Taking the Hidden Route: Deep Mapping of Affect via 3D Neural Networks

Claudio Ceruti[1], Vittorio Cuculo[1,2], Alessandro D'Amelio[1], Giuliano Grossi[1], and Raffaella Lanzarotti[1](✉)

[1] PHuSe Lab - Dipartimento di Informatica, Università degli Studi di Milano,
Via Comelico 39/41, Milano, Italy
{claudio.ceruti,vittorio.cuculo}@unimi.it,
alessandro.damelio@studenti.unimi.it,
{grossi,lanzarotti}@di.unimi.it
[2] Dipartimento di Matematica, Università degli Studi di Milano,
Via Cesare Saldini 50, Milano, Italy

Abstract. In this note we address the problem of providing a fast, automatic, and coarse processing of the early mapping from emotional facial expression stimuli to the basic continuous dimensions of the core affect representation of emotions, namely valence and arousal. Taking stock of results in affective neuroscience, such mapping is assumed to be the earliest stage of a complex unfolding of processes that eventually entail detailed perception and emotional reaction involving the proper body.

Thus, differently from the vast majority of approaches in the field of affective facial expression processing, we assume and design such a feedforward mechanism as a preliminary step to provide a suitable prior to the subsequent core affect dynamics, in which recognition is actually grounded. To this end we conceive and exploit a 3D spatiotemporal deep network as a suitable architecture to instantiate such early component, and experiments on the MAHNOB dataset prove the rationality of this approach.

Keywords: Deep neural networks · Continuous affect space

1 Introduction

Facial expression (FE) is the most effective modality for emotion display [8] and humans have developed specific skills to recognize even subtle expression changes [2]. FEs are generated by rapid (between 250 ms to 5 s) contraction of facial muscles. The accurate measure of FE could be delegated to the fEMG [7,19] or accomplished by computer vision techniques. In this vein, a plethora of approaches have been proposed ranging from local to holistic approaches, adopting deformation or motion-based models, being image or model-based (for an exhaustive discussion of FE methods see the survey [27]).

However, as a general comment on the large body of work that has been done in the field of affective FE detection/analysis, the vast majority of

S. Battiato et al. (Eds.): ICIAP 2017 International Workshops, LNCS 10590, pp. 189–196, 2017.
https://doi.org/10.1007/978-3-319-70742-6_18

approaches mostly rely on the classic computer vision and pattern recognition "pipeline" [27], where visual feature extraction/reduction is followed by classification (discrete emotion recognition) or regression (continuous affect detection), e.g. [20].

In this note, we take a different perspective. As motivated in Sect. 2, our main concern is in providing a suitable account of the earliest stage of FE processing, where, upon stimulus onset, a fast, automatic trigger is fed into the continuous core affect state-space of valence and arousal. To such end, in this work we adopt a deep network architecture (Sect. 3). Recently, deep networks have proven their effectiveness in solving a variety of vision tasks. Also the FE task has been faced with deep networks [4,20]. Here, different from those works, we aim at achieving at the output of this automatic feed-forward step, a suitable prior in probabilistic terms, for the latent manifold that functionally models core-affect state-space [31].

This, in turn will initiate further processing that, eventually, will lead to emotion recognition/attribution. Remarkably, it has been shown that at the neurobiological level, processes that follow this "hidden", sub-cortical step, involve both visuomotor and visceromotor pathways that are likely to be used in simulation-based mechanisms for affective expression recognition [2,17,32].

2 Background and Motivations

FE analysis goes back to 1872, when Darwin demonstrated among other things the universality of facial and body expressions [10]. About one century later Ekman and Friesen [14] postulated six primary emotions that possess each a distinctive content together with a unique and universal facial expression. These prototypic emotional displays are also referred to as *basic emotions*. In more general terms, this is the bulk of the discrete theory of emotions. Pioneering work on automatic FE analysis goes back to Mase and Pentland [23]. Since then, in computer vision and markedly in the more recent affective computing field [24] a large body of work has been done in the framework of the discrete approach [11]. This success can be easily understood if one thinks of basic emotions as categorical constructs: then the attribution of emotion simply boils down to a classification problem over a set of suitable features (e.g., computational counterparts of Ekmans's Action Units -AUs [13]- in the case of FEs [27]).

There are however other competing theories to the discrete theory of emotions. The continuous, dimensional view parsimoniously proposes the two broad dimensions of valence (pleasure/displeasure) and arousal (sleepy/activated) of affect [25], as the core (core affect) of emotion representation and processing. This describes a kind of "kernel" neurophysiological state as the basis for grounding emotion episodes. Such view is supported by the fact that many kinds of emotion data can be mapped well into such a continuous two-dimensional space. Interestingly enough, the dimensional approach has received much more attention in affect evaluation via physiological, voice or music signals rather than FEs research [11], though, more recently, continuous representations are gaining currency [18]. Componential models of emotion [15,28] argue that the rich emotions that people experience unfold through a complex set of evaluations and coping mechanisms. Such "appraisal" theories invoke a larger vocabulary of features from which a correspondingly larger set of emotions can be constructed. In some respect, the computational models derived from appraisal theories have raised interest over the years in the classic Artificial intelligence (AI) community [11]. However the use of how to exploit this theoretical approach for automatic measurement of affect is an open research question since requiring, as discussed [18], complex, multicomponential and sophisticated measurements of change.

This long dispute by competing psychological theories might be eventually reconciled, as conjectured by Dubois and Adolphs: "*One could imagine constructing a more complex framework consisting of an underlying dimensionality of valence and arousal, a more fine-grained classification into six or so basic emotion categories, and a very fine-grained and more flexible attribution based on appraisal features*" [12].

Under such circumstances, it is best to take into account recent findings in affective neuroscience that might pave the way to a thorough and principled synthesis. Coming back to FE analysis, the unfolding of emotion attribution at the neurobiological level can be summarised as follows [1,2]. Upon the onset of an emotionally meaningful stimulus, observer's response undergoes the following stages: (1) fast early perceptual processing of highly salient stimuli (120 ms); (2) detailed perception and emotional reaction involving the body (170 ms); (3) retrieval of conceptual knowledge about the emotion signaled by the expresser's face (>300 ms).

At the core of all such stages lies the activity of the amygdala and the pre-frontal cortex. It has been argued [26] that functional interactions between the amygdala and pre-frontal cortex form a potential neural substrate for the encoding of the psychological dimensions of valence and arousal, thus of the core affect.

Most interesting for the work presented here is the first stage. Initial perception of the face modulates activity in subcortical structures as well as in early visual cortices. The subcortical structures, the superior colliculus and the pulvinar thalamus, are likely to be specialized for very fast, automatic, and coarse processing of the stimulus. In particular, coarse processing of the visual motion generated by dynamic facial expressions might be relevant. Crucially, information from the pulvinar feeds into early processing within amygdala. As to the cortical structures, it would include V1, V2, and other early visual cortices, via input from the lateral geniculate thalamus.

Early visual processing may be relatively specialized to extract information about highly salient stimuli and it may rely in large part on information based on specific features that are detected. These processes are likely to be fairly automatic and obligatory. Subsequently, it would be also supported by more anterior visual regions dedicated to face processing (e.g., superior temporal gyrus for what concerns mouth movement, eye movements, and changes in facial expression).

The amygdala participates in the recognition of emotional signals via at least the subcortical route (superior colliculus, pulvinar thalamus), and the cortical route via the visual neocortex. This is consistent with LeDoux dual-route proposal [9,16,21].

It has been shown that even subliminally presented facial expressions of fear activates the amygdala and such activation appears to depend on relatively passive or implicit processing of the emotion. Indeed, it is this fast automatic perceptual mapping from the early stimulus onset to basic "core affect components" that we are addressing in this note.

3 Architecture

To account for the early trigger to affective nuclei we propose to adopt an architecture inspired to the 3D convolutional network C3D presented in [30], allowing to learn spatiotemporal features. Since dealing with a more specific task than the usual video scene classification and using a small dataset, we opted for a shallower architecture than C3D to prevent overfitting. We choose three layers each composed by one 3D convolutional block, followed by *ReLU* nonlinearities, max pooling and layer normalization [3]. The last layer is a global average pooling block [22] followed by a linear fully connected layer and a *tanh* activation that outputs the estimates for valence and arousal. The network is fed by inputs composed by 16 consecutive frames, thus capturing spatiotemporal features. Groundtruth values of valence and arousal for these 16-frame volumes are obtained averaging the corresponding single-frame valence and arousal labels.

Table 1. Size of the 3D convolutional blocks for each layer

Layer	Convolutional filter size
1	$5 \times 5 \times 5 \times 3 \times 64$
2	$5 \times 5 \times 5 \times 64 \times 128$
3	$5 \times 5 \times 5 \times 128 \times 128$

4 Experimental Results

The training and test of the spatiotemporal convolutional network we proposed, has been setup considering the 23 most expressive videos of the MAHNOB dataset [5], and referring to the continuous valence/arousal labeling we produced exploiting a novel web-based annotation tool named DANTE (Dimensional ANnotation Tool for Emotions) [6]. 13 videos have been used for training and the remaining for test.

In Fig. 1 we report an example of both the manual annotations and the automatic regression values of valence on one of the tested video (vid9 in MAHNOB). We observe that the automatic values respect the trend of the ground truth, while the range of variability is slightly reduced.

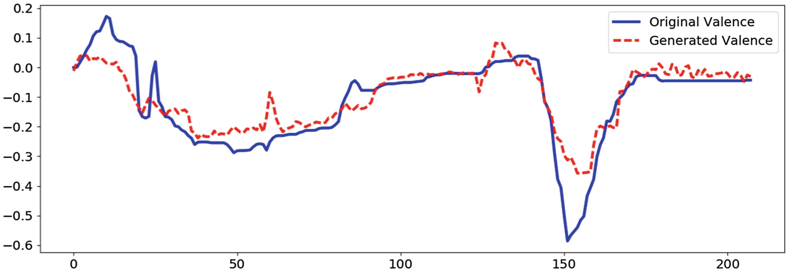

Fig. 1. An example of comparison between the ground truth (blue line) and the regressed value (red dashed line) on the Valence obtained by the trained convolutional network. (Color figure online)

As a quantitative and exhaustive evaluation of the regressor, we computed the root mean square error of the obtained emotional values, as reported in Table 2. The error is always contained, independently of the tested video, and this is reflected in the resulting low mean RMSE.

It is interesting to visualize network behavior. In Fig. 2 we show the input areas that are more informative for the regressor. We estimate these areas in a similar way as the *GRAD-Cam* method described in [29], using the mean of the feature maps of the last convolutional layer weighted by their backpropagated gradient. As we can observe the most informative areas concern the facial features, and in particular the eyes, but also, when the framing is larger, the body motion is captured and exploited to regress valence and arousal.

Table 2. Root Mean Square Error on the obtained values for each video and the mean value for the whole dataset.

Video	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	MEAN
RMSE	0.07	0.11	0.07	0.10	0.14	0.11	0.11	0.10	0.06	0.08	**0.08**

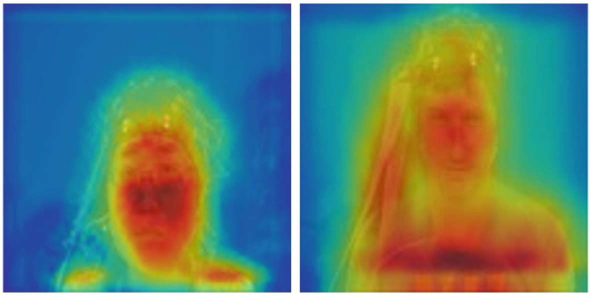

Fig. 2. Visualization of the last layer of the learnt convolutional network. Both the face and the body movements help the emotional state determination.

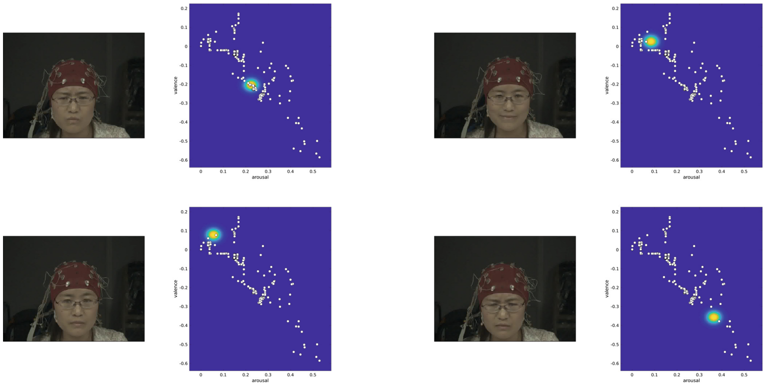

Fig. 3. Correspondences between the expresser and the "amygdala-like" activation.

Finally, in Fig. 3 we report the effect of the amygdala-like activation as a suitable prior for the latent manifold that has been learnt from valence-arousal labelling and that functionally models core-affect state-space. This very fast and automatic step it is likely to trigger the subsequent core affect representation of emotions.

5 Conclusions

We have presented a feed-forward mechanism providing a fast, automatic, and coarse processing of the earliest mapping from emotional facial expression stimuli to the core affect representation of emotions. Such mapping provides a suitable prior to the subsequent core affect dynamics, in which recognition is actually grounded.

To this end we designed a 3D spatiotemporal deep network as a suitable architecture to instantiate such early component. The network is quite shallow due to both the specificity of the task, and to the limited quantity of labeled data available.

The regression has produced satisfactory results on both valence and arousal dimensions, encouraging the integration of this preliminary step in a complete core affect model.

Acknowledgments. This research was carried out as part of the project "Interpreting emotions: a computational tool integrating facial expressions and biosignals based shape analysis and bayesian networks", supported by the Italian Government, managed by MIUR, financed by the *Future in Research* Fund.

References

1. Adolphs, R.: Neural systems for recognizing emotion. Curr. Opin. Neurobiol. **12**(2), 169–177 (2002)
2. Adolphs, R.: Recognizing emotion from facial expressions: psychological and neurological mechanisms. Behav. Cogn. Neurosci. Rev. **1**(1), 21–62 (2002)
3. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization (2016). arXiv preprint: arXiv:1607.06450
4. Bargal, S.A., Barsoum, E., Canton-Ferrer, C., Zhang, C.: Emotion recognition in the wild from videos using images. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, Tokyo, Japan, pp. 433–436, 12–16 November 2016
5. Bilakhia, S., Petridis, S., Nijholt, A., Pantic, M.: The mahnob mimicry database: a database of naturalistic human interactions. Pattern Recogn. Lett. **66**, 52–61 (2015)
6. Boccignone, G., Conte, D., Cuculo, V., Lanzarotti, R.: Amhuse: A multimodal dataset for humour sensing. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI 2017, pp. 438–445. ACM, New York, NY, USA (2017)
7. Boccignone, G., Cuculo, V., Grossi, G., Lanzarotti, R., Migliaccio, R.: Virtual emg via facial video analysis. In: Battiato, S., Gallo, G., Schettini, R., Stanco, F. (eds.) ICIAP 2017, Part I. LNCS, vol. 10484, pp. 197–207. Springer International Publishing, Cham (2017)
8. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the 6th International Conference on Multimodal Interfaces, ICMI 2004, pp. 205–211. ACM (2004)
9. Dalgleish, T.: The emotional brain. Nat. Rev. Neurosci. **5**(7), 583–589 (2004)
10. Darwin, C.: The Expression of the Emotions in Man and Animals. Oxford University Press, Oxford (1998)
11. D'mello, S.K., Kory, J.: A review and meta-analysis of multimodal affect detection systems. ACM Comput. Surv. (CSUR) **47**(3), 43 (2015)
12. Dubois, J., Adolphs, R.: Neuropsychology: how many emotions are there? Curr. Biol. **25**(15), R669–R672 (2015)
13. Ekman, P., Rosenberg, E.L.: What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). Oxford University Press, Oxford (1997)
14. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. **17**(2), 124 (1971)
15. Fontaine, J.R., Scherer, K.R., Roesch, E.B., Ellsworth, P.C.: The world of emotions is not two-dimensional. Psychol. Sci. **18**(12), 1050–1057 (2007)
16. Garrido, M.I., Barnes, G.R., Sahani, M., Dolan, R.J.: Functional evidence for a dual route to amygdala. Curr. Biol. **22**(2), 129–134 (2012)
17. Goldman, A.I., Sripada, C.S.: Simulationist models of face-based emotion recognition. Cognition **94**(3), 193–213 (2005)
18. Gunes, H., Schuller, B.: Categorical and dimensional affect analysis in continuous input: current trends and future directions. Image Vis. Comput. **31**(2), 120–136 (2013)
19. Hildebrandt, A., Recio, G., Sommer, W., Wilhelm, O., Ku, J.: Facial EMG responses to emotional expressions are related to emotion perception ability. PLoS ONE **9**(1), e84053 (2014)

20. Khorrami, P., Paine, T.L., Brady, K., Dagli, C.K., Huang, T.S.: How deep neural networks can improve emotion recognition on video data. In: 2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, pp. 619–623, 25–28 September 2016 (2016)
21. LeDoux, J.: The Emotional Brain: The Mysterious Underpinnings of Emotional Life. Simon and Schuster, New York (1998)
22. Lin, M., Chen, Q., Yan, S.: Network in network (2013). arXiv preprint: arXiv:1312.4400
23. Mase, K., Pentland, A.: Recognition of facial expression from optical flow. IEICE Trans. Inf. Syst. **74**, 3474–3483 (1991)
24. Picard, R.W.: Affective Computing. MIT Press, Cambridge (2000)
25. Russell, J.A.: Core affect and the psychological construction of emotion. Psychol. Rev. **110**(1), 145 (2003)
26. Salzman, C.D., Fusi, S.: Emotion, cognition, and mental state representation in amygdala and prefrontal cortex. Ann. Rev. Neurosci. **33**, 173–202 (2010)
27. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: a survey of registration, representation, and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(6), 1113–1133 (2015)
28. Scherer, K.R.: Emotion theories and concepts (psychological perspectives). In: Sander, D., Scherer, S.K.R. (eds.) Oxford Companion to Emotion and the Affective Sciences, pp. 145–149. Oxford University Press, Oxford (2009)
29. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: why did you say that? Visual explanations from deep networks via gradient-based localization (2016). arXiv preprint: arXiv:1610.02391
30. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the International Conference on Computer Vision, ICCV 2015 (2015)
31. Vitale, J., Williams, M.A., Johnston, B., Boccignone, G.: Affective facial expression processing via simulation: a probabilistic model. Biologically Inspired Cogn. Archit. J. **10**, 30–41 (2014)
32. Wood, A., Rychlowska, M., Korb, S., Niedenthal, P.: Fashioning the face: sensorimotor simulation contributes to facial expression recognition. Trends Cogn. Sci. **20**(3), 227–240 (2016)