HAL
open science

CAKE: Compact and Accurate K-dimensional representation of Emotion

Corentin Kervadec, Valentin Vielzeuf, Stéphane Pateux, Alexis Lechervy, Frédéric Jurie

▶ To cite this version:

Corentin Kervadec, Valentin Vielzeuf, Stéphane Pateux, Alexis Lechervy, Frédéric Jurie. CAKE: Compact and Accurate K-dimensional representation of Emotion. Image Analysis for Human Facial and Activity Recognition (BMVC Workshop), Dr. Zhaojie Ju, Sep 2018, Newcastle, United Kingdom. hal-01849908v2

HAL Id: hal-01849908

https://hal.science/hal-01849908v2

Submitted on 2 Aug 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CAKE: Compact and Accurate K-dimensional representation of Emotion

Corentin Kervadec*[1]
corentin.kervadec@orange.com
Valentin Vielzeuf*[12]
valentin.vielzeuf@orange.com
Stéphane Pateux[1]
stephane.pateux@orange.com
Alexis Lechervy[2]
alexis.lechervy@unicaen.fr
Frédéric Jurie[2]
frederic.jurie@unicaen.fr

[1] Orange Labs, Cesson-Sévigné, France
[2] Normandie Univ., UNICAEN, ENSICAEN, CNRS Caen, France

*Both authors contributed equally.

Abstract

Numerous models describing the human emotional states have been built by the psychology community. Alongside, Deep Neural Networks (DNN) are reaching excellent performances and are becoming interesting features extraction tools in many computer vision tasks. Inspired by works from the psychology community, we first study the link between the compact two-dimensional representation of the emotion known as *arousal-valence*, and discrete emotion classes (e.g. anger, happiness, sadness, *etc.*) used in the computer vision community. It enables to assess the benefits – in terms of discrete emotion inference – of adding an extra dimension to arousal-valence (usually named dominance). Building on these observations, we propose CAKE, a 3-dimensional representation of emotion learned in a multi-domain fashion, achieving accurate emotion recognition on several public datasets. Moreover, we visualize how emotions boundaries are organized inside DNN representations and show that DNNs are implicitly learning arousal-valence-like descriptions of emotions. Finally, we use the CAKE representation to compare the quality of the annotations of different public datasets.

1 Introduction

Facial expression is one of the most used human means of communication after language. Thus, the automated recognition of facial expressions – such as emotions – has a key role in affective computing, and its development could benefit human-machine interactions.

Different models are used to represent human emotion states. Ekman *et al.* [6] propose to classify the human facial expression resulting from an emotion into six classes (*resp.* happiness, sadness, anger, disgust, surprise and fear) supposed to be independent across the cultures. This model has the benefit of simplicity but could be not sufficient to address the

© 2018. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

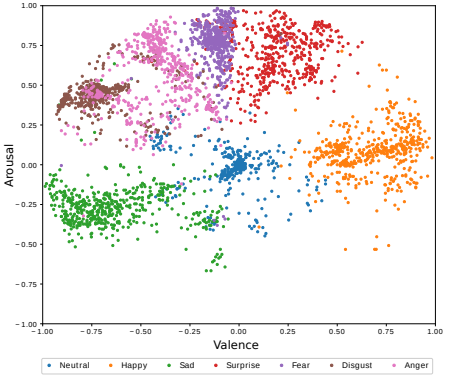

Figure 1: Comparison of the discrete and continuous (arousal-valence) representations using AffectNet's annotations [17].

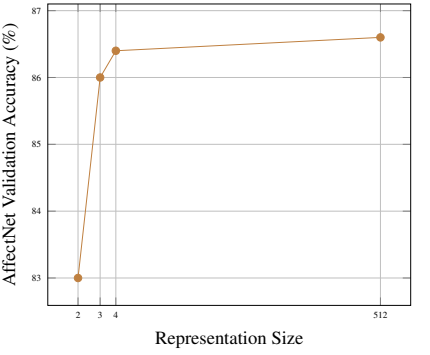

Figure 2: Influence of adding supplementary dimensions to arousal-valence when predicting emotion on AffectNet [17].

whole complexity of human affect. Moreover it suffers from serious intra-class variations as, for instance, soft smile and laughing equally belong to *happiness*. That is why Ekman's emotion classes are sometimes assembled into compound emotions [5] (*e.g.* happily surprised). Others have chosen to represent emotion with an n-dimensional continuous space, as opposite to the Ekman's discrete classes. Russel has built the *Circumplex Model of Affect* [20] in which emotion states are described by two values: arousal and valence. *Arousal* represents the excitation rate – the higher the arousal is, the more intense the emotion is – and *valence* defines whether the emotion has a positive or a negative impact on the subject. Russels suggests in [20] that all Ekman's emotions [6] and compound emotions could be mapped in the *circumplex model of affect*. Furthermore, this two-dimensional approach allows a more accurate specification of the emotional state, especially by taking its intensity into account.

A third dimension has been added by Mehrabian *et al.* [16] – the *dominance* – which depends on the degree of control exerted by a stimulus. Last, Ekman and Friesen [7] have come up with the *Facial Action Code System* (FACS) using anatomically based action units. Developed for measuring facial movements, FACS is well suited for classifying facial expressions resulting from an affect.

Based on these emotion representations, several large databases of face images have been collected and annotated according to emotion. EmotioNet [8] gathers faces annotated with Action Units [7]; SFEW [4], FER-13 [9] and RAF [15] propose images in the wild annotated in basic emotions; AffecNet [17] is a database annotated in both discrete emotion [6] and arousal-valence [20].

The emergence of these large databases has allowed to develop automatic emotion recognition systems, such as the recent approaches based on Deep Neural Networks (DNN). AffectNet's authors [17] use three AlexNet [13] to learn respectively emotion classes, arousal and valence. In [18], the authors make use of transfer learning to counteract the smallness of the SFEW [4] dataset, by pre-training their model on ImageNet [2] and FER [9]. In [1] authors implement *Covariance Pooling* using second order statistics when training on emotion recognition (on RAF [15] and SFEW [4]).

Emotion labels, FACS and continuous representations have their own benefits – simplic-

ity of the emotion classes, accuracy of the arousal-valence, objectivity of the FACS, *etc.* – but also their own drawbacks – imprecision, complexity, ambiguity, *etc*. Therefore several authors have tried to leverage the benefits of all these representations. Khorrami *et al.* [11] first showed that neural networks trained for expression recognition implicitly learn facial action units. Contributing to highlighting the close relation between emotion and Action Units, Pons *et al.* [19] learned a multitask and multi-domain ResNet [10] on both discrete emotion classes (SFEW [4]) and Action Units (EmotioNet [8]). Finally, Li *et al.* [15] proposed a "*Deep Locality-Preserving Learning*" to handle the variability inside an emotion class, by making classes as compact as possible.

In this context, this paper focuses on the links between arousal-valence and discrete emotion representations for image-based emotion recognition. More specifically, the paper proposes a methodology for learning very compact embedding, with not more than 3 dimensions, performing very well on emotion classification task, making the visualization of emotions easy, and bearing similarity with the arousal-valence representation.

2 Learning Very Compact Emotion Embeddings

2.1 Some Intuitions About Emotion Representations

We first want to experimentally measure the dependence between emotion and arousal-valence as yielded in [20]. We thus display each sample of the AffectNet [17] validation subset in the arousal-valence space and color them according to their emotion label (Figure 1). For instance, a face image labelled as *neutral* with an arousal and a valence of zero is located at the center of Figure 1 and colored in blue. It clearly appears that a strong dependence exists between discrete emotion classes and arousal-valence. Obviously, it is due in part to the annotations of the AffectNet [17] dataset, as the arousal-valence have been constrained to lie in a predefined confidence area based on the emotion annotation. Nevertheless, this dependence agrees with the *Circumplex Model of Affect* [20].

To evaluate further how arousal-valence representation is linked to emotion labels, we train a classifier made of one fully connected layer[1] (fc-layer) to infer emotion classes from arousal-valence values provided by AffectNet [17] dataset. We obtain the accuracy of 83%, confirming that arousal-valence can be an excellent *2-d* compact emotion representation.

This raises the question of the optimality of this 2-*d* representation. Would adding a third dimension to arousal-valence make the classification performance better? To address this question, we used the 512-*d* hidden representation of a ResNet-18 [10] trained to predict discrete emotions on the AffectNet dataset [17]. This representation is then projected into a more compact space using a fc-layer outputting k dimensions, which are concatenated with the arousal-valence values. On top of this representation, we add another fc-layer predicting emotion classes. The two fc-layers are finally trained using Adam optimizer [12]. Adding 1 dimension to arousal-valence gives a gain of +3 points on the accuracy. It agrees with the assumption that a three-dimensional representation is more meaningful than a two-dimensional one [16]. The benefit of adding more than 1 dimension is exponentially decreasing; with +512 dimensions, the gain is only of +0.6 points compared to adding 1 dimension, as shown in Figure 2.

From these observations, the use of a compact representation seems to be consistent with discrete emotion classes, as it enables an accuracy of 83% and 86% – respectively for a 2-*d*

[1] By "fully connected layer" we denote a linear layer with biases and without activation function.

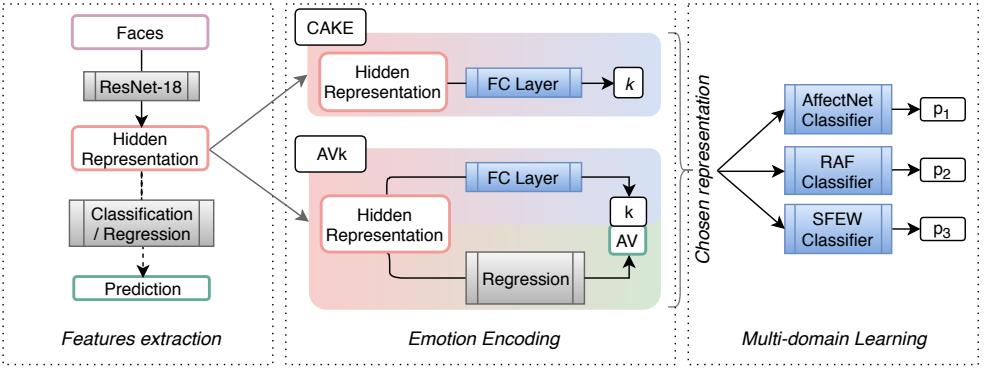

Figure 3: Our approach's overview. Left: we use a ResNet-18 previously trained for discrete emotion recognition or arousal valence regression to extract 512-d hidden representations from face images. Center: using these hidden representations, CAKE or AVk representations (center) are learned to predict discrete emotions. Right: the learning process is multi-domain, predicting emotions on three different datasets with three different classifiers. Gray blocks are non-trainable weights while blue blocks are optimized weights.

and a 3-*d* representation – and it even may allow to describe affect states with more contrast and accuracy. Even if arousal-valence is a good representation for emotion recognition, the question of its optimality has not been answered by these preliminary experiments. In other words, is it possible to learn 2-*d* (or 3-*d*) embedding better than those built on arousal-valence? We positively answer this question in Section 2.2.

2.2 Learning Compact and Accurate Representations of Emotions

Based on the previous observations, this section proposes a methodology for learning a compact embedding for emotion recognition from images.

Features extraction The basic input of our model is an image containing one face displaying a given emotion. We first extract 512-*d* features specialized in emotion recognition. So as to, we detect the face, align its landmarks by applying an affine transform and crop the face region. The so-obtained face is then resized into 224×224 and fed to a ResNet-18 [10] network (Figure 3, *Features extraction*). The face image is augmented (*e.g.* jittering, rotation), mostly to take the face detector noise into account. We also use cutout [3] – consisting in randomly cutting a 45×45 pixels sized patch from the image – to regularize and improve the robustness of our model to facial occlusions. Our ResNet outputs 512-*d* features, on top of which a fc-layer can be added. At training time, we also use dropout [21] regularization. The neural network can be learned from scratch on two given tasks: discrete emotion classification or arousal-valence regression.

Compact emotion encoding Compact embedding is obtained by projecting the 512-*d* features provided by the ResNet-18 (pretrained on discrete emotion recognition) into smaller k-dimensional spaces (Figure 3, *Emotion Encoding*) in which the final classification is done. The *k* features may be seen as a compact representation of the emotion, and the performance

of the classifier can be measured for different values of k. CAKE-2, CAKE-3, *etc.*, denote such classifiers with $k = 2$, $k = 3$, *etc.*

In the same fashion we can train the ResNet-18 using arousal-valence regression. In this case, the so-obtained arousal-valence regressor can be used to infer arousal-valence values from novel images and concatenate them to the k features of the embedding. Thus we reproduce here the exact experiment done in Section 2.1 in order to assess the benefit of a third (or more) dimension. The difference is that arousal-valence are not ground truth values but predicted ones. These methods are denoted as AV1, AV2, AV3, *etc.* for the different values of k.

Domain independent embedding As we want to ensure a generic compact enough representation, independent of the datasets, we learn the previously described model jointly on several datasets, without any further fine-tuning.

Our corpus is composed of AffectNet [17], RAF [15] and SFEW [4], labelled with seven discrete emotion classes: *neutral*, *happiness*, *sad*, *surprise*, *fear*, *disgust* and *anger*. Our training subset is composed of those of AffectNet (283901 elts., *95.9%* of total), RAF (11271 elts., *3.81%* of total) and SFEW (871 elts., *0.29%* of total). Our testing subset is composed of the subsets commonly used for evaluation in the literature (*validation* of SFEW and AffecNet, *test* of RAF).

To ease the multi-domain training, we first pre-train our features extractor model on AffectNet and freeze its weights. Then we apply the same architectures as described before, but duplicate the last fc-layer in charge of emotion classification in three dataset-specific layers (Figure 3, *multi-domain learning*). The whole model loss is a modified softmax cross entropy defined as follows:

$$Loss = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{3} w_{class}^{i,j} w_{dataset}^{j}\, E(y^i, \hat{y}^{i,j}) \tag{1}$$

where j is ranging in [AffectNet, RAF, SFEW], y^i is the label of ith element, $\hat{y}^{i,j}$ is the prediction of the jth classifier on the ith element, E is the softmax cross entropy loss, N is the number of elements in the batch, w_{class}^{i} is a weight given to the ith element of the batch depending on its emotion class and $w_{dataset}^{j}$ is a weight given to the jth classifier prediction. Each sample of the multi-domain dataset is identified according to its original database, allowing to choose the correct classifier's output when computing the softmax cross entropy.

The w_{class} weight is defined as: $w_{class}^{i,j} = \frac{N_{total}^{j}}{N_{class}^{i,j} \times nbclass}$ where N_{total}^{j} is the number of elements in the jth dataset, $N_{class}^{i,j}$ is the number of elements in the class of the ith element of the jth dataset and *nbclass* is the number of classes (7 in our case). The goal here is to fix the important class imbalance in the dataset by forcing to fit the uniform distribution, as previously done by [17].

The $w_{dataset}$ weight permits to take the imbalance between dataset's sizes into account.

$$w_{dataset}^{j} = \begin{cases} \frac{1}{\log N_{total}^{j}} & sample \in j^{th} dataset \\ 0 & sample \notin j^{th} dataset \end{cases} \tag{2}$$

We thus define a global loss enabling to optimize the last two layers of our model (namely *Emotion Encoding* and *Multi-domain Learning* in Figure 3) on the three datasets at the same

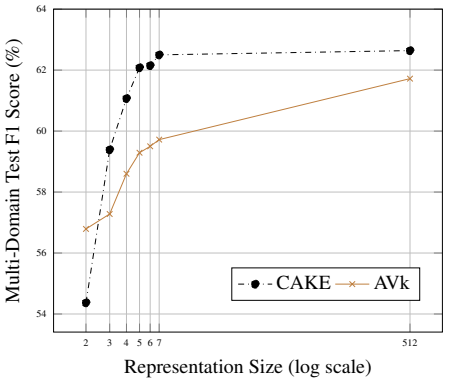

Figure 4: Influence of representation size on the multi-domain F1 score.

Dataset	Rep. & Dim.		F1 Score
Affect -Net	CAKE-3	3	**58.1 ± 0.5**
	AV1	3	55.6 ± 0.5
	AV	2	55.8 ± 0.0
	CAKE-2	2	52.1 ± 0.4
SFEW	CAKE-3	3	**34,1 ± 1.0**
	AV1	3	30.2 ± 0.8
	AV	2	33.3 ± 0.1
	CAKE-2	2	28.0 ± 0.8
RAF	CAKE-3	3	**64.4 ± 0.5**
	AV1	3	63.0 ± 0.9
	AV	2	61.2 ± 0.2
	CAKE-2	2	60.6 ± 1.9

Table 1: Evaluation of compact representations on AffectNet, SFEW, RAF.

time. The dimension k (or $k+2$ in the case of the arousal-valence approach) can easily be changed and help to evaluate the interest of supplementary dimensions for emotion representation.

3 Experiments

3.1 Evaluation Metrics

We measure the classification performance with the *accuracy* and the ***macro*** *F1 Score* (3). *Accuracy* measures the number of correctly classified samples. Instead of accuracy, we prefer *macro F1 score* which gives the same importance to each class:

$$F_{1macro} = \frac{1}{N_c}\sum_{i}^{N_c} F_{1i} \quad F_{1i} = 2\frac{prec_i \cdot rec_i}{prec_i + rec_i} \quad prec_i = \frac{tp_i}{tp_i + fp_i} \quad rec_i = \frac{tp_i}{tp_i + fn_i} \tag{3}$$

where i is the class index; $prec_i$, rec_i and F_{1i} are the precision, the recall and the F1-score of class i; N_c is the number of classes; tp, fp and fn are the true positives, false positives and false negatives rates. All scores are averaged on 10 runs, with different initializations, and given with associated standard deviations, on our multi-domain testing subset.

3.2 Compactness of the Representation

We first evaluate the quality of the representations in a multi-domain setting. Table 1 reports the F1-score of CAKE-2, AV, CAKE-3 and AV1 trained on three datasets with three different classifiers, each one being specialized on a dataset as explained in Section 2. Among the 2-d models (AV and CAKE-2), AV is better, taking benefits from the knowledge transferred from the AffectNet dataset. This is not true anymore for the 3D models, where CAKE-3 is better than AV1, probably because of its greater number of trainable parameters.

To validate the hypothesis of the important gain brought by adding a third dimension, we run the "CAKE" and "AVk" experiments with different representation sizes. To simplify

	Rep. Dim.	RAF [15]	SFEW [4]	AffectNet [17]
Covariance Pooling [1]	2000	79.4	-	-
	512	-	58.1	-
Deep Locality Preserving [15]	2000	74.2	51.0	-
Compact Model [14]	64	67.6	-	-
VGG[15]	2000	58.2	-	-
Transfer Learning [18]	4096	-	48.5	-
ours (CAKE-3)	3	68.9	44.7	58.2
ours (Baseline)	512	71.7	48.7	61.7

Table 2: Accuracy of our model regarding state-of-the-art methods. The size of the representation is taken into account. Metrics are the average of per class recall for RAF and accuracy for SFEW and AffectNet.

the analysis of the results, we plot in Figure 4 a multi-domain F1-score, *i.e.* the weighted average of the F1-scores according to the respective validation set sizes. We observe that the gain in multi-domain F1-score is exponentially decreasing for both representations – note that the representation size axis is in log scale – and thus the performance gap between a representation of size 2 and size 3 is the more important. We also observe that "CAKE" representations still seem to yield better results than "AVk" when the representation size is greater than 2.

These first experiment shows that a very compact representation can yield good performances for emotion recognition. It also is in line with the "dominance" dimension hypothesis, as a third dimension brought the more significant gain in performance. After 3 dimensions, the gain is much less significant.

3.3 Accuracy of the Representation

To evaluate the efficiency of the CAKE-3 compact representation, we compare its accuracy with state-of-the-art approaches (Table 2) on the public datasets commonly used in the literature for evaluation (*validation* of SFEW and AffecNet, *test* of RAF). In order to get a fair comparison, we add a *"Rep. Dim."* column corresponding to the size of the last hidden representation – concretely, we take the penultimate fully connected output size. We report the scores under the literature's metrics, namely the mean of the per class recall for RAF [15] and the accuracy for SFEW [4] and AffectNet [17]. To the best of the author's knowledge no other model has been evaluated before on the AffectNet's seven classes.

CAKE-3 is outperformed by Covariance Pooling [1] and Deep Locality Preserving [15]. Nevertheless, it is still competitive as the emotion representation is far more compact – 3-*d* *versus* 2000-*d* – and learned in a multi-domain fashion. Moreover, we gain 1 point on RAF when we compare to models of same size (2 millions parameters), *e.g.* *Compact Model* [14]. These results support the conclusion made in 3.2, as we show that a compact representation of the emotion learned by small models is competitive with larger representations. This finally underlines that facial expressions may be encoded efficiently into a 3-*d* vector and that using a large embedding on small datasets may lead to exploit biases of the dataset more than to learn emotion recognition.

Our experiments also allow to perform a cross-database study as done in [15]. This study consists in evaluating a model trained on dataset B on a dataset A. Thereby we obtain

		Dataset		
		AffectNet	SFEW	RAF
Classifier	AffectNet	**58.1** *(± 0.5)*	27.6 *(± 2.6)*	53.8 *(± 0.6)*
	SFEW	35.1 *(± 2.1)*	**34.1** *(± 1.0)*	47.3 *(± 1.2)*
	RAF	51.8 *(± 0.4)*	31.5 *(± 1.7)*	**64.4** *(± 0.6)*

Table 3: Cross-database evaluation on CAKE-3 model (F1-Score).

Table 3 with the evaluation of each classifier on each dataset. Results on SFEW [4] – trained or evaluated – are constantly lower than others, with a higher standard deviation. This could be due to the insufficient number of samples in the SFEW training set or more probably to the possible ambiguity in the annotation of SFEW compared to AffectNet and RAF. Confirming this last hypothesis, the *RAF classifier* has the better generalization among the datasets. It is in line with the claim of Li *et al.* [15] that RAF has a really reliable annotation with a large consensus between different annotators. Finally, it also underlines the difficulty to find a reliable evaluation of an emotion recognition system because of the important differences between datasets annotations.

3.4 Visualizing Emotion Maps

Visualizations are essential to better appreciate how DNN performs classifications, as well as to visualize emotion boundaries and their variations across datasets. Our visualization method consists in densely sampling the compact representation space – 2-*d* or 3-*d* – into a mesh grid, and feeding it to a formerly trained model – AV, CAKE-2 or CAKE-3 – in order to compute a dense map of the predicted emotions. Not all the coordinates of the mesh grid belong to real emotions and some of them would never happen in real applications.

The construction of the mesh grid depends on the model to be used. For the AV and the CAKE-2 models, we have simply built it using 2d vectors with all values ranging in intervals containing maximum and minimum values of the coordinates observed with real images. As the CAKE-3 model is dealing with a three-dimensional representation, it is not possible to visualize it directly on a plane figure. To overcome this issue we modify CAKE-3 into a CAKE-3-Norm representation where all the coordinates are constrained to be on the surface of the unit sphere, and visualize spherical coordinates. Even if CAKE-3-Norm shows lower performances (about 2 points less than CAKE-3), the visualization is still interesting, bringing some incentives about what has really been learned.

Figure 5 shows the visualization results for CAKE-3-Norm, AV and CAKE-2 representations (*resp.* from top to down). Each dot is located by the coordinates of its compact representation – $(arousal, valence)$ for AV, (k_1, k_2) for CAKE-2 and spherical coordinates (ϕ and θ) for CAKE-3-Norm – and colored according to the classifier prediction. The per class macro F1-score is displayed inside each emotion area.

First, each compact representation – CAKE-2, CAKE-3-Norm and AV – exhibits a strong consistency across the datasets (in Figure 5, compare visualizations on the same row). Indeed, the three classifiers show a very similar organization of the emotion classes, which is demonstrating the reliability of the learned representation. Thereby, the *neutral* class – in blue – is always placed at the origin and tends to neighbor all other classes. It is in line with the idea of neutral as an emotion with a very low intensity. Nevertheless, we can witness small inter-dataset variations, especially on SFEW [4] (in Figure 5, middle column) with

disgust and *fear* – *resp.* brown and purple – which are almost missing. This underlines the disparities of annotations across the datasets and confirms the need of multi-domain frameworks when wishing to achieve a more general emotion recognition model.

Second, we can analyze variations between the different representations for a given dataset (in Figure 5, compare visualizations on the same column). As AV is based on arousal-valence, we observe the same emotion organization as in Figure 1. Especially, as the majority of the AffectNet's training (and validation) samples have a positive arousal, the classifier do not use the whole space (in Figure 5, second row: see green, blue and orange areas) unlike CAKE-2 and CAKE-3 which are not constrained by arousal-valence.

We can find many similarities between these three representations, but the most impressive come across when comparing CAKE-2 and AV. Despite the inequality of scaling – which causes the *neutral* area (blue) to be smaller in CAKE-2 – AV and CAKE-2 compact representations are very close. Indeed, the area classes are organized exactly in the same fashion. The only difference is that for AV they are disposed in a clockwise order around *neutral* whereas for CAKE-2 they are disposed in an anticlockwise order. This observation shows that a DNN trained on the emotion recognition classification is able to learn an arousal-valence-like representation of the emotion. It contributes – along with Khorrami [10] who points that DNNs trained to recognize emotions are learning action units [7] – to bring the dependence across the emotion representations in the forefront.

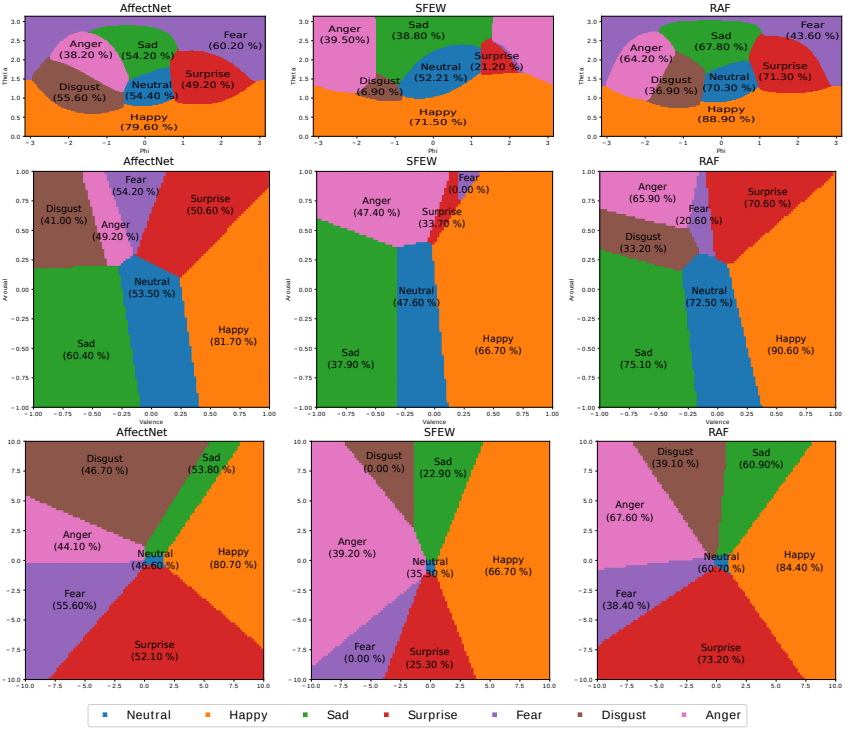

Figure 5: Visualization of CAKE-3-Norm, AV and CAKE-2. Rows indicate evaluated representation – *resp.* from top to down: CAKE-3-Norm, AV, CAKE-2 – and columns indicate datasets – *resp.* from left to right: AffectNet [17], SFEW [4] and RAF [15].

4 Conclusion

This work proposes a comprehensive analyze on how a DNN can describe emotional states. To this purpose, we first studied how many dimensions are sufficient to accurately represent an emotion resulting from a facial expression. We then conclude that three dimensions are a good trade-off between accuracy and compactness, agreeing with the arousal-valence-dominance [20][16] psychologist model. Thereby, we came up with a DNN providing a 3-dimensional compact representation of emotion, learned in a multi-domain fashion on RAF [15], SFEW [4] and AffecNet [17]. We set up a comparison with the state-of-the-arts and showed that our model can compete with models having much larger feature sizes. It proves that bigger representations are not necessary for emotion recognition. In addition, we implemented a visualization process enabling to qualitatively evaluate the consistency of the compact features extracted from emotion faces by our model. We thus showed that DNN trained on emotion recognition are naturally learning an arousal-valence-like [20] encoding of the emotion. As a future work we plan to also apply state-of-the-art techniques – as Deep Locality Preserving Loss [15] or Covariance Pooling [1] – to enhance our compact representation. In addition, nothing warranty that the learned CAKE bears the same semantic meanings as arousal-valence-dominance does: further interpreting the perceived semantic of the dimensions would therefore be an interesting piece of work.

References

[1] Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Covariance pooling for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 367–374, 2018.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[3] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[4] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2106–2112. IEEE, 2011.

[5] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, page 201322355, 2014.

[6] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[7] Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976.

[8] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5562–5570, 2016.

[9] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer, 2013.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Pooya Khorrami, Thomas Paine, and Thomas Huang. Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 19–27, 2015.

[12] Diederik P Kingma and Jimmy Lei Ba. Adam: Amethod for stochastic optimization. In *Proc. 3rd Int. Conf. Learn. Representations*, 2014.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[14] Chieh-Ming Kuo, Shang-Hong Lai, and Michel Sarkis. A compact deep learning model for robust facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2121–2129, 2018.

[15] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 2584–2593. IEEE, 2017.

[16] Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.

[17] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 2017.

[18] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 443–449. ACM, 2015.

[19] Gerard Pons and David Masip. Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition. *arXiv preprint arXiv:1802.06664*, 2018.

[20] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[21] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.