Learning Emotional-Blinded Face Representations

Alejandro Peña, Julian Fierrez, Aythami Morales

BiDA Lab, Universidad Autónoma de Madrid, Spain

Àgata Lapedriza

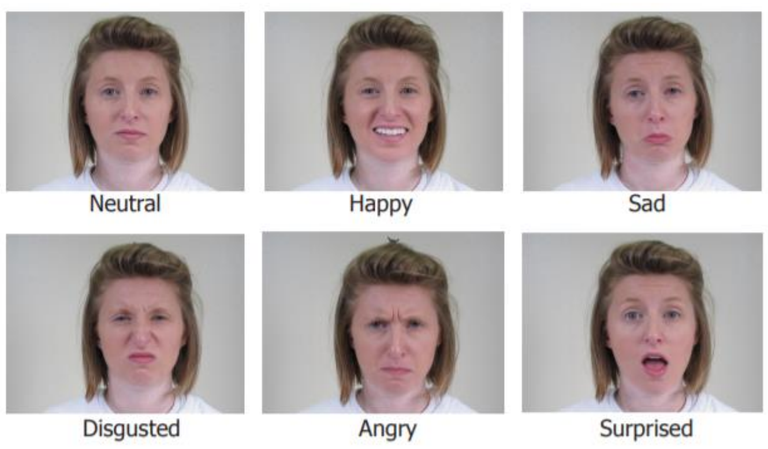
Universitat Oberta de Catalunya, Spain
Massachusetts Institute of Technology, USA

Abstract

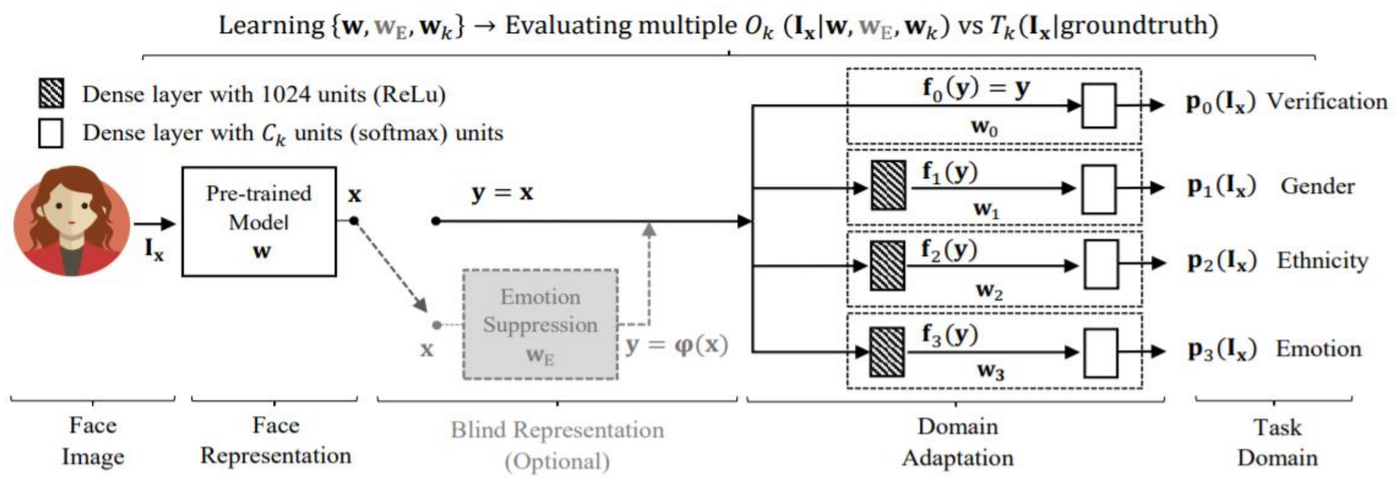
We propose two face representations that are **blind** to facial expressions associated to **emotional responses**. This work is in part motivated by new international regulations for **personal data protection**, which enforce data controllers to protect any kind of sensitive information involved in automatic processes. The advances in Affective Computing have contributed to improve human-machine interfaces but, at the same time, the capacity to monitorize emotional responses triggers potential risks for humans, both in terms of fairness and privacy. We propose two different methods to learn these **expression-blinded facial features**. Our experiments demonstrate that it is possible to eliminate information related to emotion recognition tasks in the face representations, while the performance of subject verification, gender recognition, and ethnicity classification are just slightly affected.

How emotions are expressed in face images

We study facial expressions [3] related to 6 different emotional responses.



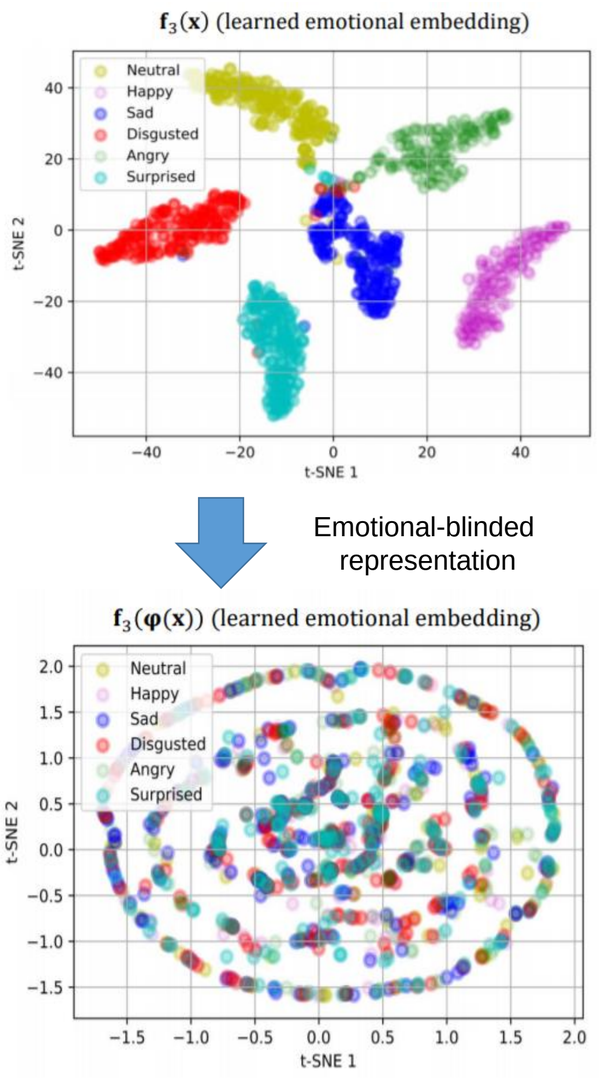
Proposed framework



We employ a **privacy-preserving** learning framework including domain adaption from a pre-trained face representation network to different face analysis tasks, with and without the **emotional-blinded representation**.

Emotional-blinded face representations

- We have adapted two existing methods to the problem of learning emotional-blinded representations, namely **SensitiveNets** [1] and **Learning Not To Learn** [2], and applied them effectively to the face recognition pre-trained model ResNet-50.
- We evaluated the performance of both the original representation **x** and the emotional-blinded **φ(x)** on subject verification in **LFW**, gender and ethnicity and classification in **DiveFace** [1], and emotion classification in **CFEE** [3]. Our experiments demonstrate that we can remove emotion related information, while preserving the performance in the other tasks.



Domain	x	$\varphi_{SN}(x)$	Diff. SN	$\varphi_{LnL}(x)$	Diff. LnL
ID	96.8 %	96.3 %	↓ 1 %	59.4 %	↓ 75 %
Gender	99.2 %	98.9 %	↓ 1 %	72.7 %	↓ 53.9 %
Ethnicity	98.8 %	98.6 %	↓ 1 %	67.4 %	↓ 47.9 %
Emotion (NN)	88.1 %	59.6 %	↓ 40 %	41.6 %	↓ 65 %
Emotion (SVM)	88.1 %	16.7 %	↓ 100 %	25 %	↓ 88.2 %
Emotion (RF)	77.4 %	58.3 %	↓ 31 %	44.7 %	↓ 53.8 %

Blind representations: Training fairer classifiers

- **Case study**: Attractiveness classification in presence of **facial expresion biases** using CelebA dataset.
- We design a 40K images training set, where 70% of attractive people are **smiling**, while 70% of unattractive people don't (both groups presenting 20K images).
- The emotional-blinded representations **reduce** the gap between both groups, and therefore improve the Equal of opportunity criterion.

Method (training)	Acc	TPR Smi.	TPR No Smi.	Eq. Opp.
x (unbiased)	77.26 %	84.55 %	82.47 %	2.08 %
x (biased)	76.23 %	84.17 %	66.70 %	17.47 %
$\varphi_{SN}(x)$ (biased)	74.50 %	81.87 %	73.58 %	8.29 %
$\varphi_{LNTL}(x)$ (biased)	76.62 %	86.97 %	73.70 %	13.27 %

Conclusions

The development of automatic systems capable of **reading emotions** without consent triggers a potential risk to user's privacy. In this work, we have adapted two methods for the purpose of generating **facial representations** that are blind to **facial expressions** associated to **emotional responses**. Our experiments demonstrate that is posible to reduce the performance of emotion classifiers while maintaining competitive performance in other face analysis tasks.

Main references

[1] A. Morales, J. Fierrez, R. Vera-Rodriguez and R. Tolosana, "SensitiveNets: Learning agnostic representations with application to face images", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[2] B. Kim, H. Kim, K. Kim, S. Kim and J. Kim, "Learning not to learn: Training deep neural networks with biased data", in *IEEE CVPR*, 2020.

[3] S. Du, Y. Tao and A. Martínez, "Compound facial expressions of emotions", in *Proceedings of the National Academy of Science*, 2014.

Transferable Technology (patent pending) ⟫ **See www. sensitivenets.com to learn more**