Large Margin Loss for Learning Facial Movements from Pseudo-Emotions

Andrei Racoviteanu
aracoviteanu@imag.pub.ro
Mihai Badea
mihai-sorin.badea@upb.ro
Corneliu Florea
corneliu.florea@upb.ro
Laura Florea
laura.florea@upb.ro
Constantin Vertan
constantin.vertan@upb.ro

Image Processing and Analysis Laboratory
University "Politehnica" of Bucharest
Bucharest, Romania

Abstract

In this paper we propose a large margin based loss function that enables information transfer from an unsupervised domain to a supervised one. The proposed methodology is applied in the context of face expression analysis. Categorical expressions are easier to understand and mutually exclusive, yet annotation is difficult and arguable. In contrast, facial movements encoded as action units have gained wider acceptance. Our strategy assumes self labelling images in the wild with pseudo-emotions to better learn action units. The proposed method is tested in two challenging scenarios with expressions in the wild, showing improved performance with respect to the baseline.

1 Introduction

Human-computer interaction has gained significant momentum during the last years. The human face carries important cues in inter-human interactions, thus being a powerful mean of communication. Facial expressions can include cues about the state of mind, motivation or intention of the subject. Aiming to mine this information, there has been a lot of research conducted in the area of automatic face expression recognition, both from psychologists and computer vision researchers.

The most popular system for analyzing expressions is called the Facial Action Coding System (FACS) and has been proposed by Ekman *et al*. [9]. The classification, based on face anatomy, measures the visible facial muscle movements in terms of so called Action Units (AUs). A facial expression is a combination of muscle movements with different intensities. In the expression analysis field, the FACS system has the benefit that its determination is more *objective*, as compared to *emotions*, which are a pure *subjective* aspect.

Following extensive research, Ekman *et al*. [9] found evidence that supports the universality of six basic facial expressions linked to six discrete emotions (happiness, sadness, fear,

anger, surprise and disgust). Yet the expressions may be genuine, as result of emotion, or posed, where the subject tries to replicate a set of movements that corresponds to what is agreed to be the pose for a specific emotion.

However, even with such a simple classification, it was shown that the human annotation in the case of facial expression is hard. One person needs more than 100 hours of training for recognizing action units with a decent accuracy [1]. The limit to get a FACS certification for annotation is 70% [8], but one would expect that a non-trained annotator will achieve less. For comparison, on CIFAR-10, an untrained user can go higher than 90% accuracy.

This is a reason why datasets of images acquired in the wild with face expressions of genuine emotions do not usually contain the manual annotation on emotions or on AUs annotations. The manually annotated datasets are small, acquired in laboratory conditions and with simulated expressions usually at maximum intensity (apex). Bigger, "in the wild" datasets with expressions have been proposed, but they are labelled either completely by automatic, non-ideal solutions, or partially labelled by non-experts. Both cases lead to limited sets of labels and, thus, encourage methods to go beyond pure supervision and to use, additional, unlabelled data.

If one aims to take advantage of the benefits of the deep learning techniques in the problem of face analysis, it needs large annotated datasets. Since images with faces in expressions are easy to be found on the Internet, but labelling is difficult for genuine emotions, semi-supervised and transfer learning methods can be used to improve the performance of the supervised learning algorithms by using the available, unlabelled data. Challenges in analyzing faces refer to the high inter-class similarities and, respectively, the high intra-class variations. Due to these challenges, one needs to construct efficient loss functions in order to achieve not only separable, but discriminative features that will ensure that the faces will be classified based on expressions and not based on facial appearance or looks.

In this paper, we claim the following contributions: $(\cdot)$ First, we propose a new loss function that combines a classical loss computed on the last, decisional, layer with a large margin loss computed on a relevant embedding taken from a lower layer. $(\cdot)$ Since the large margin assumes clustering the data, it may be naturally extended into a semi–supervised or domain transfer strategy using a self-label solution. We define a mechanism to use it for action units detection, in which case we impose classes in the form of pseudo-emotions.

2 Related Work

The proposed method uses a clustering loss into a domain adaptation method (i.e. an extension of semi–supervised learning) to the problem of action units detection firstly and, in subsidiary, to that of expression recognition. We review recent works in each direction.

Loss function for better deep features discrimination. In conjunction to deep learning, several different types of loss function were proposed in the last years and were shown to work on the problem of face recognition: Wen *et al.* [25] proposed a loss function, called *center loss*, to minimize the intra-class distances between the deep features; Liu *et al.* [19] learned angularly discriminative features with the angular softmax loss in order to achieve smaller maximal intra-class distance than minimal inter-class distance; Zhang *et al.* [28] developed a loss function for long tailed distributions; Zheng *et al.* [34] showed that normalizing the deep features with the so-called Ring Loss leads to improved accuracy. All these methods were shown to give good results on face recognition tasks, where very large annotated datasets like MegaFace are available.

The same problem is found in expression recognition tasks, but coupled with smaller annotated databases. In this direction, Cai *et al*. [3] added a new term to the center loss for imposing arge inter-class distances and tested it on facial expressions with promising results.

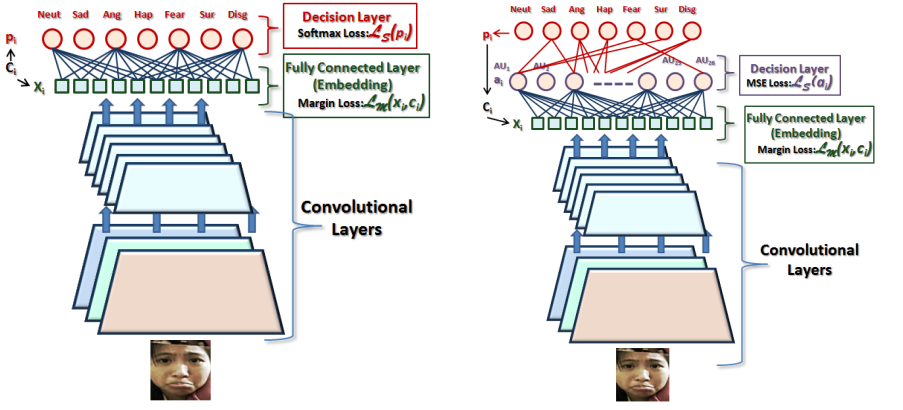
Domain Adaptation and Semi-Supervised Learning. When data from two domains are available, the concept of domain transfer or domain adaptation appeared as an alternative to the increase of the amount of information over which a learner may be trained directly in order to improve its prediction capabilities. Many previous solutions and alternatives have been introduced and we refer the reader to the recent review by Wang *et al*. [24]. Shortly, the domain transfer is feasible and the resulting learner has improved performance if one domain is adapted to the other, so to ensure the transfer.

If the two domains have the same distribution of the input data, but one has no labels, the problem is of Semi-Supervised Learning (SSL). Some SSL techniques were reported to be very effective on standard benchmarks such as CIFAR10/100, MNIST, SHVN, ImageNet. A simple solution, derived from the concept of self training, uses the model trained on the labelled part of the data to label the unlabelled part and to further propagate the so called Pseudo-Labels [16]. Another approach uses an association between the labelled and unlabelled data by means of nearest neighbor to retrieve a better embedding for a deep learning solution [11]. Thus, in the context of deep learning, the unsupervised part may influence either the decision, or a previous, but relevant, layer.

Face Expression Recognition. A lot of effort was put in research for automatic face expression recognition. Most of the proposed methods classify the expressions in the six/seven prototypical classes proposed by Ekman *et al*.: happy, sad, surprise, fear, anger, disgust (and contempt). Representative methods can be found in several surveys of the domain as [21], or [4]. Lately, the proposed solutions rely on deep learning [15, 26, 32, 33], the challenge being the recognition of genuine expressions on images acquired in the wild. For such tasks, an example of dataset with in the wild images is Real-world Affective Faces Database (RAF-DB) [17, 18] which contains a little over 15000 images annotated via crowdsourcing. Due to the challenge of manual annotations, transfer learning or semi-supervised learning stood at the base of some proposed solutions. Du *et al*. introduced a semi-supervised multiview deep generative framework for emotion recognition based on variational autoencoders [6]. Zhang *et al*. [29] proposed a so-called enhanced collaborative SSL in order to assess the performance degradation problem connected to SSL.

Action Units Detection and Intensity Estimation. Since there is controversy among psychologists regarding the connection between expression and emotions, some stating that the theory of the six basic emotions is too simple, many researches have been focused on the action unit (AU) detection and, more recently, on the AU intensity estimation. Consequently, datasets with images in the wild with AU annotations like EmotioNet [10] were introduced. Yet manual annotation is hard and it is available only for a small subset, with the remainder usable for learning into a context of semi–supervised or domain transfer.

Overall, in the recent years many researches used deep architectures to explore such datasets. AU detection with deep learning was done by Zhao *et al*. [30] and Corneanu *et al*. [5]. For AU intensity estimation most of the methods rely on supervised learning. Kaltwang *et al*. [12] proposed using a latent tree model for learning the intensities of different AUs. Benitez-Quiroz *et al*. [2] used deep nets, trained with a global loss in a supervised manner to recognize AUs in the wild. A combination of conditional random fields and copula functions was proposed by Walecki *et al*. [23]. Variational autoencoders were used by Tran *et al*. [22] for the same task. Recently, spectral clustering was used to structure unlabelled data followed by supervised classification in a reduced set [31]. Thus far, due



Learning expressions Learning AU from pseudo emotions

Figure 1: The schematic of the two versions of the proposed method.

to the difficulty of the problem and limited availability of annotated databases, there is left much place for improvement with respect to the reported performance.

3 Large Margin Loss for Information Transfer

From a technical point of view, we propose a methodology to train a deep network in a semi–supervised manner in a classification problem with mutually exclusive categories. In this scenario, we ask the network to include a layer that may act as an embedding or a feature descriptor. The unlabelled part of the dataset contributes to finding good features (in the sense that they provide large margins between the classes clouds), while the last (classification) layer is developed solely by the supervised part. This method is suitable for recognizing one of the basic expressions, under the assumption that they are mutually exclusive.

In the second step, we extend the method to address action units. The problem of detecting/estimating the intensity of AU is one where the classes are not mutually exclusive, thus the previous definition of large margin does not hold. In this sense, we will use a set relations in which the AUs aggregate in pseudo-emotions. The two embodiments of the proposed method may be followed in figure 1.

3.1 Framework for Semi-supervised Learning

Large decision margin in supervised learning. For a mutually exclusive class problem, Wen *et al*. [25] introduced the center loss that explicitly reduces the intra-class variations by pushing embedding samples towards their corresponding class centers in the feature space (embeddings) during training. The centers are updated in each iteration using Stochastic Gradient Descent (SGD). If the embedding is $\mathbf{x}_i$, with y_i the label of the same instance, the

center loss is:

$$\mathcal{L}_C = \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{c}^j\|_2; \quad \mathbf{c}^j = \frac{\sum_{i=1}^{N} \alpha_i^j \mathbf{x}_i}{\sum_{i=1}^{N} \alpha_i^j}; \quad \alpha_i^j = \begin{cases} 1 & , y_i = j \\ 0 & , y_i \neq j \end{cases} \tag{1}$$

where $\mathbf{x}_i$ describes an instance of class j ($y^i = j$), $\mathbf{c}^j$ is the centroid of the same class j and α_i^j shows the membership of the data i to class j. Further extensions [3, 19] of this method sought ways to enforce also large distances between class centroids using the cosine distance. One problem with theseextensions is that if there is no explicit intervention over the other class centroids, the contribution of the "large margin" idea is limited as the network may set centroids to fixed positions and simply scale down data. An alternative way is to use Euclidean L_2 distances over normalized embeddings:

$$\mathcal{L}_{\mathcal{M}} = \sum_{i=1}^{N} \left(\left\| \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2} - \frac{\mathbf{c}^j}{\|\mathbf{c}^j\|_2} \right\|_2 - \frac{1}{C-1} \sum_{k=1,k\neq j}^{C} \left\| \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2} - \frac{\mathbf{c}^k}{\|\mathbf{c}^k\|_2} \right\|_2 \right) \tag{2}$$

Denoting the normalized vector ($\hat{\mathbf{x}}_i = \frac{\mathbf{x}}{\|\mathbf{x}\|}$), the loss can be rewritten as:

$$\mathcal{L}_{\mathcal{M}} = \sum_{i=1}^{N} \left(\left\| \hat{\mathbf{x}}_i - \hat{\mathbf{c}}_j \right\|_2 - \frac{1}{C-1} \sum_{k=1,k\neq j}^{C} \|\hat{\mathbf{x}}_i - \hat{\mathbf{c}}_k\|_2 \right) \tag{3}$$

where C is the number of classes. Eq. (3), specifically, imposes that one instance of normalized data should be close to its class center and far from the other centers thus enforcing a large margin on the embedding. Its behavior is illustrated in figure 2(a). The normalization of the data ensures that the loss $\mathcal{L}_{\mathcal{M}}$ is bounded and in the implementation numerical instability is prevented.

Overall, the network is trained using the loss computed as a weighted sum:

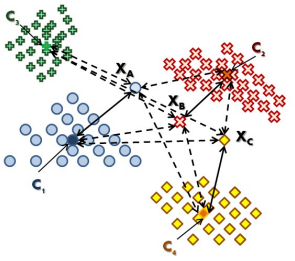
$$\mathcal{L} = \lambda_1 \mathcal{L}_{\mathcal{S}} + \lambda_2 \mathcal{L}_{\mathcal{M}} \tag{4}$$

where λ_1, λ_2 are weighting constants, $\mathcal{L}_{\mathcal{S}}$ is the decisional loss which may be either cross entropy for classification, or mean square error for regression. In the backward propagation one needs to compute the derivative of the loss with respect to the current d- element of the D-dimensional embedding as:

$$\frac{\partial \mathcal{L}_{\mathcal{M}}}{\partial x_d} = \left(2(\hat{\mathbf{x}}_i - \hat{\mathbf{c}}_j) - \frac{2}{C-1} \sum_{k=1,k\neq j}^{C} (\hat{\mathbf{x}}_i - \hat{\mathbf{c}}_k) \right) \cdot \frac{\partial \hat{\mathbf{x}}_i}{\partial x_d}; \quad \frac{\partial \hat{\mathbf{x}}_i}{\partial x_d} = \frac{1 - \hat{\mathbf{x}}_i^2}{\|\mathbf{x}\|_2} \tag{5}$$

Semi–supervised learning A solution for the usage of unlabelled data in the context of deep learning is inspired by Pseudo-Labels [16]. There, the predictor itself infers pseudo-labels of unlabelled examples by seeking maximum confidence and the network is trained with the difference between the actual confidence and the maximum confidence.

In our case, as the network predicts some label for any data, it will contribute to its centroid update. Compared to the purely supervised training of a centroid based strategy, where the centroid is computed based on the annotated labels, here the centroid is computed based on *self–predicted labels*. We should emphasize again that this embodiment requires mutually exclusive classes.



(a)

$$
\begin{aligned}
AU_6 + AU_{12} &= \text{Happiness} \\
AU_1 + AU_4 + AU_{15} &= \text{Sadness} \\
AU_1 + AU_2 + AU_{5B} + AU_{26} &= \text{Surprise} \\
AU_1 + AU_2 + AU_4 + AU_5 + AU_7 + AU_{20} + AU_{26} &= \text{Fear} \\
AU_4 + AU_5 + AU_7 + AU_{23} &= \text{Anger} \\
AU_9 + AU_{15} + AU_{16} &= \text{Disgust} \\
AU_{R12A} + AU_{R14A} &= \text{Contempt}
\end{aligned}
$$

(b)

Figure 2: **(a)** The behavior of the large margin loss: Given four classed (as set of points), each will have its centroid. Given new points $(\mathbf{x}_A, \mathbf{x}_B, \mathbf{x}_C)$ which are labelled by the network to be in the classes 1,2,4, the feature–space will be computed such that the distance to the corresponding class (with continuous arrows) to decrease, while distances to other centroids (depicted with dashed line) to increase. **(b)** Connection between action units and expressions, as proposed by Ekman *et al*. [8]. In the EmotioNet only some AUs are available and we did not use Contempt.

3.2 Pseudo Emotions

Given the two systems for describing the human face, namely into categorical expressions and, respectively, with AUs, there were multiple efforts to construct a set of formulas that would equal each of the fundamental expressions with a (weighted) sum of AUs. Ekman *et al*. [8] introduced a set of such formulas, which is presented in figure 2(b). Yet these formulas are not unanimously accepted and other versions have been found relevant too [7, 20]. Thus, one may conclude that there is not a single universal set of formulas and any such set only provides a partial matching between basic expressions and action units.

Using one set of formulas, one will enforce the AUs to match a set of unique face expressions. However, as there is no definitive psychological evidence that these expressions do match the universal emotions, we will call them pseudo-emotions. In the given equations, to ensure balanced emotion intensity, one should normalize the expressions with the number of contributing AUs.

3.3 Exploiting features for emotions into AU tasks

The problem of using a deep network to detect or estimate the intensity of action units existing in a facial image may be treated as multi-class, non-exclusive, classification or as vectorial regression problem. In such a case, the definition of centroid as the average of embeddings of data having that class becomes unclear since the "class" is not defined. Nevertheless, we may use pseudo-emotions as means to define the centroids.

Formally, given a data set and its embeddings $\mathbf{x}_i$, these are transformed by the network in the final layer, where the outputs $\mathbf{a}_i$ are computed. A set of formulas (e.g. here the pseudo-emotions – figure 2 (b)) are used to transition from the outputs, $\mathbf{a}_i$, to the pseudo–class probabilities, $\mathbf{p}_i$:

$$\mathbf{p}_i^k = f_k(\mathbf{a}_1, \dots \mathbf{a}_M), \quad k = 1 \dots C \tag{6}$$

The pseudo–class is defined as the maximum confidence argument: $u_i^{pred} = \text{argmax}_k \mathbf{p}_i^k$. In this case, the network is trained with the following combined loss:

$$\mathcal{L} = \lambda_3 \mathcal{L}_{\mathcal{S}} + \lambda_4 \mathcal{L}_{\mathcal{M}}; \quad \mathcal{L}_{\mathcal{M}} = \sum_{i=1}^{N} \left(\left\| \hat{\mathbf{x}}_i - \hat{\mathbf{c}}_j \right\|_2 - \frac{1}{C-1} \sum_{k=1, k \neq j}^{C} \left\| \hat{\mathbf{x}}_i - \hat{\mathbf{c}}_k \right\|_2 \right);$$
$$\mathbf{c}^j = \frac{\sum_{i=1}^{N} \alpha_i^j \mathbf{x}_i}{\sum_{i=1}^{N} \alpha_i^j}; \quad \alpha_i^j = \begin{cases} 1 & , u_i^{pred} = j \\ 0 & , u_i^{pred} \neq j \end{cases} \tag{7}$$

In the supervised problem of *AU* recognition, $\mathcal{L}_{\mathcal{S}}$ is the mean square error between the vector of annotated intensity AU over a face image and the network prediction. In the two domains problem, for unlabelled data, $\mathcal{L}_{\mathcal{S}} = 0$ since we have no class definition and thus the softmax is not relevant. λ_3, λ_4 are weighting constants used to balance the magnitude of the losses; one strategy is to decrease λ_4 over iterations since the good embeddings were formed in the earlier stages.

3.4 Generality

While we have exemplified the method for the specific task of action units (facial movements) estimation, the method is general as long as the following conditions are met: (1) There is a subject–unlabelled domain that may be structured by self training via mutually exclusive classes recognition. (2) The target domain, which requires a different structure, can be expressed by some logical association with respect to the reference domain (i.e. eq. (6) exists). The association is exploited to learn a relevant set of embeddings that are further easily expanded into the decision layer of the target domain by a relevant loss function.

4 Results

4.1 Databases and Scenarios

We will consider two scenarios: recognizing the six basic expressions on static images and detecting action units. Both are performed on images in the wild. Example of images may be followed in figure 3.

For the task of recognizing the basic facial expressions we will use the Real-world Affective Face Database (RAF-DB) [18], [17] as the labelled part of the data and the first subset from MegaFace benchmark [13] as the unlabelled part of the data, as they both contain images with faces in the wild. RAF-DB contains 15349 color facial images, of large resolution and is divided in 12271 images for training and 3078 images for testing. Each image is labelled with one of the seven basic emotions via crowdsourcing by at least 40 trained annotators. The first subset from MegaFace dataset contains approximately 311000 facial images that do not contain labels regarding expressions.

For the task of action units detection we used the EmotioNet database [10]. This dataset contains 1M images collected from the Internet, 50000 being annotated with binary labels of multiple AUs. From these 25000 images are used as the test/train partition as suggested in the original paper that introduced the dataset [10]. We consider them the labelled part of the data. The unlabelled part of the data consists on 400000 images from the remainder of the same database.

Figure 3: Examples from the two databases used for test: the top row shows expressions from RAF-DB, while the bottom row illustrates AUs from EmotioNet.

4.2 Implementation

The images were preprocessed as follows: the face was detected by MTCNN [27]; the cropped face was re–scaled at 227×227; color normalization was applied on each plane. The method was implemented on TensorFlow and library defaults were used (training from scratch, weight sparsity regularization, etc.). Network update was performed using Adam. λ_1 and λ_2 were selected such that $\lambda_1 \mathcal{L}_{\mathcal{S}} \approx 0.8 \cdot \lambda_2 \mathcal{L}_{\mathcal{M}}$; the same choice for λ_3 and λ_4. We have focussed on the AlexNet [14] architecture and for a fair comparison, from prior solutions, we refer to works that have used this architecture; there are works that have relied on much larger networks, overall the performance may get higher, but often that is the merit of the architecture.

Regarding the training process, while the purely supervised version converged in at most 150 epochs, for the semi-supervised/domain transfer version up to 350 epochs were needed. In particular the convergence of the centroid towards a stable position was slow and kindly see figure 4 for an illustration of this behavior.

4.3 Recognizing Face Expressions

The performance on the RAF-DB database may be followed in table 1. On this database two type of accuracies are typically computed [17]: the overall accuracy denoted by Acc. and the average of the diagonal of the confusion matrix, denoted by Avg. Acc. One may notice that compared to the baseline AlexNet trained solely on the supervised part with soft-max, the improvement in either accuracy is at least 10%. When reported to an AlexNet trained purely in the supervised manner with a center/island loss strategy, the improvement smaller, but non-negligible.

The large margin loss proposed provides better performance, in a semi–supervised framework, than the previously introduced center loss [25] and the island loss [3]. Our performance is close to the supervised solution [32], yet one should also notice, that given the weights of the layer, they performed additional optimization to select the layer better correlated with the aimed output.

Table 1: Accuracy obtained on the classification problem on the RAF-DB database. Prior work: Feat.Sel.Net - feature selection network [32], Our proposal uses Large Margin (LM).

Method	Framework	Avg. Acc.	Acc.
AlexNet - [17]	Superv	55.60	68.90
AlexNet + Feat.Sel.Net [32]	Superv	72.46	81.10
AlexNet + Island loss [3]	Superv	57.1	75.08
AlexNet + Center loss [25]	SSL	63.15	78.81
AlexNet + Island loss [3]	SSL	64.53	78.81
AlexNet + LM loss	SSL	67.26	79.85

Table 2: F1 score [%] while detecting action units on the EmotioNet database. The framework (FW) is either supervised (Sv), semi-supervised (SSL) or transfer (T). "Avg small" is the average over the reduced set of AU:{1,4,5,6,12,25,26}, Avg full is over the entire set.

Method	FW	AU_1	AU_2	AU_4	AU_5	AU_6	AU_9	AU_{12}	AU_{17}	AU_{20}	AU_{25}	AU_{26}	AU_{43}	Avg. small	Avg. full
AlexNet [31]	Sv	24.2	n/a	34.7	**39.5**	73.1	n/a	86.8	n/a	n/a	88.5	45.6	n/a	56.1	n/a
AlexNet cen. loss [25]	Sv	**34.4**	30.3	55.3	33.3	69.10	**46.1**	79.3	27.8	**32.3**	84.4	43.2	48.8	57.9	48.8
AlexNet +WSC [31]	SSL	25.3	n/a	34.5	39.3	**75.6**	n/a	**87.4**	n/a	n/a	**88.8**	47.4	n/a	57.0	n/a
AlexNet + Isl. loss [3]	T.	30.4	29.5	**56.7**	30.6	66.7	44.1	77.3	26.7	23.8	83.9	47.3	43.9	56.14	46.7
AlexNet + LM loss	T.	34.1	**31.1**	56.6	33.9	71.0	45.1	78.1	**30.9**	25.3	83.8	**50.9**	47.2	**58.33**	**49.0**

4.4 Detecting AUs

Performance on the EmotioNet database may be followed in table 2. The standard measure used for evaluation is F1 score [2] defined for action unit AU_i as follows:

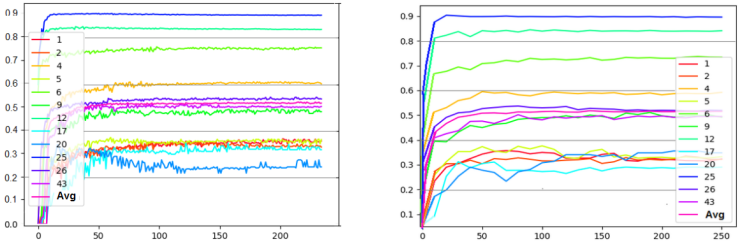
$$F1_i = 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{8}$$

where $Precision_i$ is the fraction of the annotations of AU_i that are correctly recognized (i.e., number of correct recognitions of AU_i divided by the number of images with detected AU_i), and $Recall_i$ is the number of correct recognitions of AU_i over the actual number of images with AU_i.

For comparison we recall the performance of [31] from both purely supervised and their proposal of semi-supervised (entitled Weakly Supervised Clustering - WSC). Since they report only a subset of the AUs present on EmotioNet, we report two averages, one on the small set and one on the full set. Also we report the performance when the standard island loss and center loss are used. Again the main difference is that we ensure that embeddings are normalized and thus the margin is relative to the data.

To give an upper boundary of the peformance, we also note that training for detection one AlexNet for each AU provides slightly better values, but with the cost of dramatically increased memory and time [2]. As one may see, the proposed method manages to reach the best results when a unique AlexNet is used to estimate simultaneously all AUs. Compared to standard AlexNet, we improve the F1 score by more than 2%, which is also the performance gained by the use of deep learning methods with respect to the classical approach [31].

The behavior of the method while training in the supervised and respectively while doing



Supervised Learning AU from pseudo emotions

Figure 4: The behavior while training. While doing transfer, the convergence is slower but more robust.

transfer scenarios may be followed in figure 4.

5 Discussion

In this paper we proposed a method for analysis of facial expressions using data from two domains: one labelled and one unlabelled. We also introduced a strategy to use mutually exclusive self-predicts of the trained network to explore the unsupervised domain and enforced useful embeddings as intermediated by the product of the training. We showed that it is possible to extend the strategy from the mutually exclusive classification to a multi-label regression problem by adding a layer of transition between the two domains. As long as the equations implementing the transition were logical, the learned embeddings were useful for the analysis of the labelled domain.

The proposed method originated in the usage of center loss function as a way to enforce a relevant layer beyond the decisional one. It is in a series of methods that ask simultaneously for both compact clusters and large margins between different clusters. Since the clustering is an unsupervised approach, it may be extended to unlabelled domain. We took things one step further and we proved that the strategy may be successfully applied to a domain transfer problem. Our results showed consistently improved performance with respect to strong supervised baselines.

Acknowledgment. This work was partially supported by the Romanian Ministry of Innovation and Research, UEFISCDI, project SPIA-VA, agreement 2SOL/2017, grant PN-III-P2-2.1-SOL-2016-02-0002. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

References

[1] M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Measuring facial expressions by computer image analysis. *Psychophysiology*, 36(2):253–263, 1999.

[2] F. Benitez-Quiroz, Y. Wang, and A. Martinez. Recognition of action units in the wild with deep nets and a new global-local loss. In *ICCV*, pages 3990–3999. IEEE, 2017.

[3] J. Cai, Z. Meng, A.S. Khan, Z. Li, J. O'Reilly, and Y. Tong. Island loss for learning discriminative features in facial expression recognition. In *FG*, pages 302–309, 2018.

[4] C. Corneanu, M. Oliu Simón, J. Cohn, and S. Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE T. PAMI*, 38(8):1548–1568, 2016.

[5] C. Corneanu, M. Madadi, and S. Escalera. Deep structure inference network for facial action unit recognition. In *ECCV*, 2018.

[6] C. Du, C. Du, H. Wang, J. Li, W.L. Zheng, B.L. Lu, and H. He. Semi-supervised deep generative modelling of incomplete multi-modality emotional data. In *ACM MM*, pages 108–116, 2018.

[7] S. Du, Y. Tao, and A. Martinez. Compound facial expressions of emotion. *Proc. of the Nat. Academy of Sciences*, 111(15):E1454–E1462, 2014.

[8] P. Ekman and E.L. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the FACS*. Oxford Scholarship, 2005.

[9] P. Ekman, W. Friesen, and J. Hager. *Facial action coding system: Research nexus*. Network Research Information, Salt Lake City, 2010.

[10] C F. Benitez-Quiroz, R. Srinivasan, and A. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *CVPR*, pages 5562–5570, 2016.

[11] P. Haeusser, A. Mordvintsev, and D. Cremers. Learning by association-a versatile semi-supervised training method for neural networks. In *CVPR*, pages 89–98, 2017.

[12] S. Kaltwang, S. Todorovic, and M. Pantic. Latent trees for estimating intensity of facial action units. In *CVPR*, pages 296–304, 2015.

[13] I. Kemelmacher-Shlizerman, S. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, pages 4873–4882, 2016.

[14] A. Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014.

[15] C.-M. Kuo, S.-H. Lai, and M. Sarkis. A compact deep learning model for robust facial expression recognition. In *CVPRW*, pages 2121–2129, 2018.

[16] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshops*, 2013.

[17] S. Li and W. Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE TIP*, 28(1):356–370, 2019.

[18] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, pages 2852–2861, 2017.

[19] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, pages 212–220, 2017.

[20] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPR Workshops*, pages 94–101, 2010.

[21] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE T. PAMI*, 37(6):1113–1133, 2015.

[22] D. Tran, R. Walecki, O. Rudovic, S. Eleftheriadis, B. Schuller, and M. Pantic. Deepcoder: Semi-parametric variational autoencoders for automatic facial action coding. In *ICCV*, pages 3209–3218, 2017.

[23] R. Walecki, O. Rudovic, V. Pavlovic, B. Schuller, and M. Pantic. Deep structured learning for facial action unit intensity estimation. In *CVPR*, pages 5709–5718, 2017.

[24] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[25] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *Computer Vision – ECCV 2016*, pages 499–515, 2016.

[26] Z. Yu and C. Zhang. Image based static facial expression recognition with multiple deep network learning. In *ACMI*, pages 435–442, 2015.

[27] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Proce. Letters*, 23(10):1499–1503, 2016.

[28] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao. Range loss for deep face recognition with long-tailed training data. In *ICCV*, pages 5419–5428, 2017.

[29] Z. Zhang, J. Han, J. Deng, X. Xu, F. Ringeval, and B. Schuller. Leveraging unlabeled data for emotion recognition with enhanced collaborative semi-supervised learning. *IEEE Access*, 6:22196–22209, 2018.

[30] K. Zhao, W. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. In *CVPR*, pages 3391–3399, 2016.

[31] K. Zhao, W. Chu, and A. Martinez. Learning facial action units from web images with scalable weakly supervised clustering. In *CVPR*, pages 2090–2099, 2018.

[32] S. Zhao, H. Cai, H. Liu, J. Zhang, and S. Chen. Feature selection mechanism in cnnsfor facial expression recognition. In *BMVC*, 2018.

[33] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan. Peak-piloted deep network for facial expression recognition. In *ECCV*, pages 425–442, 2016.

[34] Y. Zheng, D. Pal, and M Savvides. Ring loss: Convex feature normalization for face recognition. In *CVPR*, pages 5089–5097, 2018.