Facial Expression Recognition

Group 521

Team members

Shelly Gamlielly, shelly.gamlielly@mail.huji.ac.il

Yerus Mandfro, yerus.mandfro@mail.huji.ac.il

Advisor

Mr. Or Sharir, or.sharir@mail.huji.ac.il

Organization *Alyn Hospital*

Table of Contents

Abstract ---------- **3**

Introduction ---------- **3-4**

Related Work ---------- **4-5**

Methods and Materials ---------- **5-7**

Solution and improvement ---------- **7**

Description of approach and key components ---------- **8-11**

Architecture ---------- **9-11**

Evaluation and verification ---------- **11-15**

Conclusion ---------- **16**

Future Work ---------- **17**

Bibliography Referenced ---------- **18**

Appendix ---------- **19**

Abstract

Our goal is to help children who are patients in the ALYN Hospital. Some of the children use life-supporting machines such as ventilators, and cannot express themselves verbally. Instead, they rely on an alternative form of communication using facial expressions. Recognizing facial expressions requires continuous eye contact and familiarity with each facial expression to understand its meaning. In these cases, a facial expression recognition system that will voice the word equivalent to a facial expression could improve the children's quality of life by helping them communicate effectively with their environment.
Last year, a different group worked on the same problem, where they used a neural network to learn the facial expressions used by a particular child. The system relies on the naive assumption that it would be easy to acquire multiple images per facial expression for every child using the system. It turned out that this is not the case because the children's disabilities make it considerably difficult for them to perform facial expressions for the system, causing the hospital's staff not to use the system.
Building upon last year's effort, we propose an alternative solution that only relies on an unsupervised video of children's faces as they go about their day. To achieve that, we leverage an unsupervised domain adaptation method that allows us to take a facial expression system trained in a controlled environment on people without disabilities (in our case fellow students), and then transfer it to the domain of faces of a specific child. We hope this streamlined process will make it more accessible for these children, and end-up improving their daily lives.

Introduction

Most children who use life-supporting machines such as ventilators suffer from neuromuscular diseases such as SMA and TNNT1. These diseases damage the nerve of the muscles throughout the body. The use of ventilators also makes it considerably difficult for a child to produce sound or voice, therefore they rely on an alternative form of communication using facial expressions.
Communicating it’s a fundamental aspect during social and emotional development also in order to acquire education or for simply daily life. Therefore we would like to find a way to make communication possible and effortless in a way that meets the child's needs.
Recognizing facial expressions requires continuous eye contact and the ability to understand the meaning of facial expression.
Those requirements are difficult to implement for several reasons. One of them is the fact that in class, for example, there are many kids drawing a teacher’s attention. If a child who communicates using facial expressions needs something or wants to answer a question, the teacher might not notice his expression, unless there is eye contact. In addition, there is a lack of convention mapping facial expressions to language. Therefore it is difficult to understand the expressions, especially for those who are not familiar with them. We are interested in identifying common and easily recognizable facial expressions in a way that there is a large number of patients who

agree on them, although most of the conventions on facial expressions vary among children. Examples of conventional facial expressions are long blink, eyebrow lifting, eyebrow squeezing, tongue removal, smile, sideways glance, light head nod (up and down), mouth opening.

Recognizing Alyn's kids' facial expressions has two interesting aspects.
Because each child has a number of limited movements that he can perform there is not one uniform expression for everyone for a specific word therefore there are different expressions for the same word. For example, some children can only move their heads up and down when others can move it left and right.
The second aspect is the natural expressions that the children in Alyn have which are not related to a word. For example, there are facial expressions that the children make even without trying to communicate such as blinking that can be translated as a facial expression to say "yes" or "no".

First, we focused on product characterization for the customer. We made an effort to understand the children's medical condition in order to get to know the target population better. Later, we focused on the practical part of getting better prediction results on children from Alyn hospital. Since the system wasn't designed for those kids, we had to make sure the predictions are accurate on them. For this reason, we met the medical staff of Alyn hospital, introduced the system, and installed the software on the children's computers. During the session, we taught the staff how to use the software, train it, and tag the images of the kids. After a period of time, we let the staff explore the software, we got their feedback. We realized that the data collection process was not possible in the current format and so we decided to make changes that will be described later. Our previous work allowed us to better characterize the problem and find a suitable strategy to continue the project from the point where it stopped, using domain adaptation methods.

Related Work

Eye-tracking is a sensor technology that makes it possible to control a computer or other device using eye positions and eye movement. This ability is vital for people who are unable to communicate verbally or use their hands. Gazespeaker. [4] is a free software designed to help people with disabilities to communicate and interact with their environment and the web uniquely with their eyes using a keyboard.

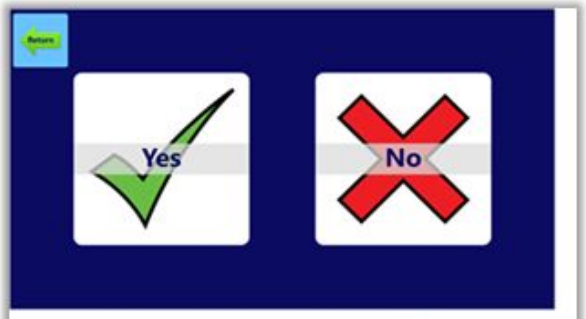

Figure 2 : Alyn Hospital uses communication software that allows the child to choose the message as yes or not. The image above illustrates the use of the system. Alyn's kids use facial expressions to give a quick answer when yes / no questions are asked. But the communication software is unable to detect facial expressions.

Figure 1 : The Eye-tracking system and software allow the user to write and choose a message. This system includes a camera focusing on one or both eyes and records their movement as the viewer looks at stimulus

Communication using the software Gazespeaker doesn't answer a child's needs. One of the main reasons for this is the fact that it takes a long time to set up the system for each individual and every movement might require resetting. In addition, using the keyboard with eye tracker for writing is a complex mission for a child with disabilities, and only a few are capable of writing. Writing using an eye tracker takes a few minutes which is too long for expressing a feeling or a need. Also, the system requires a child to be in front of a computer for a long time, thus his field of view is reduced and he can't use the computer for other tasks such as learning or playing games.

Methods & Materials

Our project is an extension of an existing project. The previous project focused on recognizing facial expressions using a simple neural network that is built and trained for each child individually. Description of previous work composed three main phases: data collection phase, training phase, detection phase.

Data collection and training phase

In this phase, the user should record short clips showing facial expressions of a child for whom the network is intended and tag the images extracted from the clips. Next, the user is required to train the neural network.

Due to movement restrictions from which the children suffer, each child expresses a particular word with a unique facial expression. Therefore the previous project members chose to train a neural network for each child with his own images. That means every child should have the software installed on the personal computer, and trained only on his images.

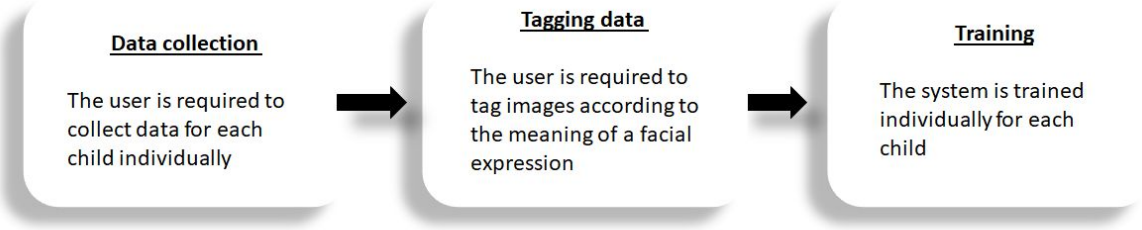

figure 3 : Previous project work phases description.

Detection phase

In this phase the software that uses the laptop camera to sample images run. Afterword prediction is performed on the saved model and the software sounds the equivalent word to a facial expression that is captured according to the prediction.

There are naive assumptions of the previous project approach that in reality turned out to be problematic. The user is required to perform complicated operations as we mentioned earlier in the three phases, data collection, tagging, and training. Therefore although the interface is relatively easy to access, for users without technology orientation it is quite difficult working with. In addition, the training phase

takes a long time (several hours) which is one of the reasons that led Alyn's staff to avoid using the software in addition to the inability to work comfortably and effectively with the software.

Collecting the data from the patients of Alyn hospital is a challenging task and thus it is not possible in the way they expected.

The data collection was designed to be performed using a GUI so the user will be able to record a video of a child displaying his unique facial expressions. After the collection phase ends, a therapist goes over the video while collecting the images where they can see the facial expression and tag them with the appropriate word.

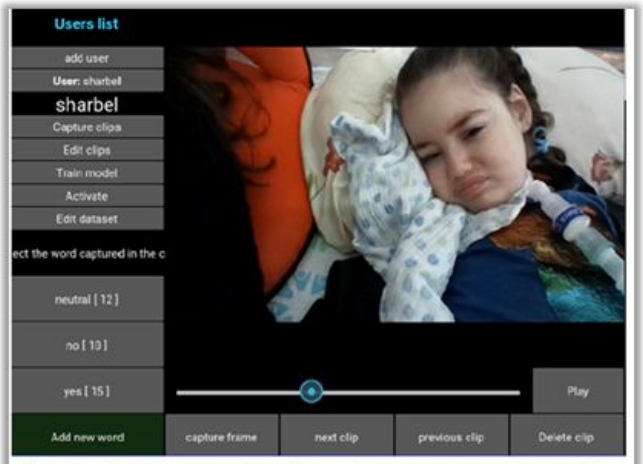

Figure 4: An example of how the user is supposed to tag an image with the appropriate word.

In order to do this, the child's therapist had to make him use his facial expression and record it. But this practice has taken a long time whereas the use of facial expression is an action that requires much effort from the child. Beyond that, the children's schedule is tight and it is difficult to find a suitable time to work with them, especially when you take as an account that recording the facial expression takes time and demands attention. In addition, since communication with the Alyn staff is limited, data collection cannot be relied on by them. In order to overcome these problems, we suggest training the system on data of other population groups, students in our case. Also in order to reduce training time we suggest training the model using a GPU on Colab instead of CPU on the personal computers of the kids. We also suggest to save the model with the highest validation accuracy and to use the same model for all the kids to make the software more comfortable for users and consume less time.

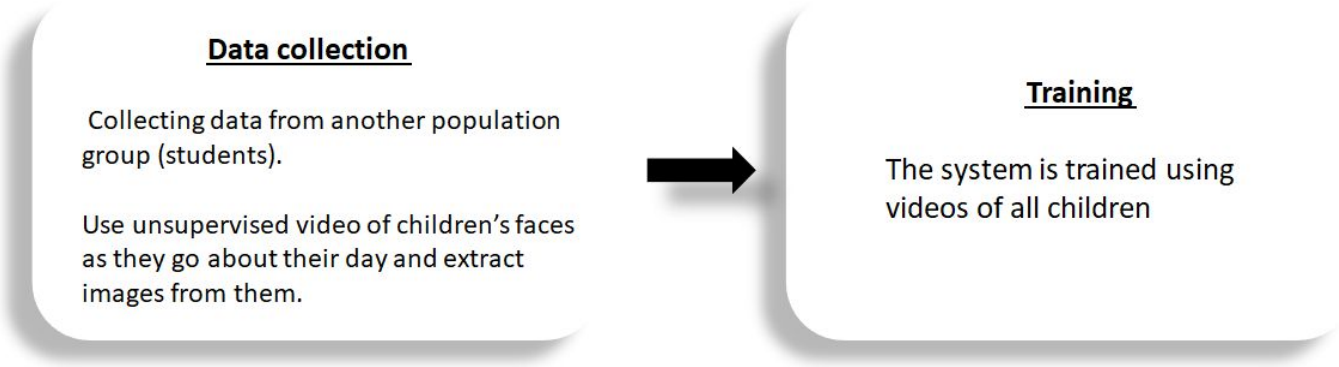

figure 5 : Our suggestion as to the first stage of collecting data and training.

Our Solution and Improvements

Data collection

Our solution is based on collecting data from another population group, in our case, it was more convenient to use data of students. The neural network learns labeled source images of students and can generalize them on unlabeled target images of children from Alyn Hospital. We use unsupervised video of children's faces as they go about their day to collect images from the target domain. We ask our fellow students to send us videos of them making facial expressions for 30 seconds so that each video contains one facial expression and we used the function VideoCapture from the package of cv2 to extract images. In the same way in order to extract images of Alyn Hospital's kids we used videos of them, only in this case the kids did not necessarily have to make specific facial expressions.

Preprocess Images

The first step we do on the preprocessing stage is to detect faces in the collected data and crop the image accordingly using the software package dlib **-** face detector and face recognition tool for image alignment (see figure 6).
In the next step, we perform augmentation that does the following - horizontal flips, gaussian blur, and sharpen for each image (see figure 7).

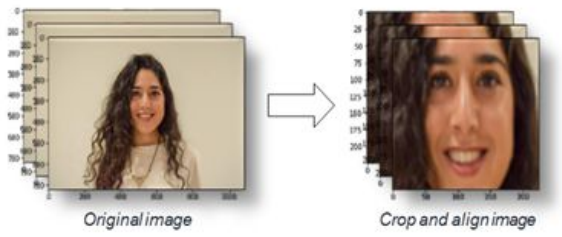

Figure 6 : Example of how the crop and align is performed using dlib.

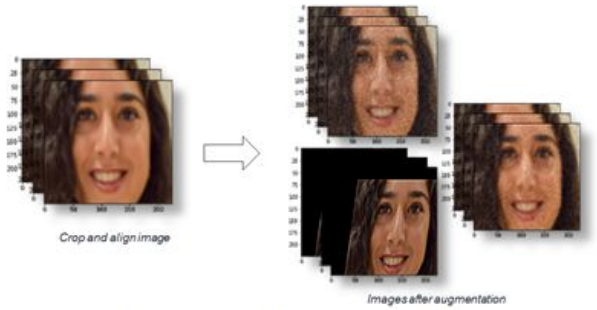

Figure 7 : Example of augmented images.

Training

Our goal is to learn a target representation, and a classifier that can correctly classify target images into one of the categories at test time, despite the lack of in domain annotations. We will discuss this further when describing our approach.

Description of approach and key components

Due to a phenomenon known as dataset bias or domain shift, recognition models trained on one large dataset do not generalize well to novel datasets and tasks. Domain adaptation methods attempt to mitigate the harmful effects of domain shift. In our case using Keras VGGface which was trained on a very large dataset.

Adversarial Discriminative Domain Adaptation [1]

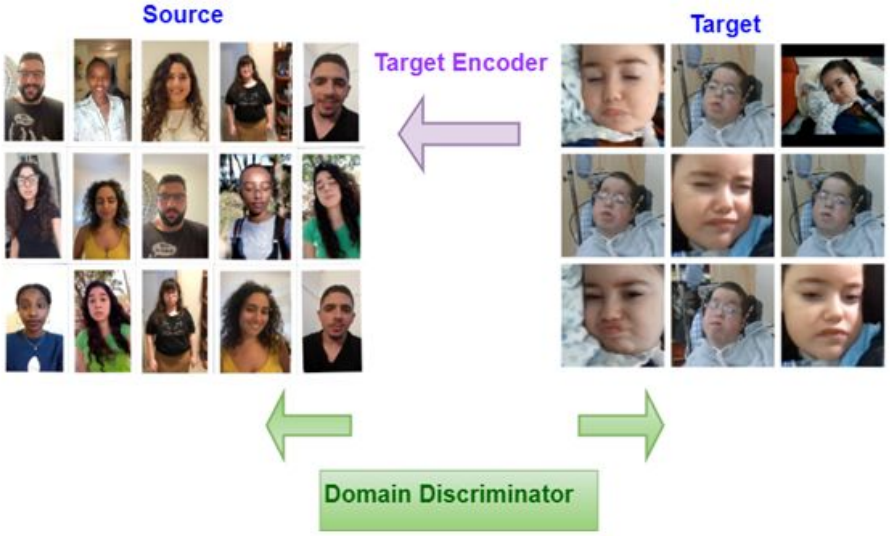

Figure 8: We propose an improved unsupervised domain adaptation method that combines adversarial learning with discriminative feature learning. Specifically, we learn a discriminative mapping of target images to the source feature space (target encoder) by fooling a domain discriminator that tries to distinguish the encoded target images from source examples.

In unsupervised adaptation, we assume access to source images and labels drawn from a source domain distribution, as well as target images drawn from a target distribution, where there are no label observations.
Our goal is to learn a target representation, and a classifier that can correctly classify target images into one of facial expressions at test time, despite the lack of in domain annotations.

Since direct supervised learning on the target is not possible, *domain adaptation* instead learns a source representation mapping, along with a source classifier, and then learns to adapt that model for use in the target domain. In *adversarial adaptive methods,* the main goal is to regularize the learning of the source and target mappings, and so as to minimize the distance between the empirical source and target mapping distributions. If this is the case then the source classification model can be directly applied to the target representations, eliminating the need to learn a separate target classifier.

Architecture

We tried two different architecture approaches.

First approach

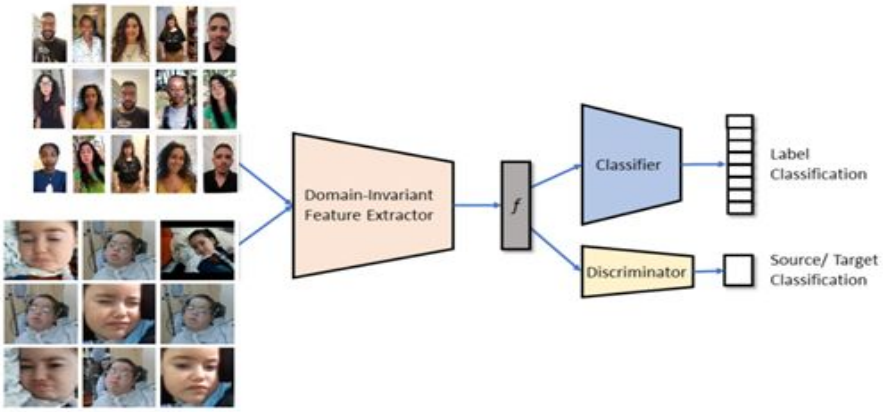

Figure 9: The architecture involves three sub-networks: A domain-invariant feature extractor that finds features in the image. Intuitively, the output of this part should tell if you have for example face or eyes in the image. A classifier that receives the encoded image and classifies it to the desired classes. A domain discriminator that classifies whether a data point is drawn from the source or the target domain.

Adversarial Domain Adaptation Model assigns class labels to images in the target domain by extracting domain-invariant features from the labeled source and unlabelled target domain images.

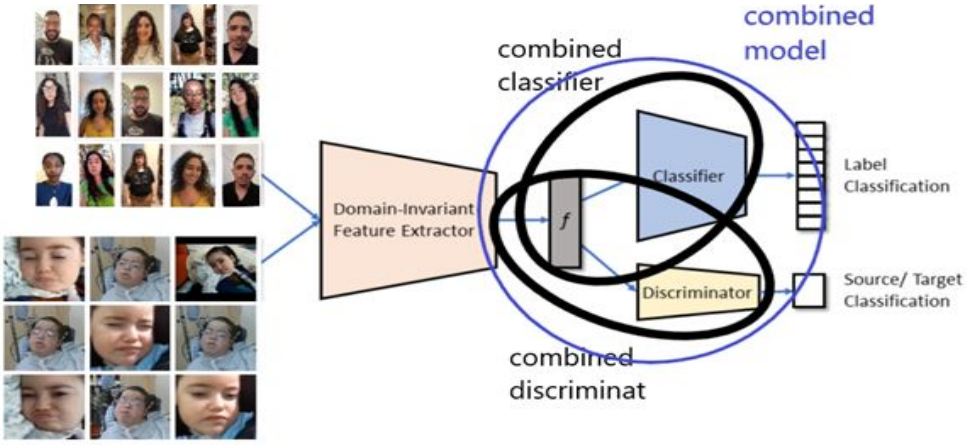

Figure 10: We use the feature extractor to extract features from both the target and source images into one feature space, we flat them and get the representation f.
Next, we build a combined classifier that receives the features as input, a combined discriminator that receives the features as input, and a combined model of the classifier and discriminator.
We train the combined model, while defining the discriminator on the combined model as non trainable. Lastly, we use the combined classifier for predicting the facial expression in each target image.

Second approach
Learn a discriminative representation using the labels in the source domain and then a separate encoding that maps the target data to the same space using an asymmetric mapping learned through a domain-adversarial loss.

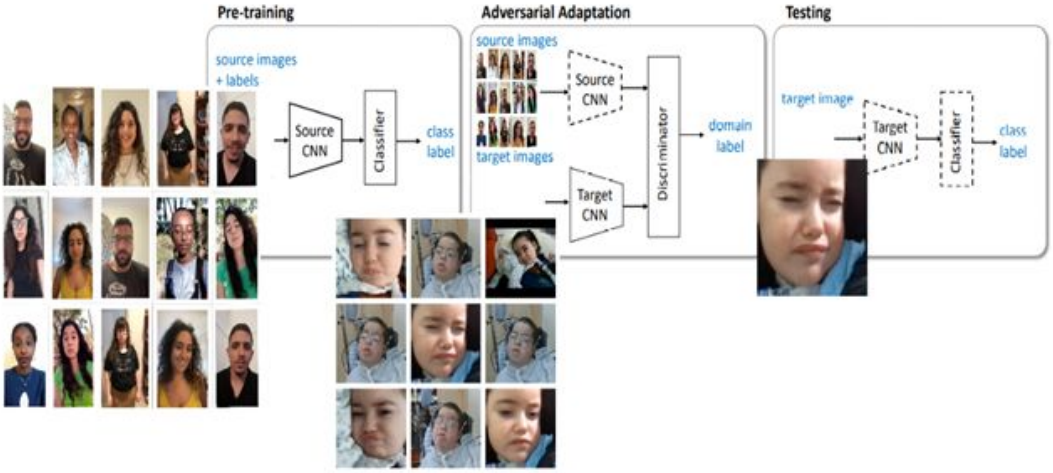

Figure 11: An overview of our proposed Adversarial Discriminative Domain Adaptation (ADDA) approach. We first pre-train a source encoder CNN using labeled source image examples. Next, we perform adversarial adaptation by learning a target encoder CNN such that a discriminator that sees encoded source and target examples cannot reliably predict their domain label. During testing, target images are mapped with the target encoder to the shared feature space and classified by the source classifier. Dashed lines indicate fixed network parameters.

Making Three Design Choices

- Whether to use a generative or discriminative base model?
- Whether to tie or untie the weights?
- Which adversarial learning objective to use?

Given that our target domain is unlabeled, it remains an open question how best to minimize the distance between the source and target mappings. The goal is to make sure that the target mapping is set so as to minimize the distance between the source and target domains under their respective mappings, while crucially also maintaining a target mapping that is category discriminative.

Method	Base model	Weight sharing	Adversarial loss
Gradient reversal	discriminative	shared	minimax
Domain confusion	discriminative	shared	confusion
CoGAN	generative	unshared	GAN
ADDA (Ours)	discriminative	unshared	GAN

Table 1: Overview of adversarial domain adaptation methods and their various properties.

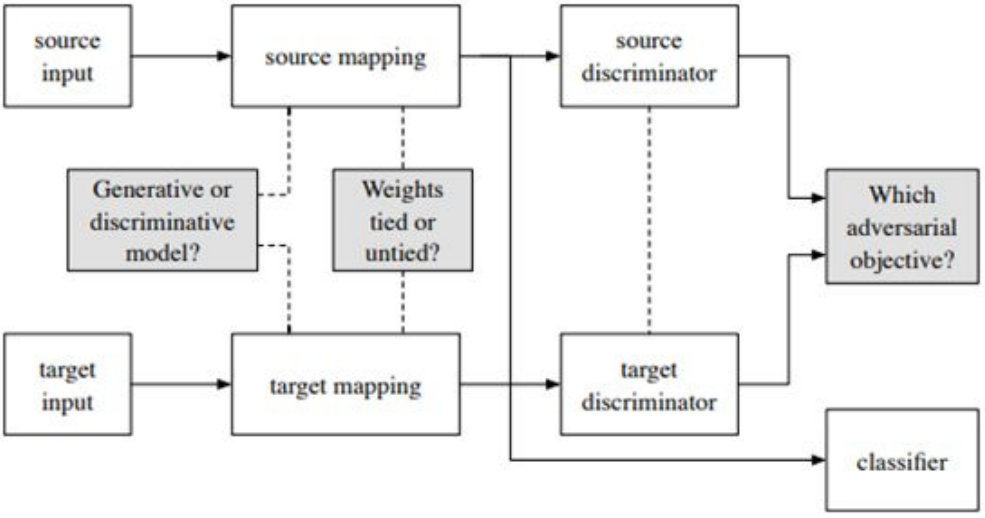

Figure 12 : Our generalized architecture for adversarial domain adaptation. Existing adversarial adaptation methods can be viewed as instantiations of our framework with different choices regarding their properties.

We use a discriminative base model, unshared weights, and the standard GAN loss. First, we choose a discriminative base model, as we hypothesize that much of the parameters required to generate convincing in-domain samples are irrelevant for discriminative adaptation tasks. Next, we choose to allow independent source and target mappings by untying the weights. This is a more flexible learning paradigm as it allows more domain specific feature extraction to be learned. We use the pre-trained source model as an initialization for the target representation space and fix the source model during adversarial training. In doing so, we are effectively learning an asymmetric mapping, in which we modify the target model so as to match the source distribution. This is most similar to the original generative adversarial learning setting, where a generated space is updated until it is indistinguishable with a fixed real space. Therefore, we choose the inverted label GAN loss.

Evaluation and verification

Evaluation Dataset

Students

The dataset used for evaluation consists of images of students, collected and tagged by us. There are four facial expressions that we trained the model on - neutral, smile, closed eyes, and brows lifting. We performed two experiments:

1. The source domain includes images of 6 students while the target domain includes images of a different student.
2. The source domain includes images of 4 students while the target domain includes images of 4 different students.

Alyn's Kids

Due to the global epidemic covid19, the department where the kids are hospitalized is closed for visitors, so we couldn't meet them in person. We asked Alyn's staff to send us all the videos and images that were saved on the kid's personal computer. We also asked them to record more videos and send us. We managed to get a few videos for training and testing and used them as target images after extracting images. It is important to understand that the main problem that got us to using unsupervised domain adaptation methods was the fact that we couldn't get tagged images of the kids. Therefore we could not rely on having labeled images of their facial expressions for testing.

We did collect a few labeled images of Lishay and Sharbel for testing. The facial expressions were: neutral, closed eyes, and lifted brows. Although the test set is not large, it might give us some idea about the model generalization ability.

Data Description:

- Number of classes : 3.
- Number of images :
  - Training images from target domain: 1440
  - Training images from source domain: 1150
  - Validation images from source domain: 290
  - Test images from target: 40

Results

Students

Classes: smile and neutral

Figure 13 : Those graphs describe our results (left) compared to previous work results (right). Right graph clarification - the training set includes images of a group of students and the validation set includes another group of students . Accuracy and Loss of source images and target images of students was calculated while training with 2000 iterations. During training the model does not have the labels of target images, we use them for evaluation. Notice that we got target accuracy - **70.225**, compared to previous work which had accuracy **50**. Since we have 2 classes, the previous work result is a guess (coin flip).

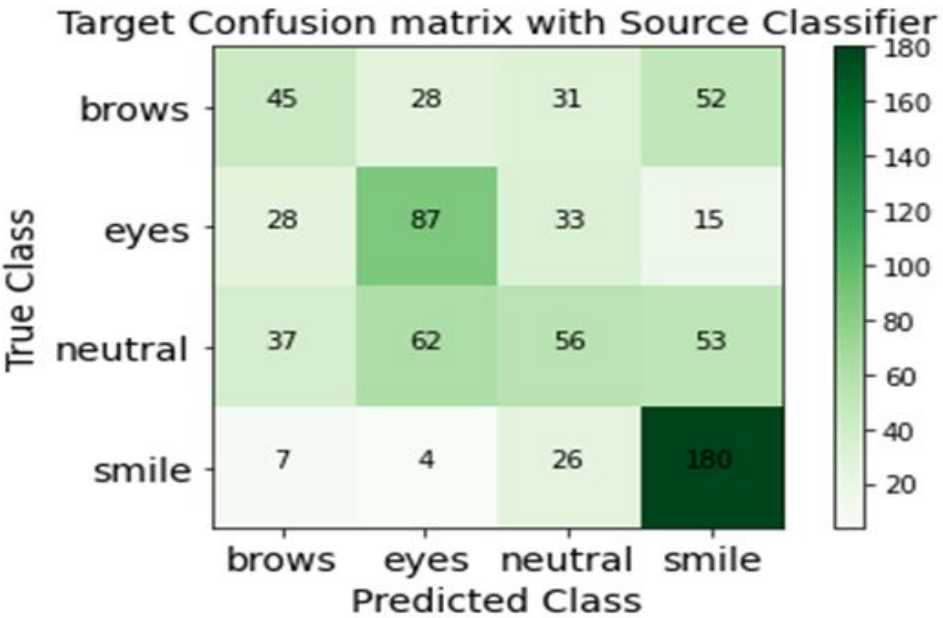

Figure 14 : Confusion matrix that allows visualization of the performance of our model. We can conclude that the model recognizes and generalizes smile very well but predicting on lifted brows did not generalize as well. Our result on target images is accuracy of 49.46% ,while previous work accuracy is 28.53%.

Class	**Smile**	**Neutral**	**Eyes**	**Brows**	**Average**
Adda target accuracy	0.8294	0.2692	0.5337	0.2884	48.01

Table 2: Adaptation results on the Students dataset, using images of 6 students from the train set as the source and another student images as target domains. We report here per class accuracy due to class imbalance.

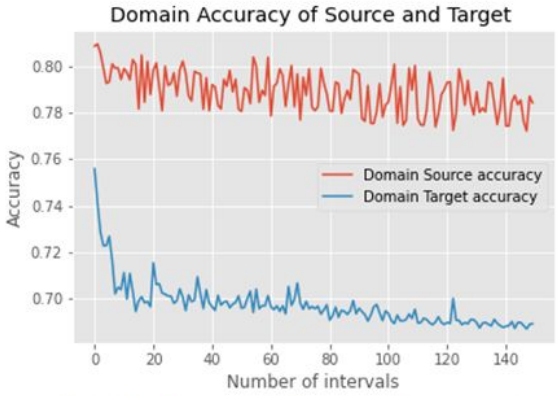

Figure 15 : This graph evaluates the accuracy of the discriminator on source images and target images, during training with 15,000 iterations. Using a pre-trained source model as an initialization for the target representation space and fixing the source model during adversarial training. Notice that the domain target is decreasing as the number of iterations increases, meaning we are successfully fooling the discriminator.

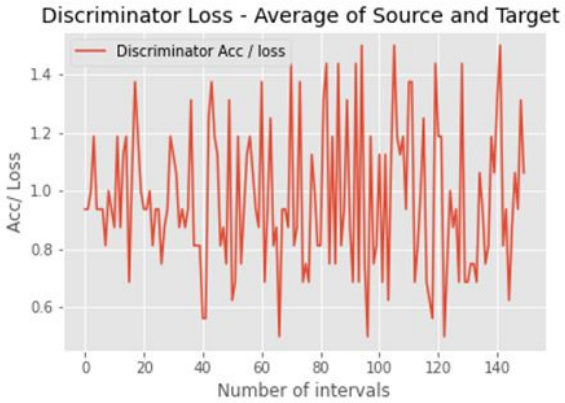

Figure 16 : This graph evaluates the average loss of the discriminator on source images and target images, during training with 15,000 iterations.

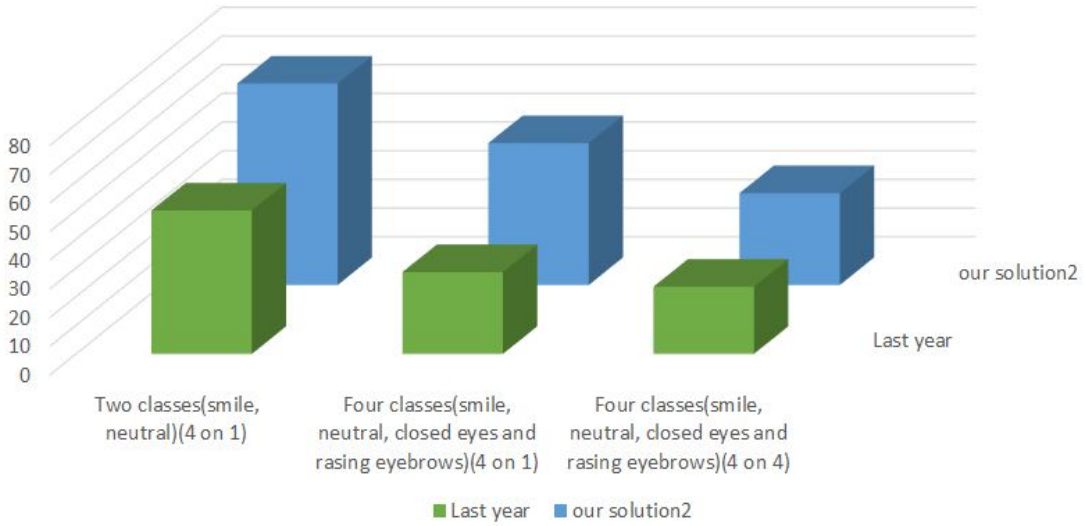

Figure 17: Final results - the dataset includes student images for both source and target domains. As you can see, we improved the results from previous work.

Alyn's Kids

We calculated the target accuracy using the source classifier on labeled images of Sharbel and Lishay. Figure 17 describes our results, while figure 18 describes the previous model's degenerated results.

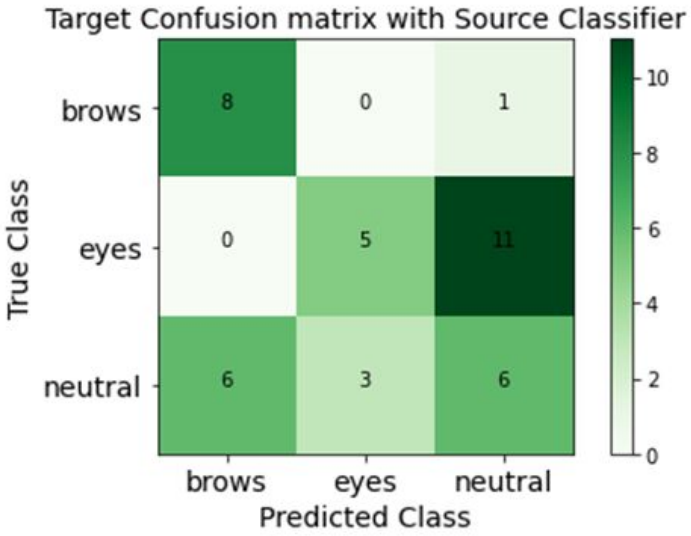

Figure 17 : Confusion matrix of target images using source classifier. We can see that the model recognizes brows very well in comparison to other facial expressions.

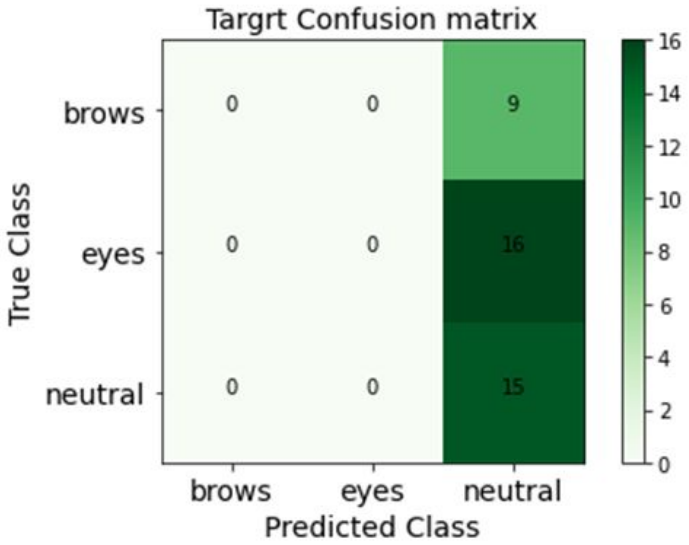

Figure 18 :The confusion matrix of target images on previous model. We can see that the model classifies all images as neutral expression.

Our result on test set of target images is accuracy *47.5%* compared to the previous work result which is *37.5%.*

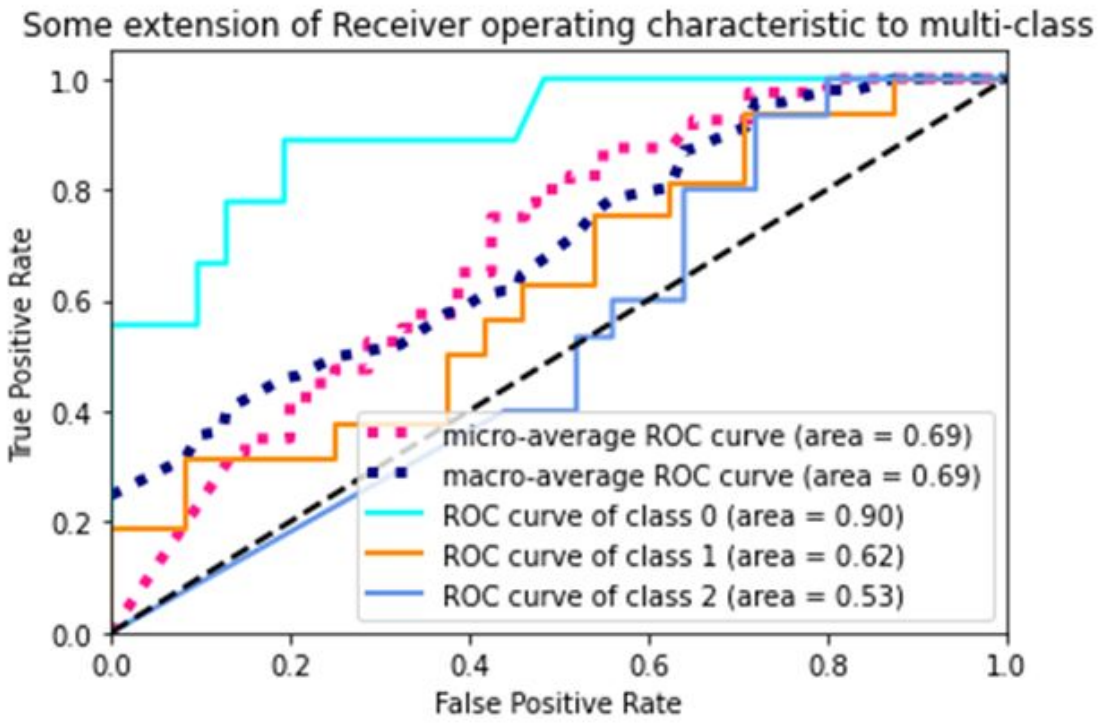

Figure 19 : Receiver Operating Characteristic (ROC) metric to evaluate classifier output quality. We computed ROC curve and ROC area for each class. We draw a ROC curve by considering each element of the label indicator matrix as a binary prediction (micro-averaging). Another evaluation measure for multi-label classification is macro-averaging, which gives equal weight to the classification of each label.

	One-vs-One ROC AUC scores	**One-vs-Rest ROC AUC scores**
macro	0.683943	0.683943
weighted by prevalence	0.649781	0.649781

Table 3: The multi-class One-vs-One scheme compares every unique pairwise combination of classes. In this section, we calculate the AUC using the OvR and OvO schemes. We report a macro average and a prevalence-weighted average.

Conclusion

We have proposed a unified framework for unsupervised domain adaptation techniques based on adversarial learning objectives. Those techniques were used for digit classification, but we used to recognize subtle facial expressions, which is considered to be a challenging task. We presented an evaluation across two domain shifts for our unsupervised adaptation approach. Our method generalizes well across a variety of facial expressions although some have better results and some less.

The goal we set ourselves is combining our system with the child's daily routine so that it will work at a high success rate. Since our project is an extension of a previous project, our main goal was to improve and extend the system they build.
Acquire multiple images per facial expression for every child using the previous system was not possible because the children's disabilities make it considerably difficult for them to perform facial expressions for the system. As a result, the hospital's staff did not use the system, leading to months of time and effort that went down the drain.
By omitting the data collection that was meant to be done by the medical staff, we achieved a goal we set ourselves to make the system user friendly. There is no need to capture images with a specific facial expression and tag it for training, but only record videos of the kids as they go about their day. To overcome the absence of labeled images of the kids we used domain adaptation methods as described above in the Method section.
The base model used in previous work might generalize well on people that look alike, but when having a large domain shift between the source and the target the results match a coin flip. When we first tested the generalization performances of our model using students' images, the results were promising. After using domain adaptation methods, we got better results, compared to the base model, in the binary case of two facial expressions - smile and neutral. To evaluate the model on the children that ought to use the system, we collected a small test set. This time we evaluated the model accuracy on three different facial expressions: neutral, closed eyes, and lifted brows. Compared to the previous work model which classified all the images as neutral, we achieved better results: accuracy of 47.5%, and micro average roc curve area 0.69.

For the purpose of sounding the relevant word to a facial expression that was recognized by the model, it is possible to use the GUI from the previous project. The user should record the voice for each child according to its meaning. Notice that facial expression could have a different meaning for each child and there are different facial expressions that each child is capable of doing. We might also use the previous project methods for recording and predicting online.

Future Work

A possible extension is adding a new feature to the system so it could record the user while he is using the computer for gaming or learning purposes. The camera will be working as long as the child uses the computer, and so we will collect more images of the child for a larger dataset. The dataset can be stored on a cloud so every month the classifier will train on the updated dataset. Notice that collecting data this way allows us to learn the child while he grows, so there is no need to collect new data every year and re-train the network on a larger dataset.

We can also explore the idea of a motion detection module that uses frame difference methods to detect the presence of motion in the frame. The current frame us compared with the previous frames pixel-wise. The difference in pixel values is noted and used to detect motion. This way we can also recognize head movements or pupil motion.

Project [3] aims at developing a security alert system. This system is based on motion detection and face recognition techniques in image processing.

A possible improvement is to make our system insensitive to changes in lighting or environments, by using a laptop camera to record the child with different lighting and background. We can add new images to the dataset and keep on training the model.

There are some new research directions such as Partial Adversarial Domain Adaptation, CoGan , Domain Confusion, and Gradient Reversal. As we described in Table 1 there are some adversarial domain adaptation methods with various properties that could be tested in the future.

This paper [2] from 2018 presents Partial Adversarial Domain Adaptation (PADA), which simultaneously alleviates negative transfer by down-weighing the data of outlier source classes for training both source classier and domain adversary, and promotes positive transfer by matching the feature distributions in the shared label space. Experiments show that PADA exceeds state-of- the-art results for partial domain adaptation tasks on several datasets.

Method	Office-31						
	A → W	D → W	W → D	A → D	D → A	W → A	Avg
ResNet [32]	54.52	94.57	94.27	65.61	73.17	71.71	75.64
DAN [7]	46.44	53.56	58.60	42.68	65.66	65.34	55.38
DANN [10]	41.35	46.78	38.85	41.36	41.34	44.68	42.39
ADDA [12]	43.65	46.48	40.12	43.66	42.76	45.95	43.77
RTN [8]	75.25	97.12	98.32	66.88	85.59	85.70	84.81
JAN [9]	43.39	53.56	41.40	35.67	51.04	51.57	46.11
LEL [25]	73.22	93.90	96.82	76.43	83.62	84.76	84.79
PADA-classifier	83.12	99.32	100	80.16	90.13	92.34	90.85
PADA-adversarial	65.76	97.29	97.45	77.07	87.27	87.37	85.37
PADA	**86.54**	**99.32**	**100**	**82.17**	**92.69**	**95.41**	**92.69**

Table 4: Accuracy of partial domain adaptation tasks. Office-31 is a most widely-used dataset for visual domain adaptation, with 4,652 images and 31 categories from three distinct domains.

Bibliography Referenced

1. Adversarial Discriminative Domain Adaptation Eric Tzeng University of California, Judy Hoffman Stanford University, Kate Saenko Boston University ,Trevor Darrell University of California, Berkeley . Our method is based on this article [1]
2. Partial Adversarial Domain Adaptation Zhangjie Cao, Lijia Ma, Mingsheng Long(B), and Jianmin Wang [2]
3. Security System Using Motion Detection and Face Recognition D.Aju1, Ashwani Agarwal2, Himanshu Jain3 , Divyanshu Bhati4 1,2,3,4SCOPE, VIT University, Vellore. [3]
4. Eye tracker [4]

Appendix

1. Our code on git : https://github.com/Shellyga/Adversarial-Domain-Adaptation-with-Keras/blob/master/4th_Year_Final_Project_Facial_Expressions_Recognition.ipynb
2. We present some explanations, definitions, and methods that will be useful in our implementation.
    1. Domain Adaptation -
       This scenario arises when we aim at learning from a source data distribution a well-performing model on a different (but related) target data distribution.
       In unsupervised domain adaptation, the learning sample contains a set of labeled source examples and a set of unlabeled target examples.

    2. Adversarial learning methods are a promising approach to training robust deep networks and can generate complex samples across diverse domains. They also can improve recognition despite the presence of domain shift or dataset bias: several adversarial approaches to unsupervised domain adaptation have recently been introduced, which reduce the difference between the training and test domain distributions and thus improve generalization performance.

       Some methods have chosen an adversarial loss to minimize domain shift, learning a representation that is simultaneously discriminative of source labels while not being able to distinguish between domains. Some proposed adding a domain classifier (a single fully connected layer) that predicts the binary domain label of the inputs.

       The Generative Adversarial Network (GAN) method is a generative deep model that pits two networks against one another: a generative model that captures the data distribution and a discriminative model that distinguishes between samples drawn from the generative model and images drawn from the training data by predicting a binary label. The networks are trained jointly using backprop on the label prediction loss in a mini-max fashion: simultaneously update a generator to minimize the loss while also updating discriminator to maximize the loss. So the generator is trained to produce images in a way that confuses the discriminator, which in turn tries to distinguish them from image examples from different domain distributions.
       In domain adaptation, this principle has been employed to ensure that the network cannot distinguish between the distributions of its training and test domain examples.
       The gradient reversal algorithm also treats domain invariance as a binary classification problem but directly maximizes the loss of the domain classifier by reversing its gradients. Gradient reversal ensures that the feature distributions over the two domains are made similar (as indistinguishable as possible for the domain classifier), thus resulting in the domain-invariant features.