

# Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition

Gerard Pons · David Masip

Received: date / Accepted: date

**Abstract** Automated emotion recognition *in the wild* from facial images remains a challenging problem. Although recent advances in Deep Learning have supposed a significant breakthrough in this topic, strong changes in pose, orientation and point of view severely harm current approaches. In addition, the acquisition of labeled datasets is costly, and current state-of-the-art deep learning algorithms cannot model all the aforementioned difficulties. In this paper, we propose to apply a multi-task learning loss function to share a common feature representation with other related tasks. Particularly we show that emotion recognition benefits from jointly learning a model with a detector of facial Action Units (collective muscle movements). The proposed loss function addresses the problem of learning multiple tasks with heterogeneously labeled data, improving previous multi-task approaches. We validate the proposal using two datasets acquired in non controlled environments, and an application to predict compound facial emotion expressions.

**Keywords** Facial emotion recognition · Facial Action Units · Multi-task learning · Convolutional neural networks

## 1 Introduction

Images of faces provide relevant information for emotion perception. As humans we can infer an accurate first impression of somebody's emotions by observing their face. Multiple applications benefit from automated

facial emotion recognition, such as human computer interaction (HCI) [4], student engagement estimation [43], emotionally aware devices [41], or the improvement of expression production in autism disorder patients [8].

The state-of-the-art in automated facial expression analysis shows excellent performance in the controlled scenario, where images are acquired in studio environments. Nevertheless, the categorization of emotions *in the wild*, is still an unsolved problem. Besides the strong intra-class variability, facial expression algorithms *in the wild* must also deal with strong local changes in the illumination conditions, out of plane rotations, large variations in pose and point of view, and low resolution imaging.

Recent advances in computer vision and particularly in object recognition suggest that new methods based on Deep Learning can improve the facial expression recognition task. Convolutional Neural Networks (CNNs) have represented a relevant breakthrough, especially since the last improvements on the ImageNet Challenge [28]. However, the amount of available data for this task is small, especially in all the possible configurations of pose, illumination and resolution. This supposes an inconvenience to exploit the training capacity of these networks, which need large amounts of training data. In this context the introduction of multi-task learning is particularly relevant, as it proved to successfully boost the performance of an individual task with the inclusion of other correlated tasks in the training process [20, 24, 47]. Thus, tasks with small amounts of data available can benefit from being trained simultaneously with other tasks, sharing a common feature representation and transferring knowledge between different domains. One of the main difficulties with multi-task approaches using different databases is the fact that not all the samples are labeled for all the tasks.

G. Pons and D.Masip  
Department of Computer Science, Universitat Oberta de Catalunya, Spain  
E-mail: {gponsro, dmasipr}@uoc.edu

In order to deal with this, classical approaches define different loss functions for each task and train alternatively for the different domains. This penalizes the tasks where sample labels are not available. One approach to solve this problem is a selective joint loss, but it only predicts probabilities for the label set to which the image belongs. Hence, in order to generate predictions for all the tasks, in this paper we propose a multi-label database-wise joint loss to overcome this problem.

This paper makes the following contributions: (i) we formalize a novel dataset-wise selective sigmoid cross-entropy loss function to simultaneously train a multi-task, multi-label and multi-domain model. (ii) We validate that this new proposal outperforms single task CNNs and the classical multi-task approach using different databases. (iii) We show that the results of the joint learning of an unlabeled task are coherently correlated to the labeled task in the case of the emotion recognition problem.

## 2 Related work

Automated emotion recognition methods from facial expression analysis in images use a subset of discrete basic emotions, as defined in [15]. Although the study of emotional states from faces is grounded in the late 19th century [12], Ekman and Friesen [15] defined a set of six basic emotions that are shared among all cultures, namely *happiness*, *surprise*, *anger*, *sadness*, *fear*, *disgust*. The production of these emotions in faces depends on specific facial muscle movements. The Facial Action Coding System (FACS) [16] defines a set of Action Units (AUs) that atomically group muscle movements. Each expression of emotion can be encoded as a combination of AU activations.

Facial expression recognition is a multidisciplinary research field, studied in machine learning, computer vision, cognitive science, psychology, neuroscience and applied health sciences. The number of computer vision researchers working in the field of facial expression analysis has increased since the early 90s, and a large amount of published works in the topic exist. Depending on the features used for the recognition task, we can distinguish two prevailing methodologies: geometric based approaches and appearance based approaches. In the first case, algorithms focus on localizing and tracking specific fiducial facial landmarks, in order to train a classifier based on distances and relative positions of these landmarks. Kotsia and Pitas [27] track a set of 119 key points to classify emotions from sequences, and Jeni et al. [25] apply a Procrustes transformation on

a reduced subset of 117 landmarks for emotion recognition. In appearance based emotion recognition, a set of features is extracted from pixel images to train a classifier. Classic examples are Gabor Wavelets [4], Histogram of Oriented Gradients (HOG) [29], and Local Binary Patterns (LBP) [46]. More complex models use Gaussian processes [17], hybrid methods using mid-level features [38] or patches from specific key locations [21]. More recently, deep learning methods have outperformed most state-of-the-art algorithms in the literature. Liu et al. [30] apply deep learning to a geometric model of facial regions for facial expression analysis, and Lu et al. [31] use Convolutional Neural Networks (CNN) on facial appearance. Recent models obtain improved results focusing on facial expression recognition *in the wild* by using committees of several CNN classifiers [26, 36]. Although most of the previous models focus on static images, facial expression analysis can benefit from temporal information. Cohen et al. [9] use Hidden Markov models on video sequences. Deep learning methods for emotion perception modeling in video include the use of CNNs for feature extraction and LSTM for learning the temporal dynamics [42]. Recent surveys on facial expression analysis can be found at [11, 33, 39].

Despite the current improvements, emotion recognition in the wild remains an open problem for the computer vision community. Results from the Emotion Recognition in the Wild (EmotiW 2015) reveal modest accuracies [13]. EmotionNet is the biggest Challenge in terms of data available, with more than 1 million images (2,000 labeled emotions and 950K unlabeled samples). The first edition of the challenge concluded that non-frontal faces still pose major difficulties to automated algorithms, and recognition rates decrease as a function of pitch and yaw rotations [5].

Deep learning methods benefit from large training datasets, but the labeling process of facial expression data requires expert FACs coders and it is a costly task. Therefore, most of the current available data sets are reduced and difficult to acquire. In this paper we conjecture that emotion recognition can be improved if we transfer knowledge from other related tasks. Particularly we focus on Multi-task learning (MTL). MTL was first studied by Caruana in [7], where he proposes to jointly learn parallel tasks sharing a common representation, and transferring part of the knowledge learned to solve one task to improve the learning of the other related tasks. Detailed analysis on the topic is carried out in some recent surveys [45, 34].

Ranjan et al. [37] propose a multi-task approach for face detection, landmark localization, pose estimation and gender recognition. This method exploits the benefits of multi-task to improve the performance of each

individual tasks. However, they use a single source of data, hence all the images must be labeled with all the tasks involved.

The closest work to this paper was published by Fourure et al. [19]. They propose a multi-task CNN for semantic segmentation of outdoor images in order to learn different tasks using images from different databases. They use a selective soft-max cross entropy with the objective of not penalizing the training of a task when feeding the model with images from another task. This approach though does not learn from similar labels. For instance, when training a database A with a class labeled *Grass*, and database B with a class labeled *Vegetation*, the method would predict the label corresponding to the database where the input image belonged to (if the image is from database A will predict it as *Grass*). In this paper we hypothesize that general accuracies will improve if we learn from both labels at the same time, and take full benefit of training related tasks using different non-homogeneous datasets.

### 3 Proposed approach

We propose a novel multi-label loss function that can be integrated into convolutional neural networks to improve the training of the emotion recognition task incorporating complementary tasks and data from different sources. This proposal can be extended to other multi-label problems, but in this article we focus on the multi-task training of emotion recognition and facial Action Units (AUs).

Emotion recognition is a problem where the model classifies the input image as one of the 7 emotion classes [15] (including the neutral class). Therefore, when addressing this task using a CNN, the computation of the output probabilities is performed by using soft-max function in the values of the last layer. This setting is not valid in our multi-task approach, since AU recognition is a multi-label problem. The classical setting for multi-label approaches is the use of the sigmoid function for the calculation of the output probabilities. However, using images from different sources and for different tasks implies having partially unlabeled data when performing the multi-task training, i.e. not all data is labeled with respect to all the tasks. Thus, this approach is not valid when dealing with images from different databases, since the sigmoid function penalizes the absence of labels in the related task. For example, optimizing the *Happy* class using samples with no AU annotations would penalize the AU6 and AU12 classes provoking a non-desired effect since these AUs together form the emotion *Happy*. We thus propose the *Selective Joint Multi-task* (SJMT) approach which defines a

novel dataset-wise selective sigmoid cross-entropy loss function to address multi-task, multi-label and multi-domain problems.

Given a set of images from  $k$  different datasets, where the label spaces are different, we define  $y_j^k$  the actual label of the  $j$ -th class in the space  $\mathcal{L}^k$ . The aim is to learn a nonlinear mapping presented as a CNN which minimizes a cross-entropy loss function for each individual sample. This function is defined as follows:

$$E(\hat{y}, y, k) = \frac{-1}{N} \sum_{j \in \mathcal{L}^k} [y_j \log \hat{y}_j + (1 - y_j) \log (1 - \hat{y}_j)] \quad (1)$$

where  $N$  is the number of classes,  $k$  is the number of different datasets, each defined on its own label set  $\mathcal{L}^k$ . The term  $y_j$  is the actual label of the  $j$ -th class, and  $\hat{y}_j$  is a sigmoid function defined as:

$$\hat{y}_j = \frac{1}{1 + e^{-\hat{p}_j}} \quad (2)$$

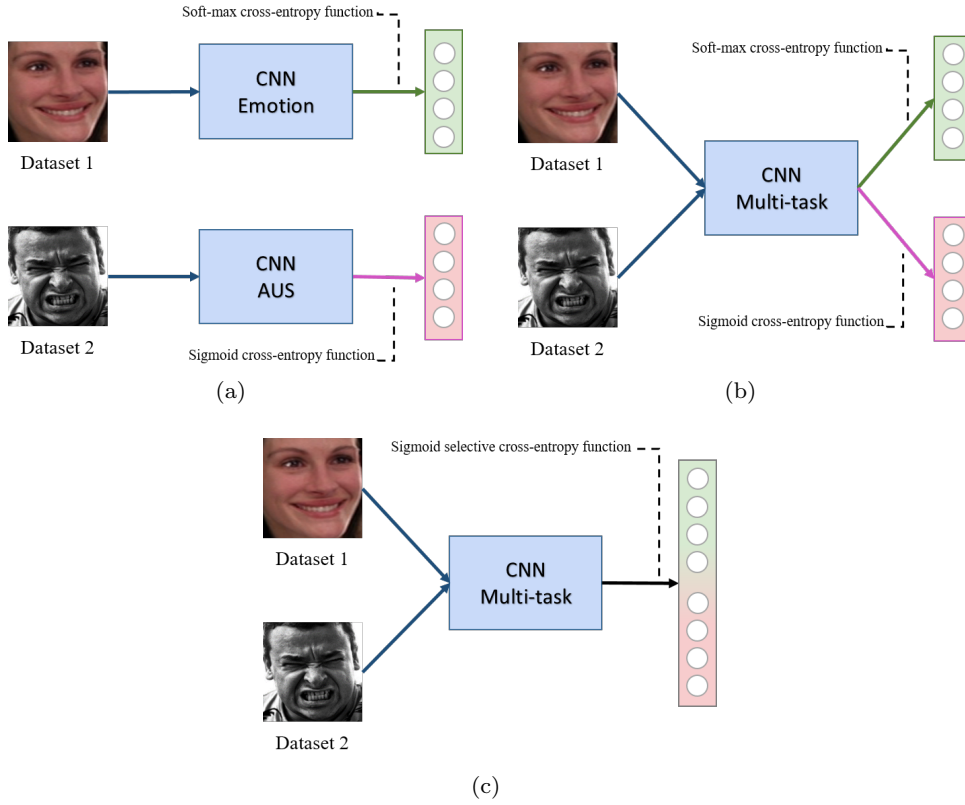
where  $\hat{p}_j$  is the output of the last layer of the CNN defined as  $f(Wh + b)$ ,  $f$  is the activation function of the layer,  $W$  and  $b$  the weights and bias of the layer and  $h$  the hidden representation of the last layer.

Figure 1 shows the different approaches compared in this article. First Figure 1(a) shows the approach where individual CNN are tailored to each task and database. Second Figure 1(b) shows the classical multi-task approach where the filters of the network are shared during training among the different tasks, taking into account that each task has an specific cross-entropy loss function and output layer, and the training is performed alternatively feeding the model with batches of images from each dataset. Finally, Figure 1(c) shows the approach proposed in this paper, where the whole network is shared among the different tasks and images from different databases are feed indistinctly thanks to the selective cross-entropy function defined in equation 1.

Even though this cross-entropy function could be incorporated in any CNN architecture, in this paper we used a ResNet [22] to implement the multi-task, multi-label and multi-database approach. Residual networks have shown great robustness in different fields [18, 35, 44] given their framework which eases the training of these networks, and enables them to be substantially deeper and, thus, obtain better performances than other networks.

### 4 Experimental results

In this paper we used the following databases for the experimental validation:



**Fig. 1** Different CNN strategies used in this paper. (a) shows the traditional approach with different CNNs for different tasks. (b) shows the classical multi-task approach with different output layers and loss functions for each individual task. (c) shows the proposed multi-task approach with a single output layer and the selective sigmoid cross entropy function.

- **SFEW2.0 database** [13], released for a competition in the 3rd Emotion Recognition In the Wild 2015 (EmotiW2015) challenge. The database was created by extracting frames from film clips with emotional content in order to obtain images in close-to-real world conditions. The images are labeled with 7 expressions (angry, disgust, fear, happy, sad, surprise, and neutral). The database consists of 958 images for training, 436 for validation, and 372 for testing. Since the labels of the test dataset are not provided, the validation set was used as test set for the experiments in this work. The database also provides the same data after being processed with a face alignment algorithm. We used this set in the experiments of this paper.
- **EmotioNet database** [6], released for the EmotioNet Challenge 2017 [5]. The database consists of 25,000 images of facial expressions with manual AU annotations. These annotations were given by expert human coders and cross referenced for verification of accuracy. From these images, a subset of 2,000 images were annotated with the 7 emotion expressions and 9 compound emotions made by the combination of two basic emotions. Since the test

dataset is private and was only available for participants of the challenge, the test set for the experiments in this work was generated randomly selecting 1,000 images (the same in all the experiments).

- **Extended Cohn-Kanade database (CK+)** [32], consists of 593 posed video sequences recorded from 123 university students ranging from 18 to 30 years old. The subjects were asked to express a series of facial displays including single or combined action units. The database also provides labels for the presence of emotions in these sequences. In this paper, we randomly generated a subset of 100 images for testing the methods. This database was acquired in controlled environments, and it has been used only to determine the parameters used for the neural networks. These parameters were consistent in all the experiments.

For ResNet [22] we used a training batch size of 128 images, and learning rate of 0.05 and exponential decay every 10,000 steps, for a total of 80,000 steps. Batch normalization and data augmentation by means of padding and random cropping images during training were used to prevent overfitting. Input data of size  $32 \times 32$  was used to pre-load CIFAR10 weights [23].

In the comparison among individual networks we also used VGG [40]. For this network we used a training batch size of 32 images, and an exponentially decayed learning rate of 0.00001 during 2,000 steps. To prevent the network from overfitting a dropout rate of 0.5 was used. In order to pre-load the trained weights used by the VGG team in the ILSVRC-2014 competition, input images of size  $224 \times 224$  were used.

All the networks trained in this paper were pre-trained using the Facial Expression Recognition (FER) 2013 database [1], which consists of 28,709 examples for training, 3,589 for validation and 3,589 for testing. All the experiments in this work were developed using TensorFlow [2] on NVIDIA GeForce GTX Titan GPU.

With regard to the detection of AUs and recognition of emotion categories, the evaluation criteria used was accuracy. Accuracy measures the number correctly classified examples, indicating whether our algorithm is able to discriminate between sample images with a certain AU/emotion present. In statistics, this difference between the measured and true value is generally called observational error, and it is defined as follows:

$$accuracy_i = \frac{TP_i + TN_i}{N} \quad (3)$$

where  $i$  specifies the class, i.e.,  $AU_i$  or the  $i$ -th emotion category,  $TP_i$  (true positives) are correctly identified test instances of class  $i$ ,  $TN_i$  (true negatives) are test images correctly labeled as not belonging to class  $i$ , and  $N$  is the total number of test images.

#### 4.1 Emotion and AU recognition with individual networks

Before validating the proposed approach in this paper we needed to validate the performance of each task trained with dedicated individual networks for such task and database. In addition, the baseline accuracy for each database is reported if known.

Thus, we trained specific networks for emotion recognition (see Figure 1(a)) with the SFEW and CK+ databases, and for AU detection with the EmotioNet and CK+ databases. Given their outstanding results in different fields, we decided to compare two of the most used architectures in image processing, VGG-16 and ResNet (with different sizes: 32 and 110). In this comparison we also added the results of other state-of-the-art commercial methods, such as the Microsoft Azure emotion recognition API [10] and the OpenFace framework for AU detection [3]. Note that these methods were not validated with the CK+ database, since we used this database only for defining the hyper-parameters

**Table 1** Results of different methods for the emotion and AU recognition tasks in percentage of accuracy.

Method	Emotion		AUs	
	SFEW	CK+	EmotioNet	CK+
Baseline	35.9%	-	-	94.5%
OpenFace	-	-	87.7%	-
Azure	33.8%	-	-	-
VGG-16	37.5%	93.0%	93.2%	95.0%
ResNet-32	37.8%	98.0%	93.2%	99.0%
ResNet-110	<b>41.3%</b>	<b>99.0%</b>	<b>93.9%</b>	<b>99.1%</b>

used in the networks we needed to train (VGG and ResNets).

Table 1 shows the accuracy obtained for each CNN for the different tasks and databases. Note that the images were not pre-processed specifically to enhance the performance in one database or another, and the same parameters were used in all the databases for a fair comparison. Hence, the most noticeable fact are the results obtained by ResNet-110, which are superior to the rest of the methods in the comparison.

#### 4.2 Multi-task learning for emotion and AU recognition

In this section we want to validate the hypothesis that the proposed SJMT method (see Figure 1(c)) outperforms the individual networks compared in the previous section as well as the classical multi-task approach (see Figure 1(b)). Given the results from the individual network comparison we decided to use the two best architectures in this comparison: ResNet 32 and 110.

For this experiment we jointly trained the multi-task approaches for emotion and AU recognition with the SFEW and EmotioNet databases respectively and the obtained results are summarized in Table 2. As occurred in the previous experiment, ResNet 110 generally obtained better results than ResNet 32. Comparing both the SJMT and classical multi-task approaches, the proposed method outperforms the classical approach in all the experiments except for AU recognition with ResNet 32. Finally, SJMT with ResNet 110 obtained the best results for both emotion and AU recognition, compared to the classical multi-task, and the single task approaches shown in Table 1

In order to analyze the benefits of using the proposed SJMT approach with related tasks, we focused on the predictions for AUs when feeding the model with images from the SFEW database, which only provides labels for emotion recognition. Thus, we grouped the images by the predicted emotion category and com-



**Table 2** Results of multi-task learning for the emotion and AU recognition tasks in percentage of accuracy. *Multi-task* refers to the classical approach (see Figure 1(b)) and *SJMT* refers to the *Selective Joint Multi-task* (SJMT) approach (see Figure 1(c))

Method	Network	SFEW (Emotion)	EmotioNet (AUs)
Multi-task	ResNet-32	39.4%	93.7%
	ResNet-110	40.3%	93.6%
SJMT	ResNet-32	40.6%	93.5%
	ResNet-110	<b>45.9%</b>	<b>93.9%</b>

	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU17	AU20	AU25	AU26
Angry	0.09	0.10	0.88	0.06	0.01	0.72	0.01	0.14	0.38	0.80	0.06
Disgust	0.73	0.00	0.66	0.00	0.41	0.00	0.05	0.00	0.00	0.01	0.00
Fear	0.21	0.16	0.06	0.91	0.16	0.01	0.02	0.02	0.13	0.98	0.43
Happy	0.00	0.05	0.06	0.04	0.82	0.12	0.93	0.00	0.01	0.98	0.10
Sad	0.05	0.03	0.08	0.11	0.37	0.21	0.41	0.69	0.04	0.06	0.03
Surprise	0.18	0.47	0.05	0.01	0.00	0.01	0.08	0.10	0.02	0.46	0.66
Neutral	0.00	0.00	0.01	0.06	0.01	0.01	0.00	0.01	0.01	0.20	0.00

**Fig. 2** Mean scores for AUs in SFEW images where only emotion labels are provided using the SJMT approach.

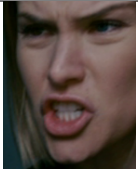

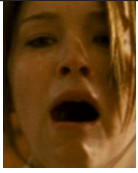

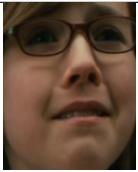
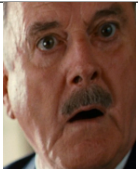
puted the mean scores for each AU as shown in Figure 2. This figure shows that when the model recognizes an emotion category, is simultaneously capable to identify the corresponding AUs in most of the cases. In order to simplify the interpretation of the figure, Table 3 shows the description of the AUs predicted by SJMT along with each emotion and a representative image from the test dataset of the SFEW database.

The table shows the coherence of the approach when recognizing the AUs in SFEW images. Given the fact that only 11 AU labels are used in the EmotioNet database (which is the database used to train the SJMT approach), some of the AUs theoretically forming the emotions cannot be recognized. A clear example is the Sad emotion, where the most discriminant AU in theory is 15 (Lip Corner Depressor) and the model recognizes it as 12 (Lip Corner Puller).

### 4.3 Compound emotion recognition

Most of the research on emotion recognition has focused on the study of the seven basic categories. We use the term “basic” to refer to the fact that such emotion categories cannot be decomposed into smaller semantic labels. However, it has been shown [14] that many

**Table 3** Description of the recognized AUs for each emotion.

Emotion	AU	Description	Example
Angry	4	Brow Lowerer	
	9	Nose Wrinkler	
	25	Lips Part	
Disgust	1	Inner Brow Raiser	
	4	Brow Lowerer	
	6	Cheek Raiser	
Fear	5	Upper Lid Raiser	
	25	Lips Part	
Happy	6	Cheek Raiser	
	12	Lip Corner Puller	
	25	Lips Part	
Sad	12	Lip Corner Puller	
	17	Chin Raiser	
Surprise	2	Outer Brow Raiser	
	25	Lips Part	
	26	Jaw Drop	

more facial expressions of emotion exist and are used regularly by humans. Compound emotions are those that can be constructed by combining basic component categories to create new ones. Hence, in this section we will validate our method on the extended problem of recognizing the following compound emotions from the EmotioNet database: *angrily disgusted*, *angrily surprised*, *fearfully angry*, *fearfully surprised*, *happily disgusted*, *happily surprised*, *sadly angry*, and *sadly disgusted*. The number of available samples is scarce for most of the classes, and we hypothesize that this emotion prediction task can take benefit from a joint learning process along with an auxiliary task such as AU recognition.

Following the experiments of the previous sections, here we compare the performance of the SJMT approach for compound emotion images with a single ded-

icated task network, in both cases training a ResNet 110. Given the unbalanced number of images per class (see Table 4), we decided to train the networks with a training set of 160 images consisting of 15 images in each class.

**Table 4** Results of single and multi-task learning for compound emotion recognition tasks in percentage of accuracy.

Compound emotion	Number images	Single ResNet-110	SJMT ResNet-110
Angrily disgusted	19	15.8%	42.1%
Angrily surprised	25	36.0%	64.0%
Fearfully angry	19	15.8%	63.1%
Fearfully surprised	17	58.8%	70.6%
Happily disgusted	486	47.6%	78.6%
Happily surprised	36	61.1%	72.2%
Sadly angry	15	13.3%	20.0%
Sadly disgusted	105	31.4%	64.7%
Mean compounds	-	34.9%	59.4%
All images	722	43.4%	73.1%

Table 4 shows that also when dealing with more complex emotions the SJMT network outperforms the single one, globally and for each of the classes. The improvement on the performance of the classes with the most examples (happily disgusted and sadly disgusted) also explains the large difference between methods when comparing globally, since much more images from the total are correctly classified. However, when comparing the mean of all the class accuracies we can see that SJMT with 59.4% considerably improves the single network performance (34.9%).

## 5 Conclusions

In this paper we presented the *Selective Joint Multi-task* approach which defines a novel dataset-wise selective sigmoid cross-entropy loss function to address multi-task, multi-label and multi-database problems. Specifically, we addressed this proposal to overcome one of the challenges with emotion recognition *in the wild*, which is the lack of large public labeled databases by including AU recognition in a multi-task approach.

Our proposed approach was assessed using the SFEW database for emotion recognition and the EmotionNet database for AU detection, and the results were compared to single task networks dedicated to each task individually and to the classical multi-task approach. SJMT obtained the best results in terms of accuracy for all the experiments. Results also showed the benefits of learning multiple correlated tasks simultaneously

by demonstrating visually that even for images without AU labels, the model is capable of inferring them in a coherent way. Finally, we evaluated the proposed method in a database of compound emotions, again obtaining improved results.

Future work will consist of studying and developing the inclusion of the landmark detection task in this multi-task approach. We believe that the addition of structural and geometrical cues to the appearance based CNNs might boost the performance of the emotion recognition task, but a new reformulation of the selective cross-entropy function needs to be done in order to handle such a problem in a joint training.

**Acknowledgements** This research was supported by TIN2015-66951-C2-2-R grant from the Spanish Ministry of Economy and Competitiveness and NVIDIA Hardware grant program.

## References

1. Facial Expression Recognition 2013 database. <https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-dataset> (2013)
2. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattemberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015). URL <http://tensorflow.org/>. Software available from tensorflow.org
3. Baltruaitis, T., Mahmoud, M., Robinson, P.: Cross-dataset learning and person-specific normalisation for automatic action unit detection. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 06, pp. 1–6 (2015)
4. Bartlett, M.S., Littlewort, G., Fasel, I., Movellan, J.R.: Real time face detection and facial expression recognition: Development and applications to human computer interaction. In: Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on, vol. 5, pp. 53–53. IEEE (2003)
5. Benitez-Quiroz, C.F., Srinivasan, R., Feng, Q., Wang, Y., Martinez, A.M.: Emotionet challenge: Recognition of facial expressions of emotion in the wild. arXiv preprint arXiv:1703.01210 (2017)
6. Benitez-Quiroz, C.F., Srinivasan, R., Martinez, A.M.: Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5562–5570 (2016)
7. Caruana, R.: Multitask learning. In: Learning to learn, pp. 95–133. Springer (1998)
8. Cockburn, J., Bartlett, M., Tanaka, J., Movellan, J., Pierce, M., Schultz, R.: Smilemaze: A tutoring system in real-time facial expression perception and production in

- children with autism spectrum disorder. In: ECAG 2008 Workshop Facial and Bodily Expressions for Control and Adaptation of Games, p. 3. Citeseer (2008)
9. Cohen, I., Sebe, N., Garg, A., Chen, L.S., Huang, T.S.: Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and image understanding* **91**(1-2), 160–187 (2003)
  10. Copeland, M., Soh, J., Puca, A., Manning, M., Gollob, D.: *Microsoft Azure: Planning, Deploying, and Managing Your Data Center in the Cloud*, 1st edn. Apress, Berkely, CA, USA (2015)
  11. Corneanu, C.A., Oliu, M., Cohn, J.F., Escalera, S.: Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (99), 1 (2016). DOI 10.1109/TPAMI.2016.2515606
  12. Darwin, C.: *The expression of emotion in animals and man*. London, England: Murray (1872)
  13. Dhall, A., Ramana Murthy, O., Goecke, R., Joshi, J., Gedeon, T.: Video and image based emotion recognition challenges in the wild: EmotiW 2015. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 423–426. ACM (2015)
  14. Du, S., Tao, Y., Martinez, A.M.: Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences* **111**(15), E1454–E1462 (2014)
  15. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. *Journal of personality and social psychology* **17**(2), 124 (1971)
  16. Ekman, P., Friesen, W.V.: *Facial action coding system: A technique for the measurement of facial movement*. CA: Consulting Psychologists Press. Ellsworth, PC, & Smith, CA (1988). From appraisal to emotion: Differences among unpleasant feelings. *Motivation and Emotion* **12**, 271–302 (1978)
  17. Eleftheriadis, S., Rudovic, O., Pantic, M.: Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE transactions on image processing* **24**(1), 189–204 (2015)
  18. Feichtenhofer, C., Pinz, A., Wildes, R.: Spatiotemporal residual networks for video action recognition. In: D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, R. Garnett (eds.) *Advances in Neural Information Processing Systems* 29, pp. 3468–3476. Curran Associates, Inc. (2016)
  19. Fourure, D., Emonet, R., Fromont, E., Muselet, D., Neverova, N., Trémeau, A., Wolf, C.: Multi-task, multi-domain learning: application to semantic segmentation and pose regression. *Neurocomputing* **251**, 68–80 (2017)
  20. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15*, pp. 1180–1189 (2015)
  21. Happy, S., Routray, A.: Automatic facial expression recognition using features of salient facial patches. *IEEE transactions on Affective Computing* **6**(1), 1–12 (2015)
  22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385* (2015)
  23. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. *arXiv preprint arXiv:1603.05027* (2016)
  24. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. *arXiv preprint 1503.02531* (2015)
  25. Jeni, L.A., Takacs, D., Lorincz, A.: High quality facial expression recognition in video streams using shape related information only. In: *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pp. 2168–2174. IEEE (2011)
  26. Kim, B.K., Lee, H., Roh, J., Lee, S.Y.: Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 427–434. ACM (2015)
  27. Kotsia, I., Pitas, I.: Facial expression recognition in image sequences using geometric deformation features and support vector machines. *Image Processing, IEEE Transactions on* **16**(1), 172–187 (2007)
  28. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, p. 2012
  29. Li, Z., Imai, J.i., Kaneko, M.: Facial-component-based bag of words and phog descriptor for facial expression recognition. In: *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, pp. 1353–1358. IEEE (2009)
  30. Liu, M., Li, S., Shan, S., Wang, R., Chen, X.: Deeply learning deformable facial action parts model for dynamic expression analysis. In: *Asian conference on computer vision*, pp. 143–157. Springer (2014)
  31. Lu, G.m., He, J.l., Yan, J.j., Li, H.: Convolutional neural network for facial expression recognition. *Journal of Nanjing University of Posts and Telecommunications* **36**(1), 16–22 (2016)
  32. 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: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 94–101 (2010)
  33. Martinez, B., Valstar, M.F., Jiang, B., Pantic, M.: Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing* (2017)
  34. Pan, S.J., Yang, Q.: A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **22**(10), 1345–1359 (2010)
  35. Pohlen, T., Hermans, A., Mathias, M., Leibe, B.: Full-resolution residual networks for semantic segmentation in street scenes. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3309–3318 (2017)
  36. Pons, G., Masip, D.: Supervised committee of convolutional neural networks in automated facial expression analysis. *IEEE Transactions on Affective Computing* (2017)
  37. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017)
  38. Sanchez-Mendoza, D., Masip, D., Lapedriza, A.: Emotion recognition from mid-level features. *Pattern Recognition Letters* **67**, 66–74 (2015)
  39. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence* **37**(6), 1113–1133 (2015)
  40. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556* (2014)



41. Soleymani, M., Pantic, M.: Emotionally aware tv. In: Proceedings of TVUX-2013: Workshop on Exploring and Enhancing the User Experience for TV at ACM CHI (2013)
42. Sun, B., Wei, Q., Li, L., Xu, Q., He, J., Yu, L.: Lstm for dynamic emotion and group emotion recognition in the wild. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 451–457. ACM (2016)
43. Whitehill, J., Serpell, Z., Lin, Y.C., Foster, A., Movellan, J.R.: The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing* **5**(1), 86–98 (2014)
44. Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A.: Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Transactions on Medical Imaging* **36**(4), 994–1004 (2017)
45. Zhang, Y., Yang, Q.: A survey on multi-task learning. *arXiv preprint arXiv:1707.08114* (2017)
46. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence* **29**(6), 915–928 (2007)
47. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2879–2886 (2012)