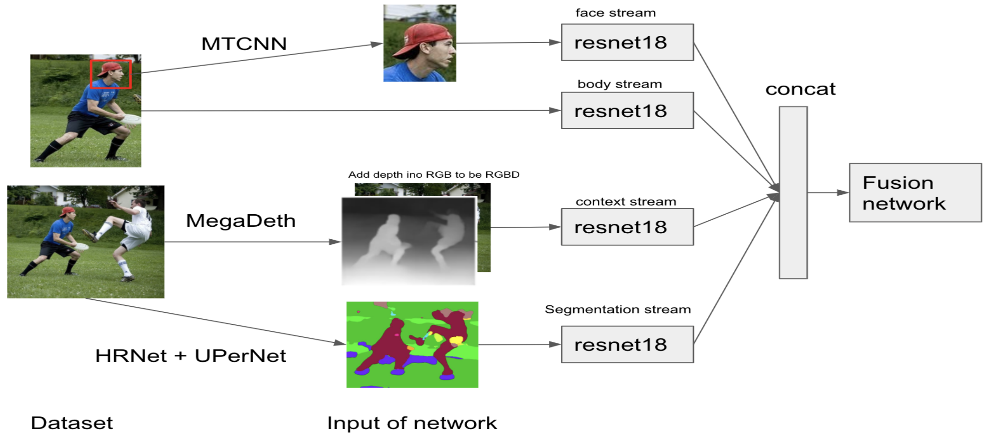


Figure 2: **MultiEmotiCon**: We use four modalities, face, body, context and segmentation. Using resnet18 to extract features from these four modalities for emotion recognition.

The MultiEmotiCon network includes five main modules, such as face, body and context and semantic streams and one fusion. Compared to the [14]'s network, both face and semantic segmentation are new streams. In the context module, we modify the input from RGB to RGB-D image. Body stream is the same as [14]'s. all streams use resnet18[23] to extract features. The features extraction by the four streams are combined with a fusion network.

The fusion network module includes two fully connected layers. The first fully connected layer is to reduce the feature's dimensions to 256. The second fully connected layers are to learn the independent features of each emotion category. The second fully connected layer is divided into two branches, one is with 26-unit features which is used to estimate 26 emotion categories, the other is with 3 unit features which is used to estimate 3 continuous emotion categories.

Next, we will introduce more details about each stream.

3.2 Face stream

For the face part, we need to select a suitable facial recognition method. In recent years, the facial detection technology has been achieving the accuracy of human-level performance. MTCNN[16] is still widely used for face detection and alignment in popular face recognition methods, and we use it to extract the facial regions in our report. According to Zhang et al.[16], the MTCNN uses a cascaded three stage-CNN for face detection and alignment, which contain less filters than the previous face detection CNN[17]. This means their carefully designed CNN is lightweight networks, which will reduce the network run time. Although some recent face detection methods have more accurate results on FDDB, such as DSFD for 0.991 [18] and Face R-Face R-FCN [20], MTCNN still includes alignment stage with not bad face detection accuracy (0.9504) [16]. In the EMOTIC datasets, because some people's faces are occluded and can not be detected by MTCNN.

$$F(x) = \begin{cases} 1, x \in \phi \\ 0, x \notin \phi \end{cases}$$

If MTCNN can detect faces, the input will be face image, otherwise the input will be the body image.

These undetected images will maintain original body image and ensure the input body features information will not be lost. Then, taking face images as input and combining them with other global features modules to infer emotion will improve the accuracy of emotion recognition.

3.3 Depth for context stream

Depth information contains the distance between the surface of objects and a viewpoint, many recent papers widely used depth map in simultaneous localization and mapping (SLAM), Visual Odometry, 3D reconstruction etc [26–28]. We consider using MegaDepth [29] to do our depth estimation, which widely collects images from the internet. This means our EMOTIC dataset is not the specific dataset for evaluating depth prediction, models trained on MegaDepth have strong generalization.

Firstly, we do not consider integrating MegaDepth into our framework. One of the reasons is that MegaDepth contains a large network that takes up huge memory. If we integrate MegaDepth into our framework, we have to adjust the parameter size of other streams. Another reason is that parameters for depth extraction do not need to be updated during the training process. So, we just use MegaDepth to prepare input data for the context stream.

Depth information in the context stream is more important than the depth value of face and body area. For context stream, depth can provide attention information such as pay more attention to context with the same depth of agent.

Our previous idea is to design an attention map generate algorithm. Input of this algorithm is depth maps. However, the idea heavily relies on the coordinates of input images, the features from CNN are position irrelevant. Another idea is to feed depth maps to CNN and fully connected layers to get features from depth maps and concatenate them with other features.

Because we noticed that the framework is huge and contains lots of parameters, we decided to add depth information effectively without adding too many parameters. This means we add the depth map as the fourth channel to the original RGB image. We expect our network will find features in depth. And these features are not independent, they are relative to the RGB channels. Because the resnet 18 of context is pre-trained, it is better to keep those trained parameters. So the method is getting parameters (size:64*3*7*7) from the first convolution layer and concatenate it with new 64*1*7*7 parameters.The RGB-D image can be shown in Figure 3.

3.4 Semantic Segmentation Stream

In the semantic segmentation part, we try to use the state-of-the-art image semantic segmentation model to generate its segmentation image (I_{seg}) for each original image. After referring to the open source project CSAILVision [22] as our semantic segmentation guidance,we decided to use the encoder-decoder structure in our work. The encoder we chose HRNet[23], a model that handles

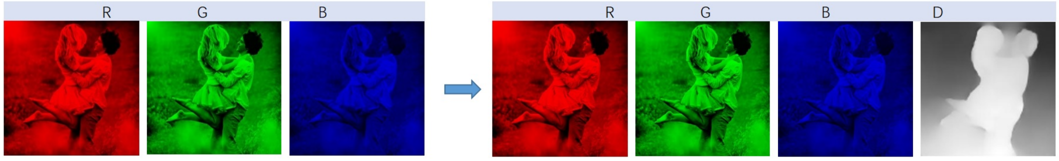


Figure 3: RGB to RGB-D

high resolution images well and achieves the state-of-the-art performance in pixel labeling tasks. By looking at the data set we used, most of the images are of high resolution, so this model is appropriate. In the decoder part, we compared the UPerNet [24] and PSPNet[30]. Since UPerNet does not use dilated convolution, it has relatively small expenses in training time and memory while it performs better than PSPNet. Finally, we select UPerNet, which uses the structure of Feature Pyramid Network (FPN) and Pyramid Pooling Module (PPM). The whole training process happens on the ADE20K dataset, which is the biggest open source semantic segmentation dataset [25]. To save time, we implement the pretrained models directly by [22]. Finally, through experiments, the addition of semantic segmentation improves the performance of emotion recognition in the scene.

4 Results

In this section, we introduce the result of our approaches and compare with baseline[14]'s performance on EMOTIC dataset.

4.1 EMOTIC dataset

The EMOTIC dataset is the specific dataset of images with people in real environments. At the same time, each image includes annotated target people's emotions. The emotion annotation contains the two dimensions, such as discrete categories and continuous dimensions. The combination between the two dimensions is good to describe what emotions of the specific person is and how level of the person's emotion is.

Discrete categories: compared to [1]'s dataset with 6 basic human emotions, the EMOTIC dataset develops more 20 emotional categories, such as Peace, Affection, Sympathy and Fatigue etc. According to the authors of EMOTIC[30], they collected affective vocabularies by books on psychology and dictionaries. Furthermore, clustering similar vocabularies into 26 categories to create the discrete dimension.

Continuous dimensions: Mehrabian[31] proposes a VDA model to describe human emotion from a comprehensive perspective. V represents Valence, which measures how pleasant or positive one emotion is. For example, happiness has a high-level valence. Otherwise, fear's valence is low-level. And A means Arousal that measures the person's agitation level. For instance, the anger is high-level when sadness has a low-level arousal. Finally, D represents Dominance to measure the control level of the condition by the person, which ranges from submission (e.g. fear) to control in feeling (e.g. Happiness).

The combination between the two dimensions is good to describe what emotions of the specific person is and how level of the person's emotion is.

The dataset utilizes Amazon Mechanical Turk to annotate the emotion of each image as the ground truth. In this dataset, it saves 23,571 images and 34,320 annotated people.

4.2 Environment

We both run our experiment and baseline program on the server (mlcv2.anu.edu.au) that is provided by our lecture. Due to the small differences between Resnet18, Resnet50 and Resnet101 on baseline's result and we have limited time to complete our experiments, not only results of baseline but also our approaches based on Resnet18.

Because of the small differences between Resnet18, Resnet50 and Resnet101 on baseline's result and limited course time, we only use Resnet18 in our experiment. Due to comparing our results and the baseline. We implement the original method with Pytorch. In our experiments, we set epochs as 15, 25, 50. We found that the 15 will get the best result. If the epoch is larger than 15, the validation errors will not decrease, and the model will show an over-fitting problem. The Figure 4 shows baseline's Training and validation errors with epochs.

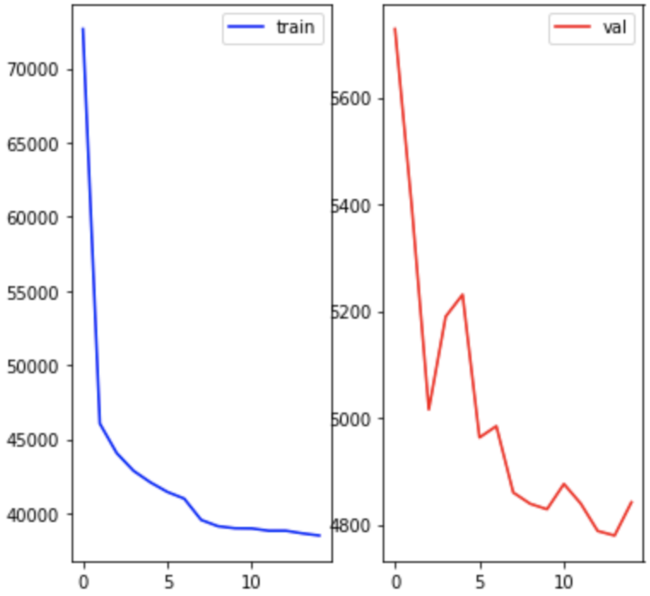


Figure 4: Baseline's Training and validation errors with epochs

4.3 Experiment Result

Due to the network architecture being end to end, it is simple to add one single method as a new stream. Thus, we can compare the effect of each of our approaches with baseline. Moreover, we will test the effect of fusion approaches.

Table 1 shows the average precision obtained on the test set per category. Next, the Table 2 demonstrates the average absolute Error obtained on the test set per continuous dimensions. We will discuss the results in the next section.

To sum up, Table 1 and Table 2 show the results for using different optimization m

5 Discussion

In this section, we will compare the results of our approach with baseline.

Discrete categories: Table 1 illustrates the average precision obtained per each category on the test set. For average precision, all single methods improve the accuracy on discrete dimensions.

Specifically, the semantic segmentation stream improves the baseline the most. According to statistics, 18 of the 26 categories by the segmentation stream have improved and the improvement of 8 categories

Emotion Categories	CNN	Ours(Depth)	Ours (Face)	Ours(Seg)	All
1. Affection	0.1781	0.1706	0.1802	0.1704	**0.1917**
2. Anger	0.0415	0.0489	0.0396	**0.0538**	0.0369
3. Annoyance	0.7573	0.0816	0.7337	**0.8815**	0.0732
4. Anticipation	0.5410	0.5388	0.5500	0.5397	**0.5507**
5. Aversion	0.0425	0.0398	0.0444	**0.0597**	0.0390
6. Confidence	0.7330	0.7330	0.7331	0.7289	**0.7367**
7. Disapproval	0.0655	**0.0691**	0.0646	0.0682	0.0649
8. Disconnection	0.2048	0.2008	0.2103	**0.2111**	0.1959
9. Disquietment	0.1291	0.1251	**0.1296**	0.1292	0.1192
10. Doubt/Confusion	0.1650	0.1643	**0.1711**	0.1658	0.1504
11. Embarrassment	0.0253	**0.0274**	0.0264	0.0221	0.0241
12. Engagement	0.8198	0.8150	0.8246	0.8250	**0.8317**
13. Esteem	0.1427	0.1344	**0.1523**	0.1461	0.1259
14. Excitement	0.6448	0.6418	0.6462	**0.6550**	0.6435
15. Fatigue	0.0637	0.0637	0.0640	0.0630	**0.0652**
16. Fear	0.0437	0.0472	0.0453	**0.0498**	0.0341
17. Happiness	0.5499	0.5470	**0.5794**	0.5668	0.5641
18. Pain	0.0278	0.0326	0.0298	**0.0425**	0.0268
19. Peace	0.1656	**0.1799**	0.1609	0.1560	0.1666
20. Pleasure	0.3083	0.0324	0.3276	0.3236	**0.3385**
21. Sadness	0.0973	0.0952	0.0997	**0.1060**	0.0811
22. Sensitivity	0.0369	0.0328	**0.0423**	0.0344	0.0400
23. sufferring	0.0638	0.0638	0.0672	**0.0872**	0.0566
24. Surprise	0.0663	**0.0708**	0.0644	0.0661	0.0664
25. Sympathy	0.1120	0.1153	**0.1233**	0.1097	0.1042
26. Yearning	0.0710	0.0722	0.0704	0.0753	**0.0771**
mean	0.2083	0.2091	0.2125	**0.2131**	0.2079

Table 1: Average Precision obtained on the test set per category

Continuous Dimensions	CNN	Ours(Depth)	Ours (Face)	Ours(Seg)	All
Valence	0.9042	0.9209	**0.8863**	0.9180	0.8938
Arousal	1.0666	1.0750	1.0718	1.0540	**1.0417**
Dominance	0.9531	0.9623	0.9480	**0.9378**	0.9715
mean	0.9746	0.9860	0.9799	0.9813	**0.9690**

Table 2: Average Absolute Error obtained on the test set per continuous dimension

are bigger than 1%. By checking the categories which have the negative optimization, we find that many images which are labelled these negative optimization categories are personal photos, which is difficult to estimate the target's emotion by segmentation. we guess the reason is that semantic segmentation makes the same instance's pixels have same intensities, it may cause that the most pixels in personal photos have same values and it is negative to extract core features. We will further analyse it in the future.

The second improvement method is Face Stream. Particularly, 20 of the 26 categories by this stream have improved, which means that the face stream is wider to estimate emotion. However, another 6 categories' labels show the negative optimization. We think the potential reason is that the proportion of pictures without face region in the 6 categories is larger than other categories' images. For Figure 5, 'Happiness' category has the highest face detection ratio which is around 83%, so the accuracy of 'Happiness' is improved due to the highest Non-occluded face. But in 'Annoyance' category, the face detection ratio is high enough, but because of the face alignment problem, the result is not good. Additionally, because we use body region to replace face region when an image cannot detect face features, it may lead to some noises in the face stream. The third method: Using RGB-D images as

the input in context stream to improve the accuracy of discrete categories is limited. We think there are two reasons. One is that depth as an input channel combined with color intensity information

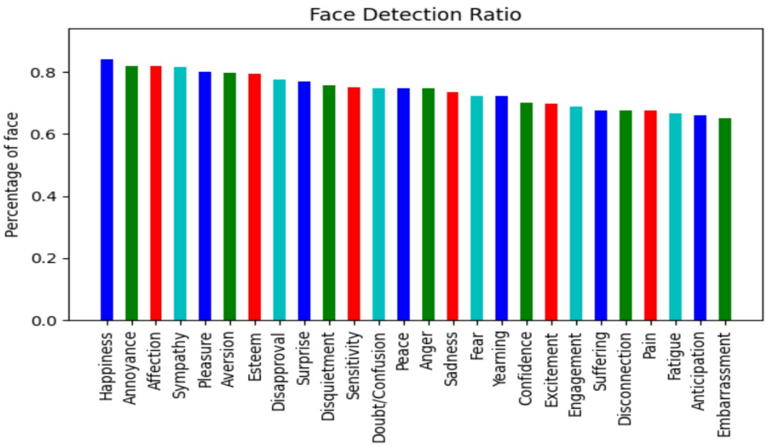


Figure 5: Face Detection Ratio

strict the feature condition. For example, a wedding dress might be a feature, and wedding dresses in different images are similar in color but have different depth distribution. The other reason is that the composite network is hard to train. It is easy to train the last fully connected layer with fixed pre-trained networks. But when we send gradient descent back to backbone networks it converges slowly.

A frustrating result is that though our three approaches improve the average precision, our fusion result is the negative optimization. Due to the limitation of research time, we have not been able to find the reason why fusion leads to the negative optimization and we will analyze it in the future.

Continuous dimension: Table 2 demonstrates the average absolute error obtained on the test set. Although the fusion result shows that it decreases a little absolute error when compared with baseline, all individual methods give the negative impact on decreasing absolute error.

We think that it is difficult to improve the accuracy on continuous dimensions with the single face information. Thus, it is normal that the face stream is useless on continuous dimensions. For semantic segmentation stream, since one potential reason is that segmentation loses the image details that can be estimated on the continuous dimensions, it causes the negative optimization. For depth, due to the RGB-D image can transfer the image from 2D to 3D, we expect that 3D image can reduce the absolute error but the truth is that the depth leads to negative effect. One possible reason is that the monocular depth estimation method is not mature, which cannot provide the high accurate depth map.

In fact, different approaches proposed in baseline's paper [14] are useless for optimizing continuous dimension.

6 Conclusion

In this paper, we address the problem of emotion recognition. Based on [14], we add two new streams, such as face and semantic segmentation, and change the input of the context stream from RGB image to RGB-D image. By EMOTIC dataset, we estimate our approaches from discrete categories and continuous dimension. For discrete categories, each individual approach shows a small improvement. For continuous dimension, our approaches have limited impacts.

To obtain significant improvement on discrete categories, we consider optimizing our current approaches in the future, such as using the newest state-of-the-art methods on depth estimation and face detection. Additionally, finding an effective approach to reduce the absolute error on continuous dimension is a good idea.

References

[1] Ekman, Paul, and Wallace V. Friesen. "Constants across cultures in the face and emotion." Journal of personality and social psychology 17.2 (1971): 124.

[2] Sown M. A preliminary note on pattern recognition of facial emotional expression[C]//The 4th International Joint Conferences on Pattern Recognition, 1978. 1978.

[3] M. S. Bartlett, G. Littlewort, I. Fasel and J. R. Movellan, "real time face detection and facial expression recognition: development and applications to human computer interaction," Computer Vision and Pattern Recognition Workshop, vol. 5, July 2003.

[4] Belhumeur, P.N., J.P. Hespanha and D.J. Kriegman, 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell., 19: 711-720.

[5] Viola, Paul, and Michael Jones. "Rapid object detection using a boosted cascade of simple features." Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001. Vol. 1. IEEE, 2001.

[6] Gunes, Hatice, and Massimo Piccardi. "Bi-modal emotion recognition from expressive face and body gestures." Journal of Network and Computer Applications 30.4 (2007): 1334-1345.

[7] Burgoon, Judee K., et al. "Augmenting human identification of emotional states in video." Intelligence Analysis Conference, McClean, VA. 2005.

[8] Kessous, Loic, Ginevra Castellano, and George Caridakis. "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis." Journal on Multimodal User Interfaces 3.1-2 (2010): 33-48.

[9] Hasani, Behzad, and Mohammad H. Mahoor. "Facial expression recognition using enhanced deep 3D convolutional neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017.

[10] Chen, Zhixing, et al. "Fast and light manifold cnn based 3D facial expression recognition across pose variations." Proceedings of the 26th ACM international conference on Multimedia. 2018.

[11] Barrett, Lisa Feldman, Batja Mesquita, and Maria Gendron. "Context in emotion perception." Current Directions in Psychological Science 20.5 (2011): 286-290.

[12] Aminoff, Elissa M., Kestutis Kveraga, and Moshe Bar. "The role of the parahippocampal cortex in cognition." Trends in cognitive sciences 17.8 (2013): 379-390.

[13] Lee, Jiyoung, et al. "Context-aware emotion recognition networks." Proceedings of the IEEE International Conference on Computer Vision. 2019.

[14] Kosti, Ronak, et al. "Emotion recognition in context." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[15] Mittal, Trisha, et al. "EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege's Principle." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[16] Zhang, Kaipeng, et al. "Joint face detection and alignment using multitask cascaded convolutional networks." IEEE Signal Processing Letters 23.10 (2016): 1499-1503.

[17] Li, Haoxiang, et al. "A convolutional neural network cascade for face detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[18] Li, Jian, et al. "DSFD: dual shot face detector." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[19] Abhishek. "Context Based Emotion Recognition using Emotic Dataset". Retrieved from https://github.com/Tandon-A/emotic. 2019.

[20] Wang, Yitong, et al. "Detecting faces using region-based fully convolutional networks." arXiv preprint arXiv:1709.05256 (2017).

[21] Zhang, Shaopeng, et al. "Joining geometric and RGB features for RGB-D semantic segmentation." 2019 International Conference on Image and Video Processing, and Artificial Intelligence. Vol. 11321. International Society for Optics and Photonics, 2019.

[22] Semantic Understanding of Scenes through ADE20K Dataset. B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso and A. Torralba. International Journal on Computer Vision (IJCV), 2018. Retrieved from https://github.com/CSAILVision/semantic-segmentation-pytorch. 2019.

[23] Sun, Ke, et al. "High-resolution representations for labeling pixels and regions." arXiv preprint arXiv:1904.04514 (2019).

[24] Xiao, Tete, et al. "Unified perceptual parsing for scene understanding." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[25] Scene Parsing through ADE20K Dataset. B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso and A. Torralba. Computer Vision and Pattern Recognition (CVPR), 2017.

[26] Weder, Silvan, et al. "RoutedFusion: Learning Real-time Depth Map Fusion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[27] Tateno, Keisuke, et al. "Cnn-slam: Real-time dense monocular slam with learned depth prediction." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[28] Yang, Nan, et al. "D3VO: Deep Depth, Deep Pose and Deep Uncertainty for Monocular Visual Odometry." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[29] Li, Zhengqi, and Noah Snavely. "Megadepth: Learning single-view depth prediction from internet photos." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[30] Zhao, Hengshuang, et al. "Pyramid scene parsing network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[31] Mehrabian, Albert, and James A. Russell. An approach to environmental psychology. the MIT Press, 1974.