Emotion Understanding in Videos Through Body, Context, and Visual-Semantic Embedding Loss

Panagiotis Paraskevas Filntisis[1], Niki Efthymiou[1], Gerasimos Potamianos[2], and Petros Maragos[1]

[1] School of E.C.E., NTUA, Greece
[2] E.C.E. Department, UTH, Greece
{filby,nefthymiou}@central.ntua.gr, gpotam@ieee.org, maragos@cs.ntua.gr

Abstract. We present our winning submission to the First International Workshop on Bodily Expressed Emotion Understanding (BEEU) challenge. Based on recent literature on the effect of context/environment on emotion, as well as visual representations with semantic meaning using word embeddings, we extend the framework of Temporal Segment Network to accommodate these. Our method is verified on the validation set of the Body Language Dataset (BoLD) and achieves 0.26235 Emotion Recognition Score on the test set, surpassing the previous best result of 0.2530.

Keywords: emotion, body, context, visual-semantic, BEEU challenge

1 Introduction

Automatic human affect recognition from visual cues is an important area of computer vision that has attracted increased interest over the last two decades, due to its many applications. Indeed, social robotics [2], psychiatric care [13], and edutainment [10] are all areas that can benefit from automatic recognition of emotion.

Most past approaches to the problem have focused on facial expressions in order to determine the emotional state of the person of interest [7,18,22]. This is reasonable due to the fact that facial expressions have been studied extensively in the psychology and emotion literature [8]. For example, the Facial Action Coding System (FACS) [9] identifies the units of facial movements, based on facial muscle groups. Combinations of the so-called action units (AUs) have also been linked with emotional states with extensions of the basic FACS such as EMFACS (Emotion FACS) [11]. On the other hand, there is no similar established coding system for body expressions, although some have been proposed [4].

Compared to facial expression based approaches, recent works have sought alternative modalities and streams of information to detect emotion; one is bodily expressions since many have highlighted the fact that the emotional state is conveyed through bodily expressions as well, and in certain emotions it is the main

Proc. 16th European Computer Vision Conference Workshops (ECCVW) - Workshop on Bodily Expressed Emotion Understanding, Aug. 2020

modality [5,15,26], or can be used to correctly disambiguate the corresponding facial expression [1]. Simultaneously, it is important to note that in cases and applications where the emotion needs to be identified, the human body is more frequently available than the face since the face can be occluded, hidden, or far in the distance. Another auxiliary stream of information besides the face and the body that can help in identifying emotions is the context and the surrounding environment of the person [16,21]. It is apparent that both the place, as well as objects and other humans can influence a person's emotions.

We should also note that inherently emotion recognition is a multi-label problem - the subject might be feeling two or more emotions. This is true, especially when considering an extended set of emotions, as in [19]. The emotions in extended sets do not have the same "semantic" distance between them. For example, anger is more close to annoyance than to happiness. Considering that previous works have showed the superiority of methods that attempt to learn a joint embedding space that contains both word embeddings and visual representations [6,12,24], we believe that trying to attach a semantic meaning to the extracted visual feature is a natural way forward.

In this paper, based on the above, we describe the method of our team in the First International Workshop on Bodily Expressed Emotion Understanding (BEEU) challenge. Our method combines Temporal Segment Networks (TSNs) [27] focusing on the body, using the context in each video as an additional stream, and also uses an extra visual-semantic embedding loss, based on GloVE (Global Vectors) [23] word embedding representations. Our experiments in the validation set verify the better performance of our method compared to the traditional TSNs, while our emotion recognition score on the test set was 0.26235.

2 Related Work

While most past approaches in visual detection of affect have been focused on facial expressions [5], recent approaches have started taking into account the body language [15] of the person in question, as well as its surrounding context/environment.

In [14], Gunes and Piccardi introduced a bimodal architecture that takes into account both upper body and facial expressions, in order to detect affect in videos. In [3], Dael et al. analyzed and classified body emotional expressions using a body action and posture coding system which was proposed in [4]. The 3D pose of children was also utilized in [20] by Marinoui et al. to detect emotions in continuous dimensions, while in [10], 2D pose was used and fused with facial expressions for child emotion recognition. Luo et al. [19] introduced a large scale video dataset (BoLD) annotated with categorical and continuous emotions, which is the one used in the BEEU challenge.

Regarding the context modality, Kosti et al. [16] introduced a large scale dataset for emotion recognition (EMOTIC) in different contexts (e.g., other people, places, or objects) and a convolutional neural network (CNN) based two-stream architecture that focused on the body and context of the subjects.

The CAER video dataset for context-based emotion recognition was presented in [17], along with a two-stream architecture which employed adaptive-fusion to merge the two steams. In [21], Mittal et al. designed a deep architecture with several branches, focusing on different interpretations of the surrounding context (e.g., environment and interaction context) to significantly increase resulting predictions in the EMOTIC dataset.

Finally, some recent works have also focused on extracting visual representations from images that present the semantic relations found in embeddings built from words. The DeViSE embedding model [12] extracted semantically-meaningful visual representations by introducing a similarity loss between the feature vector extracted from a CNN and the word embedding from a skip-gram text model. Using a similar method, Wei et al. [28] built joint text and visual embeddings as emotion representation from web images, and in [29], Ye and Li built semantic embeddings for a multi-label classification problem.

3 Dataset

The dataset used in the challenge is the BoLD (Body Language Dataset) corpus [19] consisting of 9,876 video clips of humans expressing emotion, primarily through body movements. Each clip can contain multiple characters, yielding a total of 13,239 annotations, split into a training, validation, and test set. The dataset has been annotated by crowdsourcing employing two widely accepted categorizations of emotion. The first one is the categorical annotation with a total of 26 labels first used in [16], by collecting and processing an extensive affective vocabulary. The second annotation regards the continuous emotional dimensions of the VAD (Valence - Arousal - Dominance) Emotional State Model [25]. The methods in the challenge are evaluated using the following Emotion Recognition Score (ERS):

$$\text{ERS} = \frac{1}{2}\left(\text{m}R^2 + \frac{1}{2}(\text{mAP} + \text{mRA})\right) \tag{1}$$

where mR^2 is the mean coefficient of determination (R^2) score for the three dimensional emotions (VAD), and mAP and mRA is the mean Average Precision and the mean area under receiver operating characteristic curve (ROC AUC) of the multilabel categorical predictions.

4 Model Architecture

Our model is based on the TSN architecture [27], which has been widely used in action recognition and can be seen in Fig. 1. During training, K different segments are selected from the input video, and then N consecutive frames are selected from each segment. This is done to deal with the fact that consecutive frames have usually redundant information. Traditionally, two different modalities are used, one is the spatial (RGB) modality and the second one is the optical

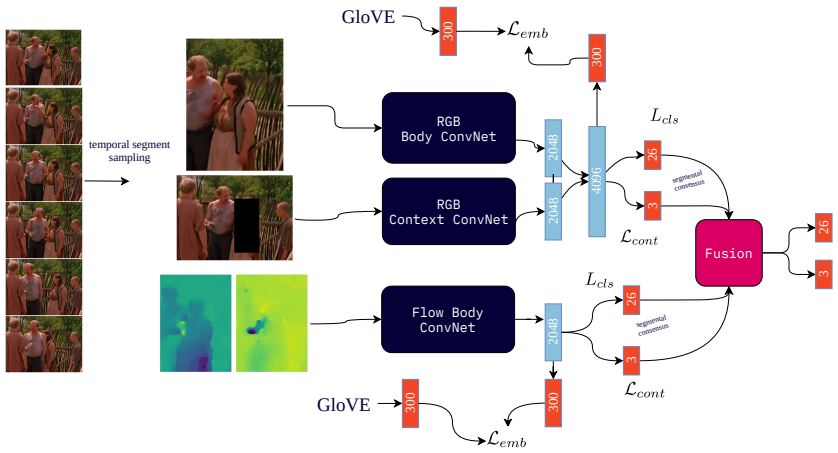

Fig. 1: TSN with two RGB spatial streams (body and context) and one optical flow stream. The final results are obtained using average score fusion.

flow. TSNs have already been shown to achieve good results for the BoLD dataset in its introductory paper [19].

In our approach, we modify the original version of TSNs mainly in two directions:

Context: We introduce one additional stream based on the context-environment surrounding the annotated human. For the RGB modality, we input the context in the network in the same way as in [21], by masking out the instance body (we set all pixels to 0). We call this stream RGB-c, and the body streams RGB-b and Flow-b. During training, the RGB-b and RGB-c streams are combined at the feature level (RGB-bc) and are trained jointly while the Flow-b TSN is trained independently.

Embedding Loss: Our second extension is the introduction of an embedding loss on the feature vector extracted by the Convolutional Neural Network (ConvNet). This is done to exploit the fact that some emotions are closer semantically to others. This is also revealed by examining the correlation matrix of the dataset labels in [19], where some labels occur more frequently in combination with others (e.g. Happiness and Pleasure, Annoyance and Anger, etc.). Due to this result, we try to attach a semantic meaning to the feature vector extracted by the backbone image network.

To implement this, we first obtain for each one of the 26 categorical labels of BoLD their 300-dimensional GloVE word embedding [23]. A PCA-projection of the 26 embeddings is shown in Fig. 2, where it is apparent that the distances between embeddings are indicative of their "semantic" distance. We then use

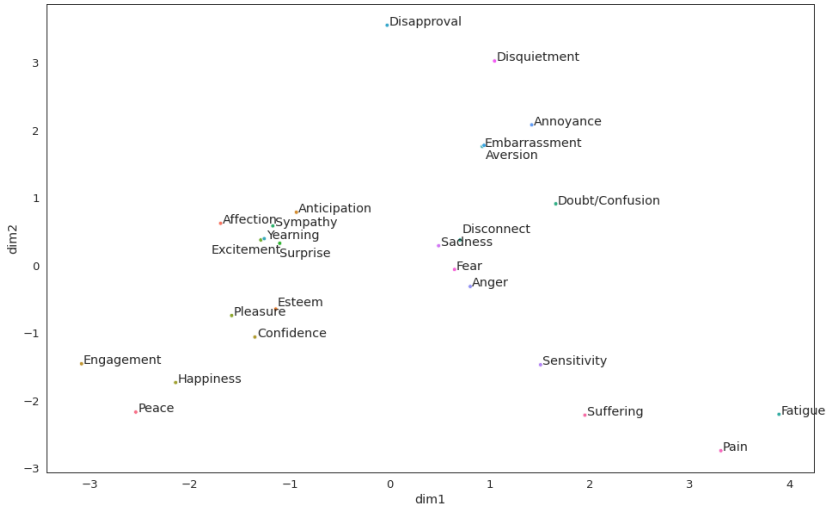

Fig. 2: PCA projection of the categorical emotions GloVE word embeddings.

a fully connected layer to map the feature extracted from the image to a 300-dimensional space and introduce the following mean-squared based loss:

$$\mathcal{L}_{emb} = ||\boldsymbol{W} f_v(\boldsymbol{x}) - \frac{1}{|K|} \sum_{y \in K} f_w(\boldsymbol{y})||_2 \tag{2}$$

where $f_v(\boldsymbol{x})$ is the feature vector extracted by applying the convNet on the image $\boldsymbol{x}$, $\boldsymbol{W}$ is a linear transformation from the space of the feature vector to the word embedding space, $f_w(\boldsymbol{y})$ is the word embedding of the label y, and K is the set of all positive labels for the image $\boldsymbol{x}$. That is, we try to reduce the Euclidean distance between the projected image feature and the arithmetic mean of the GloVE embeddings of the positive labels for image/video.

Predictions: Finally, after extracting for each sampled image its feature vector, we use two fully connected layers, one to classify to the 26 different categorical labels, and one to regress over the 3 different categorical emotions. The two TSNs are trained using the following loss:

$$\mathcal{L} = \mathcal{L}_{cls_1} + \mathcal{L}_{cls_2} + \mathcal{L}_{cont} + \mathcal{L}_{emb} \tag{3}$$

Specifically, since the dataset does not provide explicitly the multilabel targets, but the crowdsourced scores between 0 and 1, we include two different losses for the classification part: $\mathcal{L}_{cls_1}$ that is the binary cross-entropy between the predicted scores and the multilabel target (obtained after thresholding the multilabel scores at 0.5) and $\mathcal{L}_{cls_2}$ that is the mean squared error between the predicted scores and the multilabel scores. We empirically found that the inclusion of $\mathcal{L}_{cls_2}$ slightly boosted performance. For the regression part, $\mathcal{L}_{cont}$ is the

	Model	mAP	mRA	mR^2	ERS
without $\mathcal{L}_{emb}$	RGB-b	0.1567	0.6140	0.0538	0.21955
	Flow-b	0.1444	0.5914	0.0507	0.2093
	RGB-b + Flow-b	0.1623	0.6307	0.078	0.2375
with $\mathcal{L}_{emb}$	RGB-b	0.1564	0.6143	0.0546	0.21997
	Flow-b	0.1465	0.5947	0.0579	0.2142
	RGB-b + Flow-b	**0.1637**	**0.6327**	**0.0874**	**0.2428**

Table 1: Ablation experiment by training with and without $\mathcal{L}_{emb}$.

mean-squared error between the regressed values and the continuous emotions. Finally $\mathcal{L}_{emb}$ is as in (2).

5 Experimental Results

We train each TSN for 50 epochs using Stochastic Gradient Descent (SGD), with initial learning rate 10^{-3} which drops by a factor of 10 at 20 epochs[3]. The backbone networks used is a residual network (ResNet) with 101 layers for the body convNets and a ResNet with 50 layers for the context convNet. We use the default hyperparameters of TSNs: 3 segments, 1 frame from each segment for the RGB streams, and 5 frames from each segment for the optical flow stream. The consensus used for segment fusion is averaging. For each network, we select the epoch with the best validation ERS. We have also found experimentally that the partialBN (Batch Normalization) technique used in [27] gives a nontrivial boost to the performance of the network.

First, in Table 1 we present two ablation experiments regarding the addition of $\mathcal{L}_{emb}$. We can see that adding the embedding loss increases slightly the performance in the RGB-b stream, and gives a boost to the performance of the Flow-b stream.

Then, in Table 2 we present our experimental results on the validation set of BoLD including the RGB context stream. From the results we can see that including the context along with the body in the RGB modality boosts the validation ERS of the architecture. We also experimented with including the context in the Flow network, but this resulted in worse performance. Our final submission for the test set was the model with the best validation score (0.2439 employing RGB-bc + Flow-b), using 25 segments instead of 3. The results of the different metrics on the test set can also be seen in Table 2, while the final ERS is 0.26235, improving upon the previous best result of 0.2530[19].

[3] PyTorch code available at https://github.com/filby89/NTUA-BEEU-eccv2020

set	Model	mAP	mRA	mR^2	ERS
valid	RGB-c	0.1395	0.5760	0.0365	0.1971
	RGB-bc	0.1566	0.6055	0.0675	0.2243
	RGB-bc + Flow-b	0.1656	0.6266	0.0917	**0.2439**
test	RGB-bc + Flow-b	0.1796	0.6416	0.1141	**0.26235**

Table 2: Results on the validation and test set of BoLD including the RGB context stream and $\mathcal{L}_{emb}$.

6 Conclusions

In this paper we presented our method submitted at the BEEU challenge, winning first place. Our method extended the TSN framework to include a visual-semantic embedding loss, by utilizing GloVE word embeddings, and also included an additional context stream for the RGB modality. We verified the superiority of our extensions compared to the baseline on the validation set of the challenge, and submitted the best system which achieved 0.26235 Emotion Recognition Score on the BoLD test set, surpassing the previous best result of 0.2530.

Acknowledgments

This research is carried out / funded in the context of the project "Intelligent Child-Robot Interaction System for designing and implementing edutainment scenarios with emphasis on visual information" (MIS 5049533) under the call for proposals "Researchers' support with an emphasis on young researchers- 2nd Cycle". The project is co-financed by Greece and the European Union (European Social Fund- ESF) by the Operational Programme Human Resources Development, Education and Lifelong Learning 2014-2020.

References

1. Aviezer, H., Trope, Y., Todorov, A.: Body cues, not facial expressions, discriminate between intense positive and negative emotions. Science **338**(6111), 1225–1229 (2012)
2. Cavallo, F., Semeraro, F., Fiorini, L., Magyar, G., Sinčák, P., Dario, P.: Emotion modelling for social robotics applications: a review. Journal of Bionic Engineering **15**(2), 185–203 (2018)
3. Dael, N., Mortillaro, M., Scherer, K.R.: Emotion expression in body action and posture. Emotion **12**(5), 1085 (2012)
4. Dael, N., Mortillaro, M., Scherer, K.R.: The body action and posture coding system (BAP): Development and reliability. J. Nonverbal Behavior **36**(2), 97–121 (2012)

5. De Gelder, B.: Why bodies? twelve reasons for including bodily expressions in affective neuroscience. Philosophical Transactions of the Royal Society of London B: Biological Sciences **364**(1535), 3475–3484 (2009)
6. Dong, J., Li, X., Snoek, C.G.: Word2visualvec: Image and video to sentence matching by visual feature prediction. arXiv preprint arXiv:1604.06838 (2016)
7. Du, S., Tao, Y., Martinez, A.M.: Compound facial expressions of emotion. Proceedings of the National Academy of Sciences **111**(15), E1454–E1462 (2014)
8. Ekman, P., Keltner, D.: Universal facial expressions of emotion. Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture pp. 27–46 (1997)
9. Ekman, R.: What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA (1997)
10. Filntisis, P.P., Efthymiou, N., Koutras, P., Potamianos, G., Maragos, P.: Fusing body posture with facial expressions for joint recognition of affect in child–robot interaction. IEEE Robotics and Automation Letters **4**(4), 4011–4018 (2019)
11. Friesen, W.V., Ekman, P., et al.: Emfacs-7: Emotional facial action coding system. Unpublished manuscript, University of California at San Francisco **2**(36), 1 (1983)
12. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: Advances in neural information processing systems. pp. 2121–2129 (2013)
13. Gaudelus, B., Virgile, J., Geliot, S., Franck, N., Dupuis, M., Hochard, C., Josserand, A., Koubichkine, A., Lambert, T., Perez, M., et al.: Improving facial emotion recognition in schizophrenia: a controlled study comparing specific and attentional focused cognitive remediation. Frontiers in psychiatry **7**, 105 (2016)
14. Gunes, H., Piccardi, M.: A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In: Proc. ICPR. vol. 1, pp. 1148–1153 (2006)
15. Kleinsmith, A., Bianchi-Berthouze, N.: Affective body expression perception and recognition: A survey. IEEE Trans. on Affective Computing **4**(1), 15–33 (2013)
16. Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: Emotion recognition in context. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1960–1968 (2017)
17. Lee, J., Kim, S., Kim, S., Park, J., Sohn, K.: Context-aware emotion recognition networks. In: Proc. IEEE International Conference on Computer Vision. pp. 10143–10152 (2019)
18. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: Proc. IEEE computer society conference on computer vision and pattern recognition-workshops. pp. 94–101 (2010)
19. Luo, Y., Ye, J., Adams, R.B., Li, J., Newman, M.G., Wang, J.Z.: ARBEE: Towards automated recognition of bodily expression of emotion in the wild. International Journal of Computer Vision **128**(1), 1–25 (2020)
20. Marinoiu, E., Zanfir, M., Olaru, V., Sminchisescu, C.: 3D human sensing, action and emotion recognition in robot assisted therapy of children with autism. In: Proc. CVPR. pp. 2158–2167 (2018)
21. Mittal, T., Guhan, P., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D.: Emoticon: Context-aware multimodal emotion recognition using frege's principle. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14234–14243 (2020)

22. Mollahosseini, A., Hasani, B., Mahoor, M.H.: AffectNet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)
23. Pennington, J., Socher, R., Manning, C.D.: GloVE: Global vectors for word representation. In: Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014)
24. Ren, Z., Jin, H., Lin, Z., Fang, C., Yuille, A.L.: Multiple instance visual-semantic embedding. In: Proc. BMVC (2017)
25. Russell, J.A., Mehrabian, A.: Evidence for a three-factor theory of emotions. Journal of Research in Personality **11**(3), 273–294 (1977)
26. Tracy, J.L., Robins, R.W.: Show your pride: Evidence for a discrete emotion expression. Psychological Science **15**(3), 194–197 (2004)
27. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European Conference on Computer Vision. pp. 20–36. Springer (2016)
28. Wei, Z., Zhang, J., Lin, Z., Lee, J.Y., Balasubramanian, N., Hoai, M., Samaras, D.: Learning visual emotion representations from web data. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13106–13115 (2020)
29. Yeh, M.C., Li, Y.N.: Multilabel deep visual-semantic embedding. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(6), 1530–1536 (2020)