Annealed Label Transfer for Face Expression Recognition

Corneliu Florea
corneliu.florea@upb.ro
Laura Florea
laura.florea@upb.ro
Mihai Badea
mihai.badea@upb.ro
Constantin Vertan
constantin.vertan@upb.ro
Andrei Racoviţeanu
andrei.racoviteanu@upb.ro

Image Processing and Analysis Laboratory (LAPI), University Politehnica of Bucharest
Splaiul Independenţei 313, Bucharest, Romania

Abstract

In this paper we propose a method for recognizing facial expressions using information from a pair of domains: one has labelled data and one with unlabelled data. As the two domains may differ in distribution, we depart from the traditional semi–supervised framework towards a transfer learning approach. In our method, which we call Annealed Label Transfer, the deep learner explores and predicts labels on the unsupervised part, yet, in order to prevent too much confidence in its predictions (as domains are not identical), the global error is regularized with a randomization input via an annealing process. The method's evaluation is carried out on a set of four scenarios. The first two are standard benchmarks with expression faces in the wild, while the latter two have been little attempted before: face expression recognition in children and the study of the separability of anxiety-originated expressions in the wild. In all cases we show the superiority of the proposed method with respect to the strong baselines.

1 Introduction

The development of man-machine interaction emphasizes the need for computer vision methods able to accurately estimate the expression of a face in a given image. Many practical applications have been developed in this direction. We refer the reader to the recent reviews [6, 26] on the topic for an exhaustive presentation. From the multiple directions previously approached, we address the expression categorization into one of the basic sets defined by Ekman *et al*. [7]: "neutral", "anger", "fear", "disgust", "happy","sad", "surprise"; sometimes "contempt" is included. In a later test, we have also included "anxiety".

While approaching the face expression recognition, one particular aspect needs particular emphasis that is human annotation of such data is hard and costly. For a reference let us recall that the average untrained user achieves $\approx 94\%$ accuracy for image classes on CIFAR

10 [11]. In contrast, for face expression, Susskind et al. [28] showed that an experienced observer (psychology student) reached 89.2% in a 6 expression experiment; Bartlett et al. [3] and Ekman et al. [7] noted that at least 100 hours of training are needed for a person in order to get 70% accuracy in recognizing face movements (the lower limit to obtain FACS certification). Thus, due to the difficulty in annotating images, problems related to face expression analysis welcome methods and strategies that use additional unlabelled data as a substitute to more annotations, in order to augment the performance.

Contribution. In this paper we investigate strategies relying on additional unlabelled data to improve deep learning baselines in several problems of face expression recognition. As the additional data, although being similar to the labelled one, brings no guarantee of being identical, the proposed method is on the boundary between semi–supervised learning and inductive transfer learning.

In this paper we contribute by: (1) a new method for domain transfer named Annealed Label Transfer (ALT); while initial tests are in a semi–supervised framework, the later two use slightly different data distributions, and fall rather into the transfer leaning category. (2) We report systematic analysis of recognition of face expressions in children and show that transfer learning can be used to enhance performance in a problem with scarce data. (3) We report performance for recognizing the expression of anxiety in images in the wild.

Paper structure. The remainder of the paper is organized as follows: Section 2 summarizes main contributions in related directions; Section 3 describe the method from a technical point of view; Section 4 presents experimental evaluation, while the paper ends with a discussion on the achieved results.

2 Related work

While the main technical contribution is related to regularization in form of injection of randomization by annealing as a way to control the transfer of information from unlabelled data to the labelled set, the theme is face expression recognition. In the next paragraphs we review the main related works.

Random strategies for backward update in deep learning. In the latest period multiple deep learning works started to explore strategies based on injection of randomness as a way to regularize the flow of information. We refer here to techniques that use random inputs to compute the next update in the weights of a network; thus we go beyond weight shake or drop–out layers. Neftici *et al.* [22] proposed learning (in a supervised manner) with a strategy incorporating randomization adapted to each subset of parameters. Blier *et al.* [4] proposed the ALRAO strategy, which uses completely randomized learning rates (at each step) for supervised learning; it reports performance matching the standard SGD, but without prior adaptation of the range. Concurrently to this work, Jackson *et al.* [12] proposed to use randomization for the gradient direction in a semi–supervised strategy: the direction of the gradient is arbitrarily adjusted, but the method lacks a check that the advance is beneficial, although it reports improved performance on semi–supervised CIFAR10.

Face expression recognition. In the later years face expression recognition, in most of the forms and scenarios, has been dominated by deep learning methods.

Face expression recognition in the wild. In general, deep learning techniques [17, 29, 34, 35] train a single instance or an ensemble of deep networks and adapt the prediction onto a single independent image or onto a sequence. For instance, expression recognition on static images has been addressed [2] with the specific aim of mobile phone consumer applications;

there, a carefully engineered CNN-based method sought to maximize the performance on a single database, thus the method being prone to over-fitting. Multiple databases, and thus better generalization, are envisaged in a series of methods that augment the baseline performance by the usage of a modified center loss [19, 20], or of a mechanism for feature selection [34]. On occasions, multiple data are fused and the network was trained over the ensemble, fused set [30].

In the later years, the power of semisupervised learning or of the domain transfer has been exploited in face expression recognition too. Zhang et al. [33] uses a strategy that re-evaluates self labels predictions over unlabelled date at each iteration. Zeng et al. [32] used s self labelling strategy based on bottom-up propagation in a relational graph.

Face expression in children databases. While research on automatic expression recognition in adults is mature, expression recognition in children has been less studied, with few results being published only in the last period. A limitation is imposed by the reduced number of annotated databases with children, which is a consequence of our need of protecting them from malevolent intentions. The largest database available is The Child Affective Facial Expression (CAFE) [21]; yet the database was introduced in the psychology domain and little investigation from computer vision community took place. Baker *et al.* [1] used a combination of SVM and features to recognize the child expression on CAFE. On other sets, Nojavanasghari *et al.* [23] introduced a new multi–modal database and tried several feature+classifier methods on it. Very recently Khan *et al.* [15] introduced a new database with children and reported automatically obtained results. Yet, results on the largest and most prominent database, CAFE, have not been reported

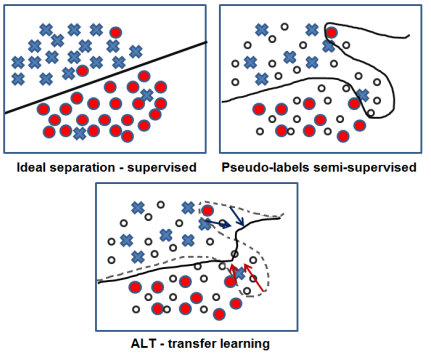
Recognizing Anxiety In emotion analysis, while on some occasions the anxiety is viewed as a subcategory of fear, there are arguments for its inclusion in its own class. Perkins *et al.* [24] performed a psychological experiment and concluded that a large group of viewers did distinguish between the expressions of anxiety and that of fear.

In terms of automatic recognition, Carneiro *et al.* [5] investigated multiple automatic ways to detect stressed persons, yet the multi–modal data require the temporal dimensions of videos. Giannakakis *et al.* [8] built a laboratory induced set of images with faces of stressed persons and reported $\approx 88\%$ accuracy based on a processing chain that included face region description with the location of keypoints given by Active Appearance Models, optical flow and K-Nearest Neighbor; the last is, to our best knowledge, the only work reporting automated recognition of expressions of stress in images. Still, the mentioned works did not include images in the wild, but only images recorded in laboratory setup.

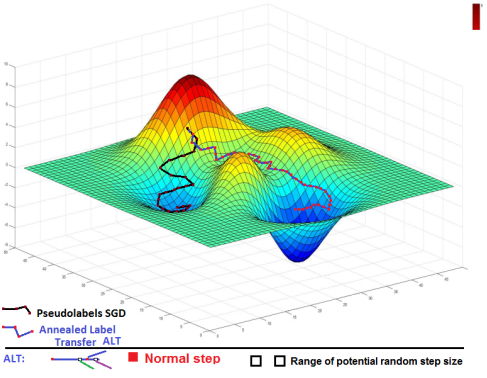
3 Method

In this paper we address a dual database framework. As learner we will use a classical deep net architecture such as AlexNet [16] or VGG-16 [27]. First, the learner transfer the knowledge (labels) from the supervised domain to the unsupervised one. Second, the learner aims to structure the unsupervised domain seeking more confidence in its predictions. However, since the two domain may not be identical, the second transfer is regularized by injection of randomization by means of an annealing process.

Formulation. In our construction, the two kinds of data are the labelled set $\{\mathcal{X}^l, \mathcal{Y}^l\} = \{(\mathbf{x}_i^l, \mathbf{y}_i^l)\}_{i=1}^N \overset{\text{iid}}{\approx} p(\mathcal{X}^l, \mathcal{Y}^l)$ and the unlabelled $\{\mathcal{X}^u\} = \{\mathbf{x}_j^u\}_{j=N+1}^{N+M} \overset{\text{iid}}{\approx} p(\mathcal{X}^u)$. The learned predictor is $f : X \to Y$, $f \in \mathcal{F}$ where $\mathcal{F}$ – hypothesis space.



(a)



(b)

Figure 1: (a) Data classes aggregation over labelled data (blue cross and red dots) and respectively unlabelled data (black dots). ALT, due to the random input, tries to randomly modify some weaker boundaries, while standard Pseudo–Labels enforce existing ones. (b) Behavior of the gradient descent: the randomization with check from the ALT algorithm assumes a random value in a given range and may find better descents.

With respect to the two databases, if $p(\mathcal{X}^l)$ is identical with $p(\mathcal{X}^u)$, the framework is that of semi–supervised learning. If $p(\mathcal{X}^l)$ and $p(\mathcal{X}^u)$ are overlapping, yet partially different, the framework is that of transfer learning. In many practical scenarios, if two distinct databases are used, one may not easily evaluate how well they match.

One solution to use the unlabelled data is provided by Lee [18] who relied on the predictor itself to infer pseudo-labels of unlabelled examples, by choosing the class with most confidence. The method is derived from entropy minimization [10] by forcing the learner to have more confident predictions on the unlabelled data.

The major limitation of methods from this category lies in the fact that it heavily builds upon the hypothesis that $p(\mathcal{X}^l)$ and $p(\mathcal{X}^u)$ are identical, with that latter being more populated in some parts and thus defining clearer borders. Yet, in cases where this assumption is not true, the use of the predictor to pseudo-label the unlabelled data may hurt the performance. To address the problem we propose the following method derived from annealing.

The supervised problem can be solved by:

$$f_\theta(\mathbf{x}) = \underset{\theta}{\operatorname{argmin}} S(\theta) = \underset{\theta}{\operatorname{argmin}} \Big(\mathcal{L}(\mathbf{y}^l; \mathbf{x}^l, \theta) + R(\theta) \Big) \tag{1}$$

where $R(\theta)$ is a regularizer, implemented in this work as the standard L_2 regularization (i.e. $R(\theta) = \alpha \sum_\theta \|\theta\|^2$). $\mathcal{L}(\cdot)$ is the loss function implemented as the cross-entropy.

Seeking a solution for a dual data set problem, one may solve:

$$f_\theta(\mathbf{x}) = \underset{\theta}{\operatorname{argmin}} S(\theta) = \underset{\theta}{\operatorname{argmin}} \left(R(\theta) + \mathcal{L}(\mathbf{y}^l; \mathbf{x}^l, \theta) + \frac{1}{M} \sum_{j=N+1}^{N+M} \sum_{m=1}^{C} L(y_m^j; f_{\theta t}^m(\mathbf{x}_j)) \right) \tag{2}$$

where C is the number of classes, while $L(y_m^j; f_{\theta t}^m(\mathbf{x}_j))$ is the loss over unlabelled data. The solution followed by this paper is summarized as Algorithm 1. A graphical interpretation

Input: Labelled inputs $\mathbf{x}_i^l$, labels y_i^l. Unlabelled inputs $\mathbf{x}_j^u$. Validation set $\mathbf{x}_k^v$ and Labels y_k^v.
Initialize: net weights θ_i randomly. Initialize: Unlabelled inputs predictions w_j^u randomly. Consider: Loss function $\mathcal{L}(y_b;\mathbf{x}_b,\theta)$. Initialize: temperature $T = T_0$.
for *epoch=1:*N_{ep}*:* **do**
  for *b=1:*N_{batch}*:* **do**
    Pass the labelled batch b: $(\mathbf{x}_b^l; y_b^l)$:
    a. Find predicts $y_{pred} = f_{\theta_b}(\mathbf{x}_b^l)$;
    b. Compute loss function $\mathcal{L}\left((y_b^l - y_{pred}); \mathbf{x}_b^l; \theta_b\right).$;
    c. Compute gradient $g^l := \nabla\mathcal{L}(y_b^l - y_{pred})$;
    d. Update net parameters $\theta_b = \theta_{b-1} + g_l(\theta_{b-1})$ using SGD ;
    e. Evaluate net update on the validation set $\mathcal{L}_v = \mathcal{L}(y^v; \mathbf{x}^v, \theta_b)$;
    Pass an unlabelled batch b: $\mathbf{x}_b^u$
    a. Find predicts $\mathbf{y}_{pred} = f_{\theta_b}(\mathbf{x}_b^u)$;
    b. Determine high confidence predicts $y_{pred}^u = \operatorname{argmax} \mathbf{y}_{pred}$;
    c. Compute loss $\mathcal{L}\left((y_{pred}^u - y_{pred}); \mathbf{x}_b^l; \theta_b\right) + T \cdot \xi$;
    d. Compute gradient $g^u := \nabla\mathcal{L}(y_{pred}^u - y_{pred})$;
    e. Attempt update on net parameters $\theta_{b2} = \theta_b + g_u(\theta_{b-1})$ using SGD ;
    f. Evaluate net update on the validation set $\mathcal{L}_{v2} = \mathcal{L}(y^v; \mathbf{x}^v, \theta_{b2})$;
    g. If the the update is potentially positive (the loss decreases) $(\mathcal{L}_v - \mathcal{L}_{v2}) > \beta(T) > 0$ keep it: $\theta_b = \theta_{b2}$;
  end
  Cooldown: Decrease the temperature $T = T - \Delta T$. Also $\beta(T - \Delta T) < \beta(T)$
end

Algorithm 1: Annealed Label Transfer algorithm. T_0 is chosen to be 0.2 of the main loss function, while ξ is a uniform random value in [-1,1]. β is a linear function with temperature : $\beta(T) = 0.75 \cdot T$.

for the dual domain approach may be found in figure 1(a), while for the gradient descent in subfigure (b).

The Algorithm 1 introduces a randomization which has a continuously decreasing maximum amplitude. The temperature update is chosen such that temperature becomes 0 before the end of training: $\Delta T = \frac{T_0}{0.75 \cdot N_{ep}}$. In such a manner, the convergence is controlled and is guided by the convergence of the SGD.

Relation with other works. While similar ideas are currently introduced, we would to emphasize some difference with respect to other work that use randomization as a regularization of the learning. The main difference w.r.t work of Neftici *et al*. [22] is that they design a strategy for each neuron/weight, while in our case the randomization is unitary, allowing the network to behave consistently. W.r.t the ALRAO work of Blier *et al*. [2] while the idea is similar, we bounded and cooled down, regressing at the later steps to the standard SGD and we use it as semisupervised/transfer learning. Closer to our method is the work of Jackson *et al*. [12] that also design a semi–supervised strategy, yet their arbitrary angle selection due to the lack checkout, may lead to quick convergence in near local minima.

4 Implementation and results

For implementation we rely on standard architectures: AlexNet and VGG-16 [16, 27] that included batch normalization. The training was done with SGD with learning rate preset to $10^{\{-3,-4-5\}}$ for 50 epochs to a total of 150. The implementation is done in Pytorch. The faces were preprocessed by cropping based on the landmarks provided by the Dlib[1] that implments the solution from [13].

4.1 Fundamental expressions of faces in the wild.

Here, we have experimented with images from RAF-DB and FER+ databases.

RAF-DB [19, 20] contains facial color images in the wild, which are, often, large enough such that cropped faces require downsizing to 224×224. The database is annotated by at least 40 trained annotators per image and divided into 12271 training images and 3078 testing images. It is labelled for seven basic emotions. On the RAF-DB database, prior works reported standard accuracy and the average of the main diagonal of confusion matrix, denoted as average accuracy.

FER+ is derived from FER2013 [9] and contains 28709 training images, 3589 validation (public test) and another 3589 (private) test images, in the wild. FER+ images have 48×48 pixels, are gray-scale and contain only the face. Barsoum et al.[2] noted the high noise in original labels and performed some "cleaning", by removing the images missing faces and providing labels by aggregating the opinion of 10 non-specialist annotators. While FER+ is more reliable, given the quality of images, we report also results on FER2013. Compared to RAF-DB, the images are small, gray and have been annotated less rigorously.

The unlabelled data for both experiments is a subset of the MegaFace database [14], containing $\approx$ 311.000 images with faces randomly selected from Internet. The MegaFace images contain a face in the wild that has an expression, but there is no information about it. Examples from the three datasets are shown in Figure 2.

As said, all databases (RAF-DB, FER+/FER2013 and MegaFace) contain images randomly acquired from the Internet, thus without obvious bias between the datasets. The framework is potentially one of semi–supervised learning.

Results achieved on the expression recognition are presented in table 1 subset (a) while experimenting on normal color images (RAF-DB) and in Table 1 (b) while using small gray images (FER2013/+). First, to establish a baseline, we report the performance of several architectures trained solely on the labelled data (supervised). We also cite previously published, carefully engineered prior art methods [17, 19, 30, 34] and respectively [2, 29, 34] tested on the same databases.

In both cases, our ALT strategy improves with respect to the use of purely Pseudo-Labels method, which serves as reference; also there is improvement with respect to the ALRAO [4] strategy for selecting learning rate, thus arguing for the power of ALT method.

On the FER 2013 database, the inconsistency of the annotations previously noted [2] leads to overall worse results when compared to the FER+ version. In these cases good performance is achieved by solutions [2, 31] that focus on a single database. However the proposed semisupervised method manages to improve the performance until it reaches the top value for the better annotated version, namely FER+.

[1] Available at dlib.net.

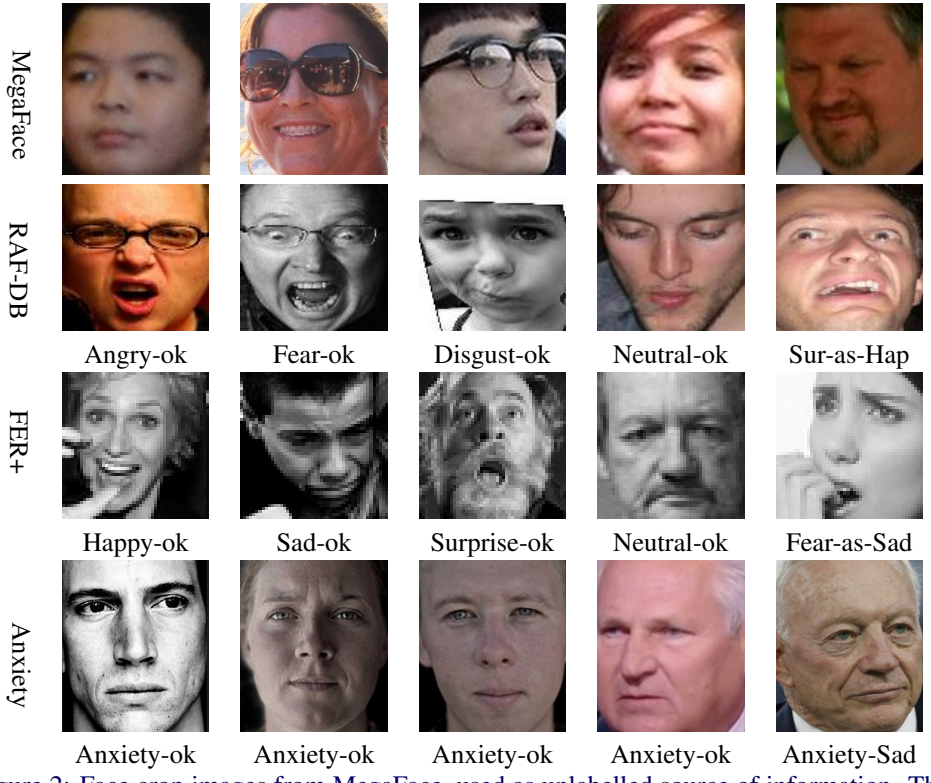


Figure 2: Face crop images from MegaFace, used as unlabelled source of information. The following rows contain images from the RAF-DB database (second row), FER+ and, respectively images with stress/anxiety; first four columns are correct recognition examples while the last contains an error. .

On the RAF-DB database, we manage to improve considerably over previously reported purely supervised solutions: for Alexnet, our semi–supervised version improves with $\approx 20\%$ in both Avg. Acc. and overall Acc w.r.t [19]; for VGG-16 the improvement is similar, given [19]; w.r.t to more recent works [17, 34], the improvement is smaller, nevertheless it exists and our solution reports state of the art performance.

4.2 Recognition of Anxiety

We have collected from Internet 176 images with anxiety and, respectively, 42 images of former combatants formally diagnosed with Post Traumatic Stress Disorder (PTSD). For the first part, the process implied searching on Google after images with various keywords ("stress", "PTSD", "anxiety", "worry") alone, or in association with professions that have high incidence of stress such as "(sport) coach", "solicitor", "farmer" etc. As images are scarce, the search was in almost all major languages supported by Google Translate. It resulted in a set of ≈ 500 images. Following the process from RAF-DB, the images were then filtered by consecutive visual inspections from different expert viewers (included in the cropped form) and only the ones that passed all observers were kept. The set was divided into 120 examples in train, 20 in validation and 36+42 in test. The latter set specifically included images of combatants showing PTSD. This set was merged with RAF-DB and we

Table 1: (a) Performance within 7-class problem on the RAF-DB database. FSN - feature selection network, FSM - frame-to-sequence method, PL - standard pseudo-labelling, ALRAO - pure random policy for the learning rate (All Learning Rates At Once) [4], ALT (Annealing Label Transfer) marks our proposal. With bold we marked the best result, while with italic our best proposal. (b) Recognition rates (accuracy) within 8-class problem on the FER+ and 7-class on the FER2013 database.

	Method / Metric	Avg. Acc.	Acc.
SUPERV	AlexNet [15]	55.60	68.90
	VGG-16 [15]	58.22	70.53
	DLP-CNN [15]	74.20	84.13
	ResNet-18 [30]	–	80.00
	FSM [12]	65.52	72.21
	FSN [34]	72.46	81.10
TRANSFER	AlexNet + PL	61.8	73.5
	AlexNet + PL + ALRAO	66.5	73.5
	AlexNet + ALT	72.3	81.50
	VGG–16 + PL	74.6	83.25
	VGG–16 + PL + ALRAO	72.8	81.25
	VGG–16 + ALT	**76.5**	**84.5**

(a) **RAF-DB** results

	Method	FER+	FER2013
SUPERV	AlexNet [34]	–	61.1
	AlexNet	78.08	68.2
	FSN [34]	n/a	67.6
	VGG – Majority voting [0]	83.85	–
	VGG – Probab. label [0]	84.99	–
	FUS [29]	67.03	–
TRANSFER	AlexNet + PL + Maj. vot	80.05	69.12
	AlexNet + PL + ALRAO	80.6	68.65
	AlexNet + ALT	82.38	69.62
	VGG–16 + PL	84.35	69.27
	VGG–16 + PL + ALRAO	82.15	69.27
	VGG–16 + ALT	**85.2**	*69.85*

(b) **FER+** / FER2013 results

have experimented within a 8 class problem.

Again, the MegaFace database is the source of unlabelled data. While this collection is huge and all kinds of expression may be included, we (manually) have not found any image to express stress or anxiety. Thus the scenario is one of domain transfer.

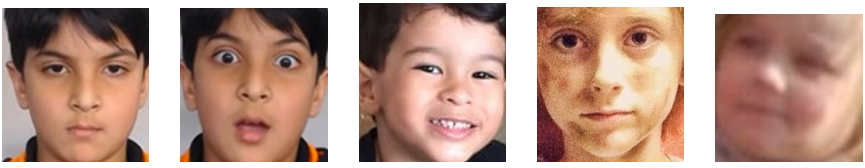
The results are in table 2. As one notices, the use of multiple domains improves the accuracy of recognizing stress, while the top performance is achieved by the ALT method.

4.3 Face Expression in children

Following the summary from recent work of Khan *et al.* [15] the largest database with children expression available is CAFE [21][2]. CAFE database contains images of children between two and eight years old. The set features 90 female children and 64 male children who were asked to pose for each of the 7 standard expressions. Since not all children were able to successfully pose for all expressions, it resulted in a set of 1192 photographs that are annotated by specialists. We have randomly divided the database in 45% for train and test each and 10% for validation using a person–wise scheme.

As unlabelled data we have selected 1389 images from the LIRIS database [15], containing 7 faces of children and all images of children from the IMDB-WIKI database [25], build for age estimation. With respect to the age, we note that based on annotations, initially were taken ≈ 3000 images of children having between 1 and 10 years; yet, after manual validation only 1154 were kept. In this scenario, while CAFE and LIRIS have both been acquired in laboratory, LIRIS contains older children, while IMDB-WIKI subset is in the wild. As there is a clear difference between the domains of the labelled (CAFE) and unlabelled (LIRIS + IMDB-WIKI) data, the framework is no longer that of semi-supervised learning.

[2] As per copyright agreement, images from the CAFE database cannot be displayed in articles. Readers are kindly asked to visit the home page: https://www.childstudycenter-rutgers.com/the-child-affective-facial-expression-se.



(a) (b) (c) (d) (e)

Figure 3: Examples of images with children: (a,b) images from LIRIS database (used as unlabelled), (c) image similar with the ones from CAFE database, (d,e) images from IMDB-WIKI database (used as unlabelled). .

Table 2: Performance (recognition rate) within 8-class problem on images of Anxiety/stress added to Raf-DB.

Method	AlexNet superv	AlexNet + PL	AlexNet + ALT	VGG-16 superv	VGG-16 + PL	VGG-16 + ALT
Anxiety Recog.	44.87	52.56	**56.41**	38.46	44.87	48.17
Overall Recog.	70.53	73.5	80.08	75.87	83.21	**84.55**

Due to the large variability of the database, the direct training and testing on the database (no matter the train/test ratios, regularization or data augmentation) resulted in a blunt over–fitting: 100% accuracy on the train and random chance on test. To improve generalization, annotated images from the FER+ database were included. Performance saturated at about 5000 added images. Thus the training set will, in fact, contain, CAFE and FER+ images. Example of images with children are given in figure 3.

As mentioned before, only Baker *et al.* [1] reported peer–reviewed, automatic results on this database using features and SVM. It trained only on CAFE images snd the result reported was obtained using 1000 images in train, but selected randomly, without ensuring person separation. To the best of our knowledge, no report using deep networks exists.

The obtained results are presented in table 3. We report the prior work's results [1], yet this was easily outperformed by any solution based on deep learning. In this case, to establish a baseline, we also report the performance of the AlexNet when it has been trained in a purely supervised manner. One may notice that using additional data helps and the strategy proposed leads to best solution, which is very close to be near perfect.

5 Discussion

While bearing usefulness in many practical applications, face expression recognition and related tasks also carry the burden of the difficulty of manual labelling, thus being an almost ideal candidate for the use of transfer learning. In many practical scenarios, users aims to produce the highest possible accuracy on some test benchmark, while having available a limited set of annotated images and many other images that lack annotations. While consistency between the two datasets is aimed, in some occasions it is hard to actually verify it (as it is the case of semi–supervised learning in RAF-DB and FER+), while in other (children expression) it is merely impossible.

Our proposal, named Annealed Label Transfer (**ALT**), showed improved results compared to the baseline in all 4 scenarios and reached top performance on the standard RAF-

Table 3: Performance (recognition rate) within 7-class problem on the CAFE Database. Result for [1] is inferred from a plot.

Method	SVM -based [1]	AlexNet - superv	AlexNet + PL	AlexNet + ALT
	62.5	83.50	90.29	**99.29**

DB database. In cases where the amount of labelled data is large enough (such as standard RAF-DB and FER+ databases), the improvement brought by transfer learning techniques is minimal or in some cases even negative. Yet, the more cautions strategy of incorporating randomization injection in the update showed that even in such cases we manage to **improve** the performance.

In this paper, we extended the exploration of the face expression recognition topic by introducing new directions in the form of recognition of anxiety/stress faces and respectively faces of children. In the first case, while debated in the psychological literature, the expression of anxiety was found by Perkins *et al*. to be different from the one of fear; our experiments in a scenario with 8 classes found a good separation of this expression too. The transfer strategy managed to bring significant improvement w.r.t to the baseline by incorporating new information.

In the case of expressions in children faces, given the here reported deep learning baselines, the use of additional information lead to improvement of up to 16%. A possible explanation lies in the scarcity of the annotated data and in the fact that the CAFE database contains expression at apex. Thus different children shows very different expressions and the additional information and regularization based on randomization lead to better generalization.

Acknowledgment. This work was partially supported by the 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] L. Baker, V. LoBue, E. Bonawitz, and P. Shafto. Towards automated classification of emotional facial expressions. In *Proc. Annual Conf. of the Cognitive Science Society*, pages 1574 – 1579, 2017.

[2] E. Barsoum, C. Zhang, C. Ferrer, and Z. Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ICMI*, pages 279 – 283, 2016.

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

[4] L. Blier, P. Wolinski, and Y. Ollivier. Learning with random learning rates. *CoRR*, abs/1810.01322, 2018.

[5] D. Carneiro, J. Carlos-Castillo, P. Novais, A. Fernandez-Caballero, and J. Neves. Multimodal behavioral analysis for non-invasive stress detection. *Expert Systems with Applications*, 39(18):13376 – 13389, 2012.

[6] C. Corneanu, M. Simón, J. Cohn, and S. Escalera. 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.

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

[8] G. Giannakakis, M. Pediaditis, D. Manousos, E. Kazantzaki, F. Chiarugi, P.G. Simos, K. Marias, and M. Tsiknakis. Stress and anxiety detection using facial cues from videos. *Biomedical Signal Processing and Control*, 31:89 – 101, 2017.

[9] I. Goodfellow, D. Erhan, P. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, et al. Challenges in representation learning: A report on three machine learning contests. In *ICONIP*, pages 117 – 124, 2013.

[10] Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In *NIPS*, pages 529 – 536, 2005.

[11] T. Ho-Phuoc. CIFAR10 to compare visual recognition performance between deep neural networks and humans. *CoRR*, abs/1811.07270, 2018.

[12] J. Jackson and John Schulman. Semi-supervised learning by label gradient alignment. *CoRR*, abs/1902.02336, 2019.

[13] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, pages 1867 – 1874, 2014.

[14] 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.

[15] R. A. Khan, A. Crenn, A. Meyer, and S. Bouakaz. A novel database of children's spontaneous facial expressions (liris-cse). *Imag. Vis. Computing*, 83–84:61 – 69, 2019.

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

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

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

[19] Shan Li and W. Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Trans. on Image Processing*, 28 (1):356 – 370, 2019.

[20] W. Li, F. Abtahi, and Z. Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *CVPR*, pages 6766 – 6775, 2017.

[21] V. LoBue and C. Thrasher. The child affective facial expression (CAFE) set: Validity and reliability from untrained adults. *Frontiers in Psychology*, 5:1532, 2015.

[22] E. Neftci, C. Augustine, S. Paul, and G. Detorakis. Event-driven random back-propagation: Enabling neuromorphic deep learning machines. *Frontiers in Neuroscience*, 11:324, 2017.

[23] B. Nojavanasghari, T. Baltrušaitis, C. Hughes, and L.-P. Morency. Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In *ACMI*, pages 137 – 144, 2016.

[24] A. Perkins, S. Inchley-Mort, A. Pickering, P. Corr, and A. Burgess. A facial expression for anxiety. *J. of Personality and Social Psychology*, 102(5):910, 2012.

[25] R. Rothe, R. Timofte, and L. Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *IJCV*, pages 144 – 157, 2016.

[26] 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.

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[28] J.M. Susskind, G. Littlewort, M. Bartlett, J. Movellan, and A. Anderson. Human and computer recognition of facial expressions of emotion. *Neuropsychologia*, 45(1):152 – 162, 2007.

[29] E. Tran, M. B. Mayhew, H. Kim, P. Karande, and A. D. Kaplan. Facial expression recognition using a large out-of-context dataset. In *WACV*, pages 52 – 59, 2018.

[30] V. Vielzeuf, C. Kervadec, S. Pateux, A. Lechervy, and F. Jurie. An Occam's razor view on learning audiovisual emotion recognition with small training sets. In *ICMI*, pages 589 – 593, 2018.

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

[32] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *ECCV*, pages 222–237, 2018.

[33] 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.

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

[35] 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.