Curriculum Learning for Facial Expression Recognition

Liangke Gui, Tadas Baltrušaitis, and Louis-Philippe Morency
Language Technologies Institute, School of Computer Science,
Carnegie Mellon University, USA

***Abstract*— Over the past few years, there has been an increased interest in machine understanding and recognition of affective states based on facial expressions. While great progress has been made, there are still a lot of challenges facing automatic emotion recognition, namely: generalizability of models across datasets, accounting for individual differences, and recognition of subtle expressions. While deep learning techniques enabled a large amount of progress in many areas of computer vision, this progress has not yet been fully translated to emotion recognition. Our work attempts to partly address that by presenting a novel learning technique for deep learning methods that leads to better generalization for emotion recognition from facial expressions.**

I. INTRODUCTION

Over the past few years, there has been an increasing interest in machine understanding and recognition of affective and cognitive mental states, especially based on facial expression analysis [28]. As the facial expression is considered the main channel of nonverbal communication, automatic facial expression analysis is used in a number of applications to facilitate human computer interaction [5], [23].

More recently, there has been a number of developments demonstrating the feasibility of automated facial behavior analysis systems for better understanding of medical conditions such as shizophrenia [34] and post traumatic stress disorders [31]. Other uses of automatic facial behavior analysis include automotive industries [8], education [22], [13], and entertainment [7]. While great progress has been made, there are still a lot of challenges facing automatic emotion recognition [21], namely: generalizability of models across datasets, accounting for individual differences, and recognition of subtle expressions.

While deep learning techniques, such as Convolutional Neural Networks (CNN), enabled a large amount of progress in many areas of computer vision [17], [33], [32], this progress has not yet been fully translated to emotion recognition. This is partly because of lack of large training datasets that such approaches tend to rely on. While tasks such as object recognition enjoy datasets with millions of diverse images [17], most datasets [19] [36], [35], [1], [18] for facial expression analysis are made up of images from tens or hundreds of subjects displaying a small range of emotions. This makes it extremely difficult to train models that generalize well within, not to mention across datasets.

This material is based upon work supported in part through a grant from the CMU-Yahoo! InMind project. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of Yahoo!

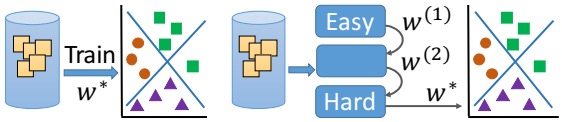

Fig. 1: **(a)** The traditional way to train a model fails to consider the complexity of facial expressions where introducing noisy or difficult samples early in training may hurt model performance. **(b)** The training data is distributed into different complexity levels based on a predetermined curriculum. The training procedure begins from easy to hard image samples (The illustration of complexity is shown in Figure 3). Then the model can be guided to achieve a better performance.

In this paper, we attempt to partly address that by presenting a novel learning technique for deep learning methods that leads to better generalization. In deep neural networks, the model parameters are learned in an iterative fashion using stochastic gradient descent and its variations. Because, the objective function in deep networks has a highly non-convex shape, the order of sample presentation for such networks is important. Curriculum learning allows the model to learn simpler instances first so they can be used as building blocks to learn more complex ones. In our work, we propose to extend the concept of curriculum learning for facial expression recognition. We introduce a specialized curriculum which follows the natural complexity found in facial expression datasets, the expression intensities. We evaluate our approach on a meta-dataset of basic emotions. The summary of our approach can be seen in Figure 1.

Our paper is structured as follows. We first overview the research done on curriculum learning and modern approaches for emotion recognition in Section II. We then present our model in Section III, followed by a description of experiments in Section IV and the results in Section V. We conclude and present future directions in Section VI.

II. BACKGROUND

We structure our background section by first discussing the recent advances in curriculum learning that our work builds on. We then follow by a discussion of modern approaches to emotion recognition from facial expressions.

978-1-5090-4023-0/17 $31.00 © 2017 IEEE
DOI 10.1109/FG.2017.68

A. Curriculum Learning and CNNs

In his seminal work Elman [12] studied the effect of a learning structure on a synthetic grammar task. His work was inspired by language learning in children and demonstrates that a neural network is able to learn the grammar when training data is presented from simple to complex order and fails to do so when the order is random. Bengio et al. [6] demonstrate that curriculum learning results in better generalization and faster learning on a synthetic vision and word representation learning tasks. Pentina et al. [26] investigate the effect of curriculum learning in a multi-task learning setup and proposed a model to learn the order of multiple tasks. Their experiments on a set of vision tasks show that learning tasks sequentially is better than learning them jointly. Finally, Vanya et al. [2] apply curriculum learning to natural image classification task by training a CNN from scratch. Our work differs from the previous approaches, by demonstrating the benefits of curriculum learning in facial expression recognition task. To our knowledge, curriculum learning has never been applied to the task of facial expression recognition

B. Emotion Recognition

While other computer vision tasks have embraced deep learning and convolutional neural networks (CNNs), they have not been as popular in emotion recognition, with the field dominated by hand-crafted features such as Local Binary Patterns and Histograms of Oriented Gradients [28], [20]. For facial expression recognition deep learning models do not always lead to best results and are at times outperformed by simpler hand crafted features [14], [4]. We review the several attempts at emotion recognition using deep networks and CNNs.

Rifai et al. [27] demonstrate how to use a Contractive Auto-Encoder to pretrain a face representation in a semi-supervised manner. Their approach attempts to disentagle the effects on representation due to emotion and other factors such as pose variation and identity. Kahou et al. [16] present a multimodal emotion recognition model, that uses CNNs for the image based emotion recognition from faces. They additionally explore a number of pipelines and training regiments to train the model. Finally, Ng et al. [24] demonstrate how to fine-tune CNNs on increasingly more task relevant datasets for better emotion recognition results. They demonstrate how a CNN trained for object recognition and then tuned for emotion recognition in small and noisy face images, followed by specific target dataset. Our work builds on top of CNN work for facial expression recognition, by applying curriculum learning to CNN training.

III. OUR APPROACH

Overview Our goal is to design an algorithm that introduces curriculum learning to deep convolutional neural networks. Our hypothesis is that starting model learning on simple examples will help with the optimization of CNNs. In the case of facial expression recognition task, the proposed training procedure is based on two assumptions: (a) facial expressions in images have different complexity level to recognize; (b) the order of training samples, prioritizing simpler samples before more complex ones, will benefit the optimization of the models. Under these assumptions, the primary challenge is how to sort image samples into a sequence of subsets that illustrates the simpler concepts first.

Formally, given a training dataset $\mathbb{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$, where $x_i \in \mathbb{R}^d$ denotes the i^{th} training samples. n is the number of training samples and y_i represents its label. The estimated label $\hat{y_i}$ is predicted by $f(x_i, \mathbf{W})$, and $\mathbf{W}$ represents the model parameters of the decision function f. Let $L(y_i, f(x_i, \mathbf{W}))$ denote the loss function which calculates the cost between the ground truth label y_i and the estimated label $\hat{y_i} = f(x_i, \mathbf{W})$. The facial expression recognition model is then optimized by:

$$\mathbf{W}^\star = \underset{\mathbf{W}}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, f(x_i, \mathbf{W})). \tag{1}$$

Here $\mathbf{W}^\star$ indicates the optimal model parameters. In our work we exploit the fact that facial expressions with high intensities are easier to recognize than those with low intensities. This leads to the proposed algorithm shown in Figure 2, we first group images into subsets based on the intensities of their facial expressions (*i.e.*, from high intensity to low intensity). We then train the model via iterative learning of increasingly complex images.

In the following sections, we demonstrate in detail: (1) the CNN architecture and objective function we use in our experiments, (2) the design of complexity function curriculum and subsets generation, and (3) the optimization procedure of our algorithm.

A. CNNs

To assess the assumption that proper data ordering could help the optimization of the model, we focus on the training of deep neural networks. Because they perform very well in other computer vision tasks (e.g., they have some translation invariance). In CNNs, the network is composed of successive convolutional layers (Conv layers), followed by fully connected layers (FC layers). In the last fully connected layer, each sample x_i is passed to a softmax function, which turns the values $\mathbf{x}_i^T\mathbf{W}$ into a valid probability distribution $[p_1, ..., p_k]$ as shown in Equation 2, where k is the kth class and $\mathbf{W} = [w_1, ..., w_k]$.

$$p(\hat{y_i} = j|\mathbf{x_i}) = \frac{\exp^{\mathbf{x}_i^T \mathbf{w}_j}}{\sum_{k=1}^{K} \exp^{\mathbf{x}_i^T \mathbf{w}_k}} \tag{2}$$

The cross-entropy function is used in our experiments to train the CNN model:

$$L = \frac{-1}{n} \sum_{i=1}^{n} [y_i \log \hat{y_i} + (1 - y_i) \log(1 - \hat{y_i})]. \tag{3}$$

The traditional way to optimize this objective function in Equation 3 is to use stochastic gradient descent (SGD) with mini-batches. When using mini-batches randomly created from the training set, one potential issue is that the image

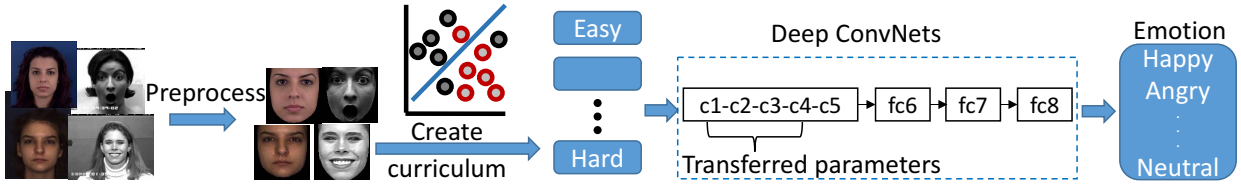

Fig. 2: Framework of our proposed algorithm. Faces are first pre-processed by face alignment and masking, Then the processed faces are ranked into different subsets based on a curriculum. In our experiments, we use a CNN with weights pre-trained on ImageNet dataset. We fix the weights of convolutional layers and fine-tune the weights of fully connected layers. The training iterations start from easy to hard subsets. In each iteration, the optimal model parameters $W^\star$ are selected for the next iteration.

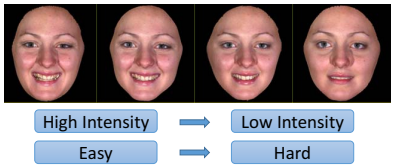

Fig. 3: Illustration of the complexity of image samples. Faces from rmotion recognition datasets have different intensity levels of facial expressions.

samples will include different levels of intensities within a single mini-batch. Such sampling strategy of mini-batches fails to consider the complexity of image samples where noisy or difficult samples can slow down convergence or even mislead the learner rather than help it establish the right location of the decision surface [6]. We create a curriculum based on the complexity of image samples and optimize the objective function from easy to hard.

B. Curriculum learning

Curriculum learning begins with an easy case, slowly extends it to the fully complex target task by taking tiny steps in the problem space, trying not to stray far from the relevant neighborhoods of the solution space [6]. To guide the model to a good local minima, a series of subsets of decreasing expression intensity that culminates in the complete data set offers a natural progression.

Suppose we have a well-defined complexity function $C(y_i, f(x_i, \mathbf{W}))$ which measures the complexity of image samples. The training data is then ranked by the complexity function. An image sample with a higher rank (*i.e.*, lower complexity) is supposed to be learned earlier. In our experiments, we define the complexity function as the facial expression intensity level of image samples. That is, higher intensity indicates lower complexity (see Figure 3). The training data is split into b subsets using this complexity function $C(y_i, f(x_i, \mathbf{W}))$: $D^{(1)}, ..., D^{(b)}$, where columns of $D^{(j)} \in \mathbb{R}^{m \times n_j}$ correspond to the samples in the j^{th} subset, n_j is the sample number in the subset and $\sum_{j=1}^{b} n_j = n$. Given two training samples $x_a \in D^i, x_b \in D^j$, the complexity will follow $C(x_a) < C(x_b), \forall i < j$.

C. Optimization

The subsets to train deep ConvNets are constructed as in Section III-B, and we use the cross-entropy loss function defined in Equation 3. We use the Baby Steps curriculum [30] for our task. That is, in the training procedure, the complexity of the training data should be increased while the simpler image samples should not be discarded. We explored other curriculum strategies (*i.e.*, switch to hard image samples and discard easy ones after certain number of iterations). These methods perform worse than the Baby Steps curriculum and their results are not included.

The training starts with the easiest subset, early stopping is used when the accuracy criteria on the validation subset does not improve in t number of epochs, the next hard subset is merged to the current dataset and the model weights $\mathbf{W}^\star$ are adopted as the initialization for the next training iteration. When the hard subset is added to the current dataset, the basic learning rate of the model is decreased in order to reduce the influence of hard image samples. The optimization procedure is shown in Algorithm 1.

Algorithm 1 CNNs with curriculum learning

Input: Input dataset $D = \{D^i\}_{i=1}^{b}$ ordered by predetermined curriculum
Output: Optimal model parameter $\mathbf{W}^\star$

```
D^train = ∅
for i = 1, .., b do
    D^train = D^train ∪ D^i
    for epoch = 1, ..., k do
        train(W, D^train)
    end for
    select best W*
    update learning rate
end for
```

IV. EXPERIMENTS

In this section, we present our experiments to evaluate whether our proposed algorithm, by considering the complexity of facial expressions, can lead to better trained CNNs. Furthermore, we examine how curriculum learning affects the training of shallower models such as multilayer perceptrons (MLP). Finally, we investigate how the design of curriculum affects the performance of our proposed algorithm.

A. Baseline Models

To explore the effect of curriculum on training emotion recognition models, we compare our approach to a number of commonly used baselines. For a fair comparison, all feature representations are evaluated using softmax classifiers.

HOG. Histograms of Oriented Gradients (HOG) descriptor [9] partitions a given image into 8×8 pixel blocks. More specifically, we use blocks of 2×2 cells which leads to 12×12 blocks of 31 dimensional histograms [3] (4464 dimensions in total to describe the face). The HOG feature is selected because it performs very well in prior work [4].

LBP. The Local Binary Pattern (LBP) descriptor [29] encodes the local texture and global shape of face images. The LBP descriptor [29] is computed by first equally dividing the face images into small 7×6 regions. Then the uniform LBP features are extracted from each sub-region and concatenated into a single, spatially enhanced feature histogram, which results in 2478 dimensions to describe the face. LBP feature and its variants are widely used features for emotion recogntiion tasks in prior work [11].

CNN off-the-shelf. As a baseline, we compare our proposed algorithm with an off-the-shelf CNN model - AlexNet [17]. We train a softmax classifier on top of the feature representations of $fc7$ layer with all the other layers fixed.

B. Datasets

As there is no large-scale, strongly labeled, and publicly available facial expression recognition dataset, we create a meta-dataset based on five datasets and evaluate our algorithm on this dataset. A brief description of these facial expression databases is as follows:

Extended Cohn-Kanade (CK+) database. CK+ database [19] contains 593 sequences across 123 subjects.

BU-3DFE. BU-3DFE database [36] contains 100 subjects, ranging from 18 to 70 years of age, with a variety of ethnic/racial ancestries.

BU-4DFE. BU-4DFE database [35] contains 606 $3D$ facial expression sequences captured from 101 subjects, with a total of $60{,}600$ captured frames. This database consists of 58 female and 43 male subjects.

MUG. MUG database [1] contains 52 subjects with 1462 image sequences in total. Each sequence varies from 50 to 160 images. The subjects display various emotions and emotional attitudes. For our experiments, we only select six basic emotions and neutral state.

Radboud Faces Database. Radboud database [18] contains 67 subjects, displaying 8 emotional expressions (including six basic emotions, neutral and contempt). Each emotion was shown with three different gaze directions and all images were taken from five camera angles simultaneously. We use only frontal faces of seven emotions in our experiments.

Dataset	CK+	BU-3DFE	BU-4DFE
Subjects	123	100	101
Images	1854	2500	1818
Dataset	MUG	Radboud	Total
Subjects	52	67	443
Images	2958	1407	10537

TABLE I: Statistics of the meta-dataset. The dataset is created by considering the balance of its composing datasets, diversity of subjects, and different intensity levels of facial expressions.

Meta-dataset. To increase the diversity of subjects and preserve the variance of intensities of different subject, we sample frames from these datasets with different intensity levels. To keep the balance of its composing datasets and prevent from over-fitting to certain subjects, we equally sample frames for each subject. The constructed dataset summary is shown in Table I. We define three levels of expression intensity: low, medium and high. BU-3DFE dataset contains 4 expression intensity levels, we map the highest level to high, two middle levels to medium and the lowest level to low. For video datasets (CK+, BU-4DFE, MUG), we do not have intensity labels, so we infer them based on their recording protocols. In CK+, all the videos start from a neutral expression and end on a peak one. We use the last frame as the high intensity one and preceding frames as the medium and low intensity ones. In BU-4DFE and MUG datasets, the videos start from a neutral expression, reach a peak expression intensity in the middle and finish with a neutral one. We use the middle frame as the high intensity one and the frames around it as the low and medium intensity ones.

C. Methodology

Data processing. We use OpenFace [3] to process the face images. In order to better analyze the texture of the face, we map the detected face to a common reference frame and remove changes due to scaling and in plane rotation. This results in a 256×256 pixel image of the face with 90 pixel interpupilary distance. In order to remove non-facial information from the image, we also perform masking of the image by using a convex hull surrounding the aligned feature points. The input and results of the procedure is illustrated in Figure 4.

Deep ConvNet architecture. Consistent with recent work, we use AlexNet [17] as our model and all experiments are carried out based on Caffe [15]. We replace the output neurons of the last fully-connected layer ($fc8$ layer) to 7 which equals to the number of predicted emotions. We initialize the weights of $fc8$ layer from a zero-mean Gaussian distribution

(a) (b)

Fig. 4: Example of face detection and alignment, followed by masking.

with standard deviation 0.01. The data augmentation in our experiments consists of generating image translations and horizontal reflections. More specifically, we extract random 224×224 patches and their horizontal reflections from the 256×256 images and fine-tune our network on these extracted patches. At test time, we extract five 224×224 patches (the four corner patches and the center patch) as well as their horizontal reflections. Then we average the predictions made by the network's softmax layer over the ten patches. The *step* learning rate update policy is used where learning rate is multiplied by a gamma factor in each step. We use 5-fold testing. In each fold, the split of training/validation/test data is 70%/10%/20%.

V. Results and Discussion

To evaluate the effectiveness of our proposed algorithm, we conduct an extensive set of experiments and compare our approach to the baseline models. We also investigate the effects of curriculum learning on not only CNNs but also multilayer perceptron (MLP). Finally, we compare three ways to create a curriculum.

A. CNNs with curriculum learning

The first task in our experiments is to test if curriculum learning (CL) helps with training better facial expression recognition models. The performance of models is measured by overall accuracy and can be seen in Table II.

Model	Accuracy
LBP	0.786
HOG	0.807
CNN off-the-shelf	0.566
Fine-tune w/o CL	0.819
Fine-tune w/ CL (hard first) (ours)	0.814
Fine-tune w CL (easy first) (ours)	**0.830**

TABLE II: Performance comparisons between our proposed algorithm with curriculum learning (CL) and the baseline models on the meta-dataset. We also include the influence of different data ordering on the model performance. Using the curriculum, our approach yields performance superior to both widely used handcrafted features (*i.e.*, LBP and HOG) and the CNN trained without the curriculum.

It is clear that the performance of our proposed algorithm outperforms the baselines by a large margin. To test if the improvements are statistically significant, we perform a t-test. CNN with curriculum learning ($M = 0.829$, $SD = 0.036$) has a significantly higher accuracy when compared to a CNN without curriculum learning ($M = 0.804$, $SD = 0.0026$); $t(4) = -27.95$, $p < 0.001$. This leads us to the conclusion that there is a significant difference in the performance of these two models. To exploit the importance of data ordering, we invert the curriculum (*easy first*) by starting from hard image samples (*hard first*). From the results we can see, a proper data ordering has a positive effect on finding good solutions in local minima (1.6% improvement compared with *hard first* curriculum).

	One-layer	Two-layer	Adaptation-layer
w/o CL	0.776	0.805	0.819
w/ CL (ours)	**0.791**	**0.828**	**0.830**

TABLE III: Performance comparisons between our proposed algorithm and the models without CL using three fine-tuning strategies. Our curriculum based training leads to consistently better performance compared to models trained without curriculum learning (CL).

Furthermore, we evaluate our proposed algorithm based on three commonly used fine-tuning strategies (see Table III). These fine-tuning strategies are (a) fine-tune $fc7$ layer with the rest of the layers fixed (*one-layer*); (b) fine-tune $fc6$ land $fc7$ layers simultaneously with the rest of the layers fixed; and (c) replace $fc7$ layer with an adaption layer (the size is set as 2048), then fine-tune $fc6$ and adaption layers simultaneously (*adaptation layer*) [25].

As there is a huge domain shift from ImageNet [10] to facial expression recognition datasets, it is interesting to note that using an adaptation layer with randomly initialized weights can achieve the transfer better. From Table III we can see that our proposed algorithm performs consistently better compared with the fine-tuning strategies without a curriculum. This leads us to conclude that our proposed algorithm is not constrained to a certain type of fine-tuning strategy but can be applied to more general neural network training.

B. Multilayer perceptron with curriculum learning

In our next experiment, we want to see if the curriculum learning benefits the shallower neural networks with hand-crafted features. To avoid over-fitting, we design a multilayer perceptron (MLP) with two hidden layers (the layer sizes are 500 and 100 respectively). We initialize the weights of hidden layers from a zero-mean Gaussian distribution with standard deviation 0.01. To reduce the influence of random initialization, we conduct five independent experiments and use the average of accuracies as the final result.

From Table IV we can see, MLP with the curriculum yields consistently better performance than MLP without the

	1	2	3	4	5	Average
w/o CL	0.838	0.837	0.838	0.840	0.834	0.837
w/ CL (ours)	0.850	0.842	0.844	0.842	0.847	**0.845**

TABLE IV: Performance comparisons between multilayer perceptron (MLP) with and without curriculum learning (CL) on five repeated experiments with independent random initialization. Ours curriculum learning strategy yields consistently better performance than the models without CL.

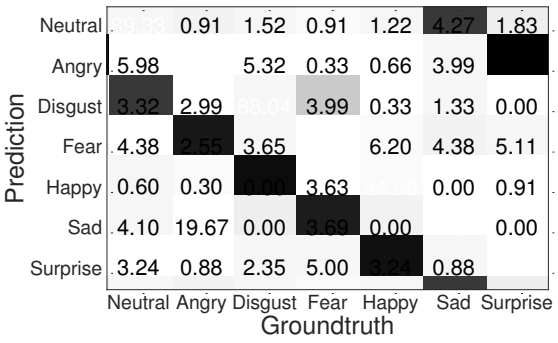

Fig. 5: Confusion matrix of our best model.

curriculum. To test if the improvements are statistically significant, we perform a t-test to compare the mean accuracies of five independent experiments: with curriculum learning ($M = 0.845$, $SD = 0.0031$), and without curriculum learning ($M = 0.837$, $SD = 0.0020$); $t(9) = -4.6289$, $p = 0.0012$; $p < 0.005$, leads us to conclude that there is a statistically significant difference between the two models. It is interesting to note that the improvements of MLP are relatively smaller compared with that in Section V-A. This could be due to the fact MLP in our experiments has a simpler network architecture with fewer parameters to learn, meaning that the model parameter are easier to learn and different data ordering does not make as big a difference in performance when compared with CNNs. However, this demonstrates the importance of training data ordering for deep and complex networks. The confusion matrix of the best performing model is shown in Figure 5.

C. Curriculum strategies

In the next set of experiments, we explore a number of ways to define a curriculum. First, we design the curriculum manually based on the datasets (*manually design*). That is, for image datasets, we infer the intensity from labels (*i.e.*, images from BU-3DFE dataset) and for video datasets, we assume the middle frame to be the highest intensity of expression.

We also explore the ways of designing the curriculum automatically by utilizing the information of facial expression recognition classifiers. Our assumption is that the complexity of the image sample can be measured by the predicted probability distribution from a given classifier. More specifically, an image sample is considered as easy (high intensity) if it is classified correctly and the model is confident in that prediction, otherwise, it is considered as a hard sample. Formally, given an image x_i and the ground truth label y_i, we define the complexity of images $C(x_i)$ as

$$C(x_i) = \begin{cases} p(\hat{y_i} = j|x_i), y_i = j, \\ \\ -p(\hat{y_i} = j|x_i), y_i \neq j, \end{cases} \tag{4}$$

where $p(\hat{y_i} = j|x_i)$ is as shown in Equation 2 and $C(x_i) \in [-1, 1]$. We compare HOG and CNN feature representations with softmax classifiers in the following experiment to construct the curriculum. We then fine-tune the AlexNet model based the curriculum. The performance is measured by the average of five independent experiments.

Method	Accuracy
Manually design	0.830
CNN	0.817
HOG	0.824

TABLE V: Performance comparisons of different ways to define a curriculum.

Using the same classifier trained on training dataset, we can rank the image samples by probability and create the curriculum based on the ranking. From Table V we can see, the curriculum created by classifiers automatically can also boost the fine-tuning performance of the CNN model in turn. This infers there is a correlation between the predicted probability distribution and the complexity of facial expressions.

VI. CONCLUSIONS

In this paper, we introduced curriculum learning to the emotion recognition task. To facilitate this study, we created a meta-dataset from five facial expression recognition datasets. Then we conducted an extensive set of experiments exploring the effect of the curriculum learning on model performance. We showed the benefits of curriculum learning for training CNN model with different fine-tuning strategies. Furthermore, we showed that curriculum learning works for shallower networks (multilayer perceptron). Our results indicated that a proper data ordering (*i.e.*, from easy to hard) can help guide the optimization of deep neural networks. Finally, even an automatically designed curriculum without human guidance outperforms a CNN model without curriculum learning.

For future work, we would like to extend our method from images to videos. More specifically, we would like to investigate how curriculum learning would affect emotion recoginition in videos. We also would like to exploit different curriculum strategies to facial expression recognition tasks.

REFERENCES

[1] N. Aifanti, C. Papachristou, and A. Delopoulos. The mug facial expression database. In *WIAMIS*, 2010.

[2] V. Avramova. Curriculum learning with deep convolutional neural networks. 2015.

[3] T. Baltru, P. Robinson, L.-P. Morency, et al. Openface: an open source facial behavior analysis toolkit. In *WACV*, 2016.

[4] T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. In *Facial Expression Recognition and Analysis Challenge*, 2015.

[5] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *CVPR Workshops*, 2003.

[6] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009.

[7] P. M. Blom, S. Bakkes, C. T. Tan, S. Whiteson, D. Roijers, R. Valenti, T. Gevers, et al. Towards personalised gaming via facial expression recognition. 2014.

[8] C. Busso and J. Jain. Advances in multimodal tracking of driver distraction. In *Digital Signal Processing for In-Vehicle Systems and Safety*, pages 253–270. Springer, 2012.

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[11] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon. Emotiw 2016: Video and group-level emotion recognition challenges. In *ICMI*, 2016.

[12] J. L. Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

[13] A. Graesser and A. Witherspoon. Detection of Emotions during Learning with AutoTutor. *CogSci*, pages 285–290, 2005.

[14] A. Gudi, H. E. Tasli, T. M. D. Uyl, and A. Maroulis. Deep Learning based FACS Action Unit Occurrence and Intensity Estimation. volume 2013, 2015.

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *MM*, 2014.

[16] E. S. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, J. Sebastian, P. Froumenty, Y. Dauphin, R. C. Boulanger-Lewandowski, Ferrari, M. Mirza, D. Warde-Farley, A. Courville, P. Vincent, R. Memisevic, C. Pal, and Y. Bengio. EmoNets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10(2):99—-111, 2016.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[18] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388, 2010.

[19] P. Lucey, J. F. 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*, 2010.

[20] B. Martinez and M. F. Valstar. Advances , Challenges , and Opportunities in Automatic Facial Expression Recognition. In *Advances in Face Detection and Facial Image Analysis*, pages 1–37. 2016.

[21] B. Martinez and M. F. Valstar. Advances, challenges, and opportunities in automatic facial expression recognition. In *Advances in Face Detection and Facial Image Analysis*, pages 63–100. Springer, 2016.

[22] B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp, and a. Graesser. Facial Features for Affective State Detection in Learning Environments. *CogSci*, pages 467–472, 2007.

[23] D. McDuff, R. el Kaliouby, D. Demirdjian, and R. Picard. Predicting online media effectiveness based on smile responses gathered over the internet. In *FG*, 2013.

[24] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *ICMI*, 2015.

[25] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.

[26] A. Pentina, V. Sharmanska, and C. H. Lampert. Curriculum Learning of Multiple Tasks. In *CVPR*, 2015.

[27] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza. Disentangling factors of variation for facial expression recognition. In *ECCV*, 2012.

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

[29] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *IVC*, 27(6):803–816, 2009.

[30] V. I. Spitkovsky, H. Alshawi, and D. Jurafsky. From baby steps to leapfrog: How less is more in unsupervised dependency parsing. In *NAACL*, 2010.

[31] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency. Automatic nonverbal behavior indicators of depression and ptsd: Exploring gender differences. In *ACII*, 2013.

[32] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013.

[33] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf. DeepFace : Closing the Gap to Human-Level Performance in Face Verification. In *CVPR*, 2014.

[34] S. Vijay, T. Baltrušaitis, L. Pennant, D. Öngür, J. Baker, and L.-P. Morency. Computational study of psychosis symptoms and facial expressions. In *CHI Workshops*, 2016.

[35] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3d dynamic facial expression database. In *FG*, 2008.

[36] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *FG*, 2006.