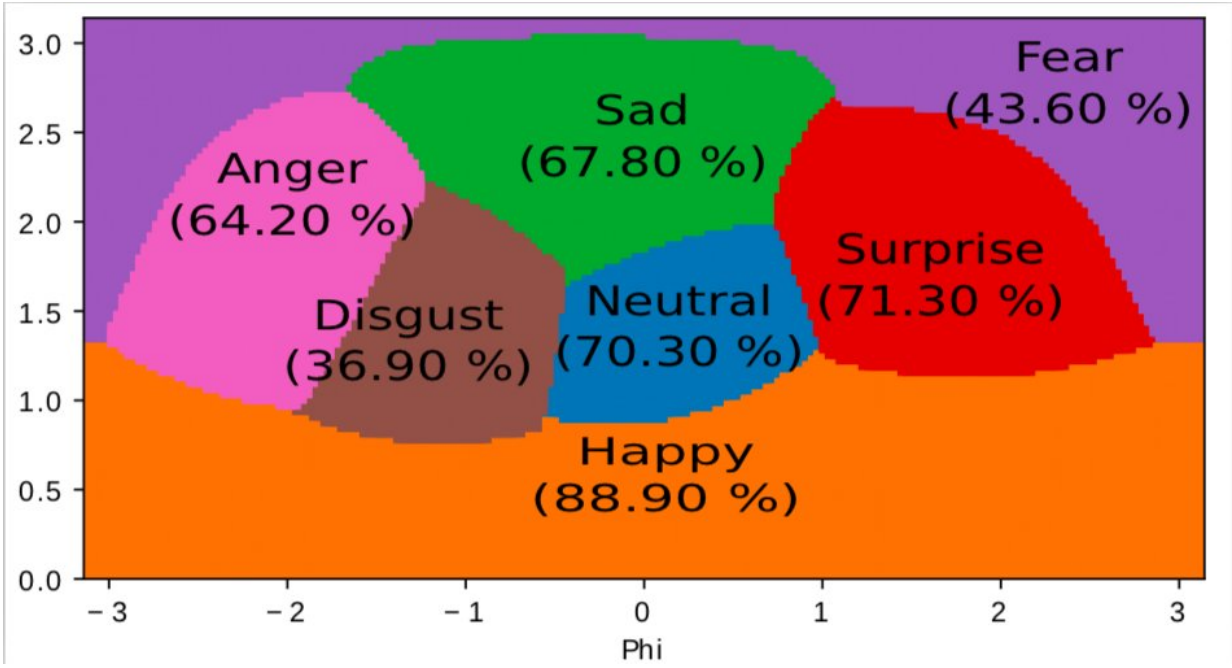
A light-weight and accurate deep learning model for audiovisual emotion recognition

August 17 2018, by Ingrid Fadelli



A representation of the internal space learned by our algorithm and used to map emotions into a 2D continuous space. It is interesting to note that even if the training data only contain discrete emotion labels, the network learns a continuous space, allowing not only to describe finely the emotional state of people but also to position emotions in relation to each other. This space bears strong similarity with the arousal valence space defined by modern psychology. Credit: Jurie et al.

Researchers at Orange Labs and Normandie University have developed a

novel deep neural model for audiovisual emotion recognition that performs well with small training sets. Their study, which was pre-published on *arXiv*, follows a philosophy of simplicity, substantially limiting the parameters that the model acquires from datasets and using simple learning techniques.

Neural networks for emotion recognition have a number of useful applications within the contexts of healthcare, customer analysis, surveillance, and even animation. While state-of-the-art deep learning algorithms have achieved remarkable results, most are still unable to reach the same understanding of emotions attained by humans.

"Our overall objective is to facilitate human-computer interaction by making computers able to perceive various subtle details expressed by humans," Frédéric Jurie, one of the researchers who carried out the study, told TechXplore. "Perceiving emotions contained in images, video, voice and sound fall within this context."

Recently, studies have put together multimodal and temporal datasets that contain annotated videos and audiovisual clips. Yet these datasets typically contain a relatively small number of annotated samples, while to perform well, most existing deep learning algorithms require larger datasets.

The researchers tried to address this issue by developing a new framework for audiovisual emotion recognition, which fuses the analysis of visual and audio footage, retaining a high level of accuracy even with relatively small training datasets. They trained their neural model on AFEW, a dataset of 773 audiovisual clips extracted from movies and annotated with discrete emotions.

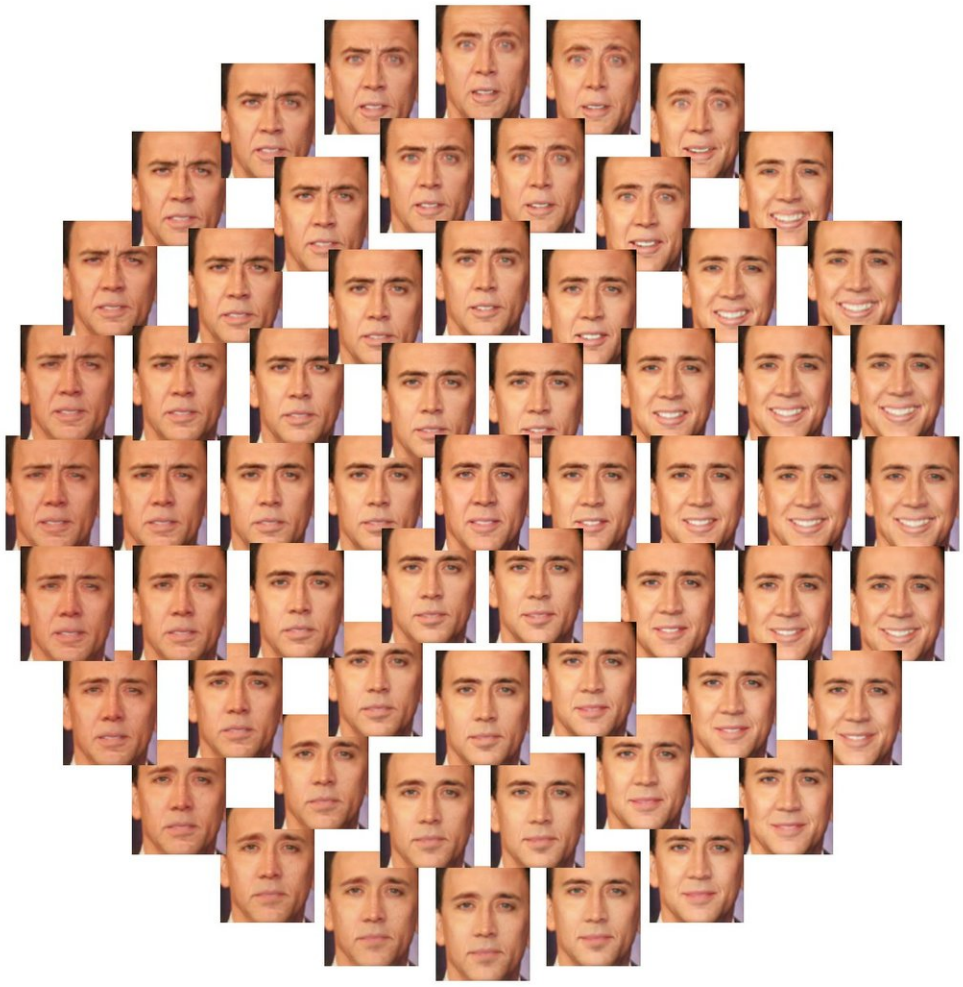


Illustration of how this 2D space can be used to control emotions expressed by faces, in a continuous way, with the help of adversarial generative networks (GAN). Credit: Jurie et al.

"One can see this model as a black box processing the video and automatically inferring the emotional state of people," Jurie explained.

"One big advantage of such deep neural models is that they learn by themselves how to process the video by analyzing examples, and do not require experts to provide specific processing units."

The model devised by the researchers follows the Occam's razor philosophical principle, which suggests that between two approaches or explanations, the simplest one is the best choice. Contrarily to other deep learning models for emotion recognition, therefore, their model is kept relatively simple. The neural network learns a limited number of parameters from the dataset and employs basic learning strategies.

"The proposed network is made of cascaded processing layers abstracting the information, from the signal to its interpretation," Jurie said. "Audio and video are processed by two different channels of the network and are combined lately in the process, almost at the end."

When tested, their light model achieved a promising emotion recognition accuracy of 60.64 percent. It was also ranked fourth at the 2018 Emotion Recognition in the Wild (EmotiW) challenge, held at the ACM International Conference on Multimodal Interaction (ICMI), in Colorado.

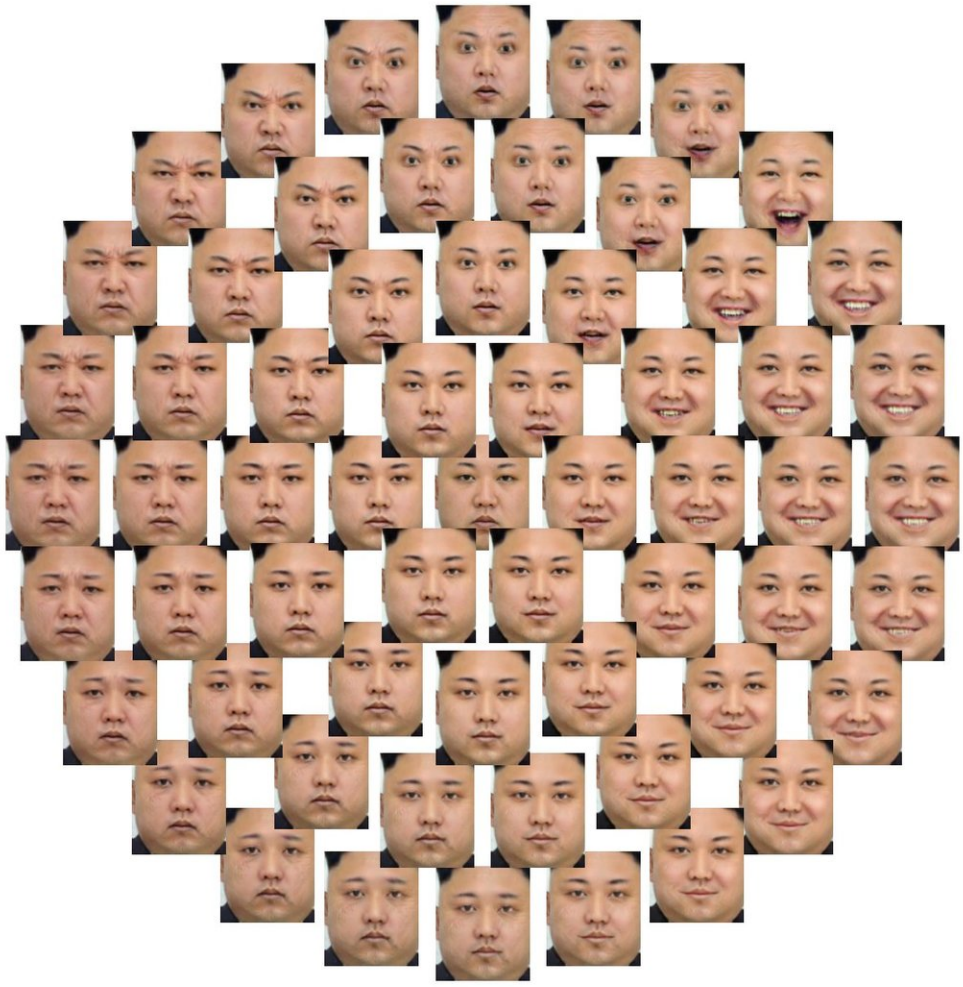


Illustration of how this 2D space can be used to control emotions expressed by faces, in a continuous way, with the help of adversarial generative networks (GAN). Credit: Jurie et al.

"Our model is proof that following the Occam's razor principle, i.e., by always choosing the simplest alternatives for designing neural networks,

it is possible to limit the size of the models and obtain very compact but state-of-the-art neural networks, which are easier to train," Jurie said. "This contrasts with the research trend of making neural networks bigger and bigger."

The researchers will now continue to explore ways of achieving high accuracy in emotion recognition by simultaneously analyzing visual and auditory data, using the limited annotated training datasets that are currently available.

"We are interested in several research directions, such as how to better fuse the different modalities, how to represent emotion by compact semantically meaning full descriptors (and not only class labels) or how to make our algorithms able to learn with less, or even without, annotated data," Jurie said.

More information: An Occam's Razor View on Learning Audiovisual Emotion Recognition with Small Training Sets, arXiv:1808.02668v1 [cs.AI]. arxiv.org/abs/1808.02668

Abstract
This paper presents a light-weight and accurate deep neural model for audiovisual emotion recognition. To design this model, the authors followed a philosophy of simplicity, drastically limiting the number of parameters to learn from the target datasets, always choosing the simplest earning methods: i) transfer learning and low-dimensional space embedding allows to reduce the dimensionality of the representations. ii) The isual temporal information is handled by a simple score-per-frame selection process, averaged across time. iii) A simple frame selection echanism is also proposed to weight the images of a sequence. iv) The fusion of the different modalities is performed at prediction level (late usion). We also highlight the inherent challenges of the AFEW dataset and the difficulty of model selection with as few as 383 validation

equences. The proposed real-time emotion classifier achieved a state-of-the-art accuracy of 60.64 % on the test set of AFEW, and ranked 4th at he Emotion in the Wild 2018 challenge.