POLITEHNICA UNIVERSITY OF BUCHAREST

Doctoral School of Electronics, Telecommunications and Information Technology

Decision No. 846 **from** 09-06-2022

Ph.D. THESIS SUMMARY

Eng. Mihai-Sorin BADEA

DEZVOLTAREA SISTEMELOR INTELIGENTE DE INTERFAŢARE VIZUALĂ OM-MAŞINĂ

THE DEVELOPMENT OF INTELLIGENT SYSTEMS FOR VISUAL HUMAN-MACHINE INTERFACES

THESIS COMMITTEE

Prof. Dr. Ing. Bogdan IONESCU Politehnica Univ. of Bucharest	President
Prof. Dr. Ing. Constantin VERTAN Politehnica Univ. of Bucharest	PhD Supervisor
Prof. Dr. Ing. Cătălin-Daniel CĂLEANU Politehnica Univ. of Timişoara	Referent
Prof. Dr. Ing. Romulus-Mircea TEREBEŞ Technical Univ. of Cluj-Napoca	Referent
Prof. Dr. Ing. Corneliu Nicolae FLOREA Politehnica Univ. of Bucharest	Referent

BUCHAREST 2022

Table of contents

Chapter 1

Introduction

1.1 Motivation

The last decades have brought a spectacular development in terms of the processing power of PCs and their interconnectivity through the Internet. This progress has led to the gradual but constant integration of software applications in most aspects of everyday life. At the same time, Artificial Intelligence came back into public attention because of a series of notable improvements in the field of Machine Learning. An important reason for this resurgence is represented by the fact that most data is now stored in a digital format. Even though the impressive results of the algorithms are not entirely explained by our current theoretical models, their applicability cannot be ignored, so that the understanding of the behavior of various popular algorithms became crucial.

Notions specific to Artificial Intelligence are overlapped with concepts from other fields, and this becomes especially clear with the development of high performance applications. Computer Vision has been dramatically changed by Machine Learning, the introduction of Convolutional Neural Networks being a milestone in recent history. Taking into account the nature of the experiments which will be presented in the paper, theoretical aspects will often be presented from the perspective of image processing.

Frequently, Machine Learning algorithms have to solve tasks which are not especially difficult for humans. However, there are numerous cases where, even for a person, prior training would be necessary. For example, tasks related to the medical field fall into this category. Also, advanced image editing isn't simple even when the user has access to highly specialized software tools. Moreover, if the significance of the data is unclear or the number of seemingly unrelated variables is large, automatic solutions can easily outperform humans.

1.2 Objectives

Based on their impressive results, Convolutional Neural Networks have become a staple in Computer Vision tasks. Their training process, however, is a difficult one, highly demanding in terms of computing power and requires a careful consideration of both quantity and quality of data. Based on this fact, a part of the approaches presented in the paper will target the way in which the data is used by the algorithm. Also, multiple supplemental loss functions will be introduced, to aid the performance of neural architectures.

From the perspective of the application domains of interest, a part of the paper is dedicated to the automatic analysis of paintings, while another one focuses on facial expressions. Even though some of the proposed solutions take into account certain particular aspects of the tasks, some can be used in general Machine Learning settings.

1.3 Structure of the paper

The next few chapters will focus on the theoretical aspects which are necessary later on. After presenting some general aspects related to Machine Learning, a closer look will be taken at Convolutional Neural Networks. Taking into account the importance of semi-supervised learning in the experiments, a chapter is dedicated to them.

The first series of applications is centered on the analysis of paintings, in Chapter 5. Afterwards, the experiments will focus on the facial expression analysis. The paper ends on Chapter 7, where the conclusions are presented.

Chapter 2

Machine Learning notions

The field of Machine Learning (ML) is developing continuously, the last couple of decades being a period of significant improvements in terms of performance. The main goal of ML is to study algorithms which are designed to solve a well-defined task, without requiring explicit programming of the system based on the particular issue of interest.

Since the beginning of the field, numerous solutions have been proposed and revised. Some of the most relevant ones, which were especially important to the experimental sections, will be presented below, alongside other theoretical aspects.

2.1 Types of learning

From the point of view of learning types, multiple categories of ML algorithms can be defined, the most important being: supervised learning, unsupervised learning and reinforcement learning. The first type requires that the used data samples X have one or more associated labels y, representing the target response from the ML system. The unsupervised category assumes that there are no labels, while reinforcement learning introduces the requirement of a temporal dimension, since the algorithm is in continuous interaction with its environment.

2.2 Image descriptors

The data samples used by ML algorithms are represented in a numeric form, usually as an array of *features*. In the case of images, the data is described by the values of the pixels they contain. In most algorithms however, these original values are processed to get higher quality features (descriptors).

Designed initially for texture analysis, Local Binary Patterns (LBP) [1], [2] have become popular in many ML tasks. In their original form, LBP descriptors process the V8 neighborhoods around each pixel and compare each value to the central pixel, to generate an encoding. They are then aggregated into a histogram, which is the final descriptor.

A particularly popular descriptor, the Histogram of Oriented Gradients (HOG) [3] was originally designed for pedestrian detection. Even though there are many variables which ultimately influence the quality of the descriptor, HOG features are always computed for different blocks in the image.

The Deep Convolutional Activation Feature descriptor (DeCAF) [4] is differentiated from the previous ones because the values are computed by using the image of interest as input for a convolutional architecture. The network is trained on a generic task and the descriptors are represented by the activations extracted from an intermediate layer, close to the output. This approach has proven to be a competitive choice, surpassing other well established descriptors in various experiments.

2.3 Machine Learning algorithms

Once the descriptors are extracted using methods such as the ones presented above, they need to be processed by a learning algorithm. In the following paragraphs, some of them will be presented, noting that convolutional solutions will be discussed separately.

Conceptually, decision trees are a simple Machine Learning solution. In each node of the tree which is not a leaf node, the feature space is divided. The final output of the system is determined by the value of the leaf nodes. To overcome some of the performance issues specific to decision trees, an ensemble solution called Random Forests (RF) was created [5]. Each tree in the RF is created with a degree of randomness to ensure diversity.

The way a Support Vector Machine (SVM) works is to find to optimal separating hyperplane. Even though in their original implementation SVMs can only solve linearly separable problems, the introduction of the *kernel trick* allowed the algorithm to be used in a considerably greater variety of tasks.

For a long time, the Multilayer Perceptron (MLP) was the most common form of neural networks. The general structure consists of an input layer, an output layer and at least one hidden layer. The value of each neuron in a layer is computed as a linear combination of all the neurons in the previous layer, on top of which an activation function is applied in order to introduce a nonlinear behavior.

Unlike the previously discussed methods, the k-Means algorithm [6] falls into the unsupervised learning category. Its main purpose is to group data together, by partitioning the feature space. To achieve this, the training phase consists in finding the k centroids for which the distance between the samples and the closest centroid is minimized.

Chapter 3

Convolutional Neural Networks

The most popular ML solutions currently used for Computer Vision are represented by Convolutional Neural Networks (CNN). They are made up of various layers, each implementing a certain operation, in order to process the input image and get the wanted output. In the following subsections, the main layer types and some important architectures will be presented.

3.1 The layers of a Convolutional Neural Network

Even though there isn't a fixed structure for the various architectures, the same layers are usually employed. A series of studies related to their behavior have led to various improvements which brought a boost in performance, usually without increasing the number of trainable parameters.

The most common layer is the convolutional one. Using the convolution operation, a part of the shortcomings typical for MLP layers are alleviated. This type of layer does not require connections between all the neurons of consecutive layers, visibly decreasing the computational load. Furthermore, using a relatively small number of weights belonging to a kernel the whole previous layer can be processed.

Going forward through the network, the precise positions of the features become less important, in some cases actually hindering training. Because of this, most convolutional architectures use a form of subsampling in the form of *Pooling* layers. They process the various neighborhoods of features belonging to the input feature maps and replace the values in each neighborhood with a single value representing a relevant statistic. The most common pooling layer is *max-pooling*, where only the largest value of a neighborhood is kept.

To prevent the unwanted behavior of overfitting the training data, some special regularization layers have been introduced. One such layer is *Dropout* [7], which nullifies with a given probability the neurons from the previous layer during training.

Another example is the *Batch Normalization* layer (BN) [8], which learns the general distribution of the activations used as input and normalizes them.

3.2 Training neural networks

The training process of CNNs is similar with the one used for MLPs. A batch of data is used as input for the network, a loss based on the predictions is computed, and an *optimizer* determines the changes which need to be applied on each weight. The adjustment of the weights is done using the *backpropagation* algorithm.

The loss computed using the network predictions is a measure of their quality, when comparing them to the labels, in supervised scenarios. For regression tasks (in which the label is a continuous value), the most common loss functions are the Mean Absolute Error and the Mean Squared Error functions. When considering classification tasks, the output is usually modelled as a probability distribution, while the loss is the Cross Entropy function, the ideal distribution being dictated by the sample's label.

Based on the value of the loss function, the optimizer adjusts all the weights in the network. A first optimizer is the *Stochastic Gradient Descent* algorithm. An important research topic targeted possible improvements, which led to solutions such as Adam [9].

3.3 Convolutional architectures

There is a significant number of architectures which have brought notable elements of novelty in the field of CNNs. In the following paragraphs, some of the most relevant architectures from the literature will be presented.

The AlexNet architecture [10] had impressive performances when it first appeared. Some of its most notable features are the use of ReLU activations, and the highly parallelized training using two GPUs.

After AlexNet was established as a reference, an increased interest for using more neural layers was noticed. The VGG family of architectures [11] explores this, displaying networks with up to 19 layers. They used kernels of size 3×3 with a stride of 1 exclusively, alongside *max-pooling* layers.

The increasing number of layers found in architectures was an issue, because the training process is affected by the *vanishing gradient* problem. To solve this, ResNet architectures [12] use a shortcut connection, which adds the input values from a block of convolution operations to their output.

Furthering the ideas introduced by ResNet, Hourglass (HG) architectures [13] suggest a different configuration for convolutional blocks. Each hourglass module consists of a block of convolution and *max-pooling* operations, followed by upsampling areas used to return to the original working resolution. These two types of areas are then connected through a shortcut connection, like in ResNet.

Chapter 4

Semi-supervised learning notions

Besides the previously discussed learning types, there are some other notable methods, with various practical uses. Semi-supervised learning (SSL) combines aspects from both supervised and unsupervised learning, and it consists of using both labeled and unlabeled samples to increase the performance of the base supervised system. While the usage of labeled samples is straightforward, the challenge of SSL algorithms comes from incorporating the unlabeled data in a meaningful fashion. A typical classification of these methods separates the approaches which minimize entropy, in which the network outputs are used directly, and the ones based on regularization techniques [14].

4.1 Formal assumptions of semi-supervised learning

In order to use SSL algorithms, there are some assumptions which need to be made, related to the structure of the training data [15]. First of all, if two samples X_1 and X_2 are similar, then it is expected that they belong to the same class (in the case of a classification task). Second, the boundaries between classes should be in areas in which there is a low density of samples. Third, it is assumed that the high dimensional space in which the data reside can be modelled as a series of lower dimensional manifolds. In this case, samples which belong to the same manifold should also have the same label [16].

4.1.1 Examples of algorithms

There are many SSL algorithms proposed by the literature, each with their own set of advantages and innovations. In this subsection, some of the algorithms which have been relevant for the experimental stages will be presented.

4.1.2 Pseudo-labels

A simple method of using unlabeled data is *Pseudo-labels* [17], which is designed for classification tasks. It consists of using the unlabeled samples in a second Cross Entropy loss function. The required labels are computed using the neuron with the maximum value from the output layer.

4.1.3 Mean Teacher

While Pseudo-labels uses unlabeled data similarly to the labeled part, the *Mean Teacher* [18] algorithm takes a different approach. A *student* network is used alongside a *teacher* (computed as the exponential moving average of the student network), and a new loss function is computed between them. The unlabeled samples are used in this new term, which is computed as the Mean Squared Error between the predictions of the two networks.

4.1.4 MixMatch

The structure of the *MixMatch* [19] algorithm is considerably more complex than the previously mentioned algorithms. The unlabeled samples are grouped with the labeled ones and new images are generated using *MixUp*. Based on the type of the image with the higher weight, the generated samples are used either in a Cross Entropy loss, or a Mean Squared Error loss.

4.2 Data augmentation

A general solution to increase network performance for supervised tasks is the use of augmentation techniques, which are applied on the input images. These operations are of much greater importance in SSL algorithms, so a few relevant examples will be provided.

The most basic types of augmentations come from image processing. Translations, rotations and flips are commonly used in the image preprocessing step. Besides these operations, other common augmentations consist of adding Gaussian noise to the image, or even cropping the input, if the source image has a higher resolution than the one the network requires.

The *MixUp* algorithm [20] is a popular augmentation choice, both because of its results and because of its simple implementation. Given two samples, a new image is generated as a linear combination of the two, using a mixing factor α. The same combination with the same α is then used for the label distributions of the two input samples.

Chapter 5

Research direction 1: Painting analysis

Typical Machine Learning tasks such as object detection usually imply processing visual information represented realistically, like the real situations in which people find themselves every day. However, humans are much more capable than automatic systems when it comes to identifying and analyzing important elements in a scene, being able to easily understand more abstract visual information. This aspect has allowed the development of visual art, in which an artist creatively utilizes various representations in order to attract the public's attention or to transmit deeper meanings.

In order to achieve significant results in the field of painting analysis, large collections of artworks need to be available in digital format. Projects such as WikiArt[1] and Art UK[2] have made hundreds of thousands of paintings easily accessible through the Internet. It should be noted that each painting has multiple labels attached to it, where it was possible. In the case of the current paper, the most important label was the *genre*. The meaning of each category is not always obvious, in some cases referring to the subject, while in other cases the manner of representation is the defining aspect. The existing overlap between the notions of genre and scene type allowed non-artistic images to be compared, to some degree, with paintings, in terms of this label.

The experimental stage which will be presented in this chapter aims to increase the performance of convolutional architectures trained to classify paintings by genre. After establishing baseline results, domain adaptation methods were employed, through the use of style transfer, to extend the size of the training dataset. Styles with a higher degree of abstraction will be treated separately.

5.1 Related Work

Research in the field of the automatic analysis of artworks has existed before CNNs were established as the most common ML approach, in works such as [21], and in many

[1] http://wikiart.org

[2] http://artuk.org

cases, the task was style recognition ([22], [23]). Before WikiArt was introduced, the commonly used datasets were considerably smaller, having under 2000 samples ([24], [25]).

In some cases, the authors approached both genre and style recognition of the paintings found in WikiArt, like in [26], where a pre-trained AlexNet architecture was used. From an adjacent perspective, efforts such as the one made by Zhou et al. [27] should be mentioned, where the authors studied scene recognition in photos, a subject which is similar to genre recognition.

5.2 The analysis of artworks in terms of domain adaptation

Domain adaptation is a method which proved to be effective when training neural networks. Using this approach, the size of the training dataset can be increased, the inner workings being described by Ben-David et al. [28]. Domain adaptation requires a source domain to be defined, alongside a target domain and a domain adaptation function. In the current case, the source domains were represented by various sets of photographs (either artistic or non-artistic), and even a subset of paintings from WikiArt. Considering the task, the considered domain adaptation functions are style transfer methods.

5.3 Style transfer algorithms

Methods which can be used to adjust the style of an image according to a reference have existed for enough time that multiple different solutions can be considered as important references. Departing from algorithms such as the one proposed by Reinhard et al. [29] or [30], approaches which operate with features at different scales stand out, such as the Laplacian [31] and the neural [32] style transfer algorithms.

The first style transfer algorithm to be used in the paper operates with a pyramidal representation of an image to achieve the transfer at multiple levels of detail. Using Laplacian filtering, at each level of the pyramid, the gradient distributions are matched between two images. Given the content source C and the style source S, for each pixel p belonging to a neighborhood v at each level of the pyramid the following computation is made:

$$\begin{aligned} C_n(p) &= r(C_{n-1}(p)) \\ r(i) &= g + sign(i-g)t(|i-g|) \\ t(i) &= F^{-1}_{\nabla S(p)} F_{\nabla C(p)}(p) \end{aligned} \tag{5.1}$$

The CNN-based style transfer algorithm proposed by Gatys et al. [32] aims to use the neural features computed at multiple levels of the forward propagation in different ways to describe the style or the content of images. Using features originating from lower levels of the network leads to generating an image which is very similar to the content source. However, if higher level features are employed, pixel level similarity decreases, while coarser spatial arrangements are kept. Similarly, for style description, the first layer are closely related to very fine details, while top layers can be used to capture higher level ideas.

The actual transfer process requires initializing a new image X with white noise, which is then adjusted using an iterative process in order to resemble C and S. The loss function separates the style and content components, which are designed according to the specifics of each type of image, the final value being computed as a linear combination of the partial loss terms (5.2).

$$L_{total}(C,S,X) = \alpha L_{content}(P,X) + \beta * L_{style}(S,X) \tag{5.2}$$

Matching the content source image C is done by minimizing the squared errors between the network activations obtained by using C and X as inputs (5.3). Style related information is computed using the Gram matrices of the activations (the dot product of two vectorized feature maps), the loss function being the squared error between the matrices corresponding to X and S at various levels in the network (5.4).

$$L_{content}(C,X,l) = \frac{1}{2}\sum_{i,j}(C^l_{ij} - X^l_{ij})^2 \tag{5.3}$$

$$E_l = \frac{1}{4N_l^2 M_l^2}\sum_{i,j}(G^l_{ij} - S^l_{ij})^2 \tag{5.4}$$

The way that the Laplacian and neural style transfer algorithms work is clearly different, but some common aspects can be noted. Both methods use filtering in order to get intermediate features, but only the neural algorithm uses learned filters, which extract increasingly complex information as the network progresses. Also, both methods operate at multiple levels of detail, either by means of a pyramidal decomposition or by using the pooling layers of a CNN.

5.3.1 Neural style transfer acceleration

In order to generate a large enough additional dataset with stylized images, long processing times are required, since generating a single pseudo-painting may take up to 20 minutes. This issue was addressed in [33], where some possible solutions were investigated. One of them consisted in changing the neural architecture from a VGG-16 or VGG-19 to a ResNet-50. After searching for an optimal combination of layers, it was found that the resulting images were visually unsatisfactory.

The second method explored in order to reduce the required generation time for a large number of pseudo-paintings consisted in finding an adequate time for an early interruption of the optimization process. After a series of verifications, a threshold of 100 iterations was chosen as an upper limit for the process.

5.4 Datasets

In the following paragraphs the datasets which were considered for the genre recognition experiments will be presented. These were made up of either paintings, photographs (artistic or non-artistic) or pseudo-paintings generated using the other datasets.

5.4.1 WikiArt

The impressive size (>250,000 paintings) and the diversity of artworks featured in WikiArt have made this dataset especially relevant for painting analysis tasks. A subset of over 80,000 paintings, used in previous studies such as [23] was chosen for the experiments, the results being reported in [34] and [35]. In a data verification stage, some of the images where discarded because they lacked the required labels, leaving 79,434 usable artworks.

For the task of genre recognition there are 42 available classes, but some of them did not feature enough samples (<200), which led to grouping them in a separate class named *Others*. The remaining 25 genres considered sufficiently well represented are: "Abstract Art", "Allegorical Painting", "Animal Painting", "Battle Painting", "Cityscape", "Design", "Figurative", "Flower Painting", "Genre Painting", "History Painting", "Illustration", "Interior", "Landscape", "Literary Painting", "Marina", "Mythological Painting", "Nude Painting", "Portrait", "Poster", "Religious Painting", "Self-Portrait", "Sketch and Study", "Still Life", "Symbolic Painting", "Wildlife Painting". An immediately noticeable difficulty is the overlap between classes, such as "Portrait" and "Self-Portrait", or "Allegorical Painting" and multiple other genres.

5.4.2 Photographs datasets

Based on the similarities between the idea of a painting's *genre* and its scene, the experiments have used the *Scene UNderstanding* (SUN) [36] dataset as a first source of extra data. From the 899 classes featured in SUN, 61 were chosen as being of immediate interest for increasing genre recognition. Considering the importance of the "Portrait" class, 5993 images were selected from the *Labeled Faces in the Wild* (LFW) [37] dataset.

An important aspect when evaluating any painting is its style, which means that images from datasets such as SUN might be insufficient for significant improvements. Although the concept of style is different for photographs than it is for paintings, the attention given to the way in which content is presented might be important for the target task, leading to the selection made by Thomas and Kovashka in [38] to be chosen for the experiments. Starting from the over 180,000 available photographs, almost 20,000 adequate samples were selected.

5.4.3 Stylized images

The samples belonging to this category represent the result of combining images from the previously mentioned datasets. Two main cases were considered, which led to new images being generated: style transfer from painting to photographs (using either the Laplacian or the neural algorithm) and style transfer from paintings which featured a high level of abstraction to paintings which have content represented realistically.

5.4.4 Experiments

The results reported for genre recognition used a subset of WikiArt and had as a starting point a set of initial experiments which were considered as reference, followed by a series of tests involving different strategies designed to improve classification quality by augmenting the training dataset. Besides them, the impact of style on the quality of genre prediction was also assessed. In previous works such as [39], only the 10 most well represented classes were considered. Because of this, an initial comparison was done to take this case into account. The test set was constructed by randomly selecting 20% of the images from the dataset, this being the same ratio used by Karayev et al. [23], although other authors preferred different proportions.

The results used as reference and the ones resulting from augmentations with classical techniques are reported in Table 5.1. Besides CNNs, other approaches, which required a separate feature extraction stage followed by a classifier, were evaluated. The two results corresponding to ResNet-34 architectures and the initial training dataset are differentiated by the hyperparameters used in [34], as opposed to [35]. The table shows that convolutional approaches have always led to better performance. The best results are reported for the case in which the dataset was significantly increased by flipping and rotating images. This scenario also required a significantly longer train time, so the style transfer was only used considering the initial image set.

5.4.5 The impact of style on classification performance

When the abstraction level of the content of a painting is increased, it is expected that the performance of the CNN should be negatively affected. Starting from this idea, the dataset was divided such that only paintings belonging to the *Cubist* and *Naive Art* styles be present in the test dataset (4,132 samples). Even though the training set was

Method	Classes	Images	Test ratio	Accuracy [%]
Saleh and Elgammal [39] Classemes + Boost	10	63,691	33%	57.87
Saleh and Elgammal [39] Classemes + ITML				60.28
Tan et. al [26] AlexNet			-	69.29
Tan et. al [26] CNN (pretrained)				74.14
ResNet-34 [34]			20%	**75.58**
pLBP + SVM [34]	26	79,434	20%	39.58
pHOG + SVM [34]				44.37
DeCAF + SVM [34]				59.05
AlexNet [34]				53.02
ResNet-34 [35]				59.1
ResNet-34 [34]				**61.64**
ResNet-34 and augmented dataset [34]				**63.58**

Table 5.1 Results for the genre recognition task. The values from experiments reported in [34] and [35] are compared with the state of the art.

noticeably larger in this case, the classification accuracy dropped by more than 10% to 50.82%, while the Top-5 accuracy dropped by approximately 8%.

Another proof of the importance of style was provided by a supplemental analysis of the base results. For classical styles, the reported results are good, while styles such as *Naive Art*, *Cubism* and *Surrealism* display noticeably worse results. Some exceptions were noted, such as *Minimalism* and *Color Field Painting*, where the content was unique for each style.

Transferred image type	Accuracy reported to the maximum number of images per class					Average improvement
	250	500	1000	5000	All	
None	19.12	27.95	35.66	54.61	61.64	0
Non-artistic photos	25.28	32.04	38.21	55.31	61.67	2.71
Artistic photos	26.72	**32.75**	39.27	53.49	61.55	2.96
Laplacian transfer on artistic photos	26.56	30.73	39.84	56.17	**61.73**	**3.21**
Neural style transfer on paintings	**26.76**	31.98	37.4	55.33	61.47	2.79
Neural style transfer on artistic photos	26.33	31.27	**39.94**	54.88	61.62	3.01
Max. improvement	7.64	4.8	4.18	1.56	0.09	—
Added images [%]	190.04	103.38	58.98	27.33	≈26	—

Table 5.2 The results of the domain transfer experiments. The row marked with *None* represents the base case, with no added images. The *Average improvement* column contains the average difference in performance for all iterations of a single type of added images. Results reported in [34].

5.4.6 Domain transfer

In order to gather the necessary images for domain transfer, multiple lengthy stages of selection or data generation were required. In the end, for this experiment, more than 30,000 pseudo-paintings were considered (generated with the neural style transfer algorithm) alongside 20,000 artistic photographs. The experiments were conducted in an iterative fashion, each iteration limiting the maximum number of images which can be used for each separate genre. Since each supplemental source consisted of a different number of samples for each class, the number of usable images was fixed for each genre. This resulted in using 2903 images for "Cityscape", 4467 for "Landscape" and 4002 for "Portrait". The results reported in Table 5.2 showcase a series of interesting cases. For starters, in every case, most of the performance comes from the paintings found in WikiArt. This trend can be attributed to the fact that paintings should have a distinct feature in order to be considered artistically relevant. Furthermore, using the entire WikiArt dataset negates any form of improvement brought by domain transfer. In order to notice any performance gain, the number of supplemental images should be at least half of the number of paintings. A more surprising result is represented by the fact that the neural style transfer algorithm had worse results when compared to the Laplacian approach.

5.5 Conclusions

Based on the experiments performed in this chapter, two main contributions are noted. First of all, the number of reported results is large, regardless of the considered stage (base analysis or domain transfer experiments). This way, a series of limitations were highlighted in relation to the usefulness of the neural style transfer as a domain adaptation function, even though the resulting images are visually interesting. Secondly, the impact of artistic style on genre recognition was assessed. Even though abstract styles are problematic for convolutional architectures, it should be noted that even humans have a hard time analyzing these kinds of paintings.

Chapter 6

Research direction 2: Facial expression analysis

The importance of interpersonal communication is undeniable in the well-running of society. There are three major components of communication, each being targeted by specific ML algorithms. The first component is the verbal one, which refers to choosing the right words to send a message. The second component is the non-verbal one which, even though it does not have the same direct informational load, offers valuable information related to the emotional state or the intentions of the speaker. By its very nature, this component requires the use of image analysis algorithms. The last component of communication, the paraverbal one, encompasses notions referring to the way in which words are spoken: emphasis on some words in a phrase, the evolution of the tone of the voice, the speed of speech, etc.

The current chapter focuses on studying the non-verbal component and its purpose is to enhance the quality of facial expression analysis algorithms. To achieve this, it was necessary to introduce some ML elements specific to the given problem, alongside some notions specific to psychology. In the following sections, a series of theoretical aspects will be explained, some particularities of using neural networks for facial analysis tasks, followed by presenting the experimental stage.

6.1 Related Work

The importance of the non-verbal component in communication is noticeable in the literature, the volume of studies referring to facial expression analysis being truly impressive. Even though there are multiple ways to approach the issue, the most popular version is to divide facial expression into 6 discrete classes, as proposed by Paul Ekman [40]: *happiness*, *sadness*, *anger*, *fear*, *surprise* and *disgust*. Besides these, the *neutral* state is considered (the lack of any expression), and sometimes the set is extended by adding the *contempt* expression. A significant number of high performing methods which appeared in the last decades are presented in papers such as the ones written by Căleanu [41] and Sariyanidi et al. [42]. Lately, solutions are oriented towards convolutional approaches, such as Kuo et al. [43], in which the authors aimed for both competitive results and a reduced memory and processing power consumption. Multiple convolutional methods are found in the study conducted by Li and Deng [44]. From the perspective of alternatives to the model with six expressions, works such as the one conducted by Zhao et al. [45] are noticed, where *Action Units* (AU) are employed.

Looking at the solutions suggested in the literature, some trends may be noticed. As is the case for facial recognition algorithms, besides using Cross Entropy, which is specific for classification tasks, supplemental loss functions are often added. Also, a method intended to enhance network performance is to use multiple types of labels (expressions and AUs) during training, taking advantage of the connection between them.

6.2 Facial expression analysis methods

Both in ML and in psychology, modelling facial expressions in an objective fashion is difficult, which is why, over time, various representation systems were introduced. The first one which needs to be mentioned is the one proposed by Ekman [40], which assumes that there is a set of facial expressions which are displayed in the same manner in every culture (listed in the previous subsection and displayed in Fig. 6.1).

A system presented as an alternative to basic expressions is represented by Action Units, which propose a more granular and objective approach to the problem. Introduced by Ekman and Friesen [46], this system defines a series of codes to various movements of the facial muscles. Even though there are 27 such AUs [47], when training neural networks only the most common 10-12 movements are taken into account.

Besides the two options mentioned above, other methods have been defined, such as dimensional models. They include multiple systems in which the analyzed expression can be represented as a point in a coordinate system, most often in a two-dimensional or three-dimensional space. The most common example is Russell's two-dimensional system [48], in which the axes represent the notions of "valence" and "arousal".

The most popular way of analyzing facial expressions is still the theory of universal expressions, even though all three options presented above have their own strengths and weaknesses. The inclination to use this model can be explained by the fact that it is simple, and it allows for a less costly labelling process. In the experiments which will be discussed in this chapter, both this version and the AU-based one will be used.

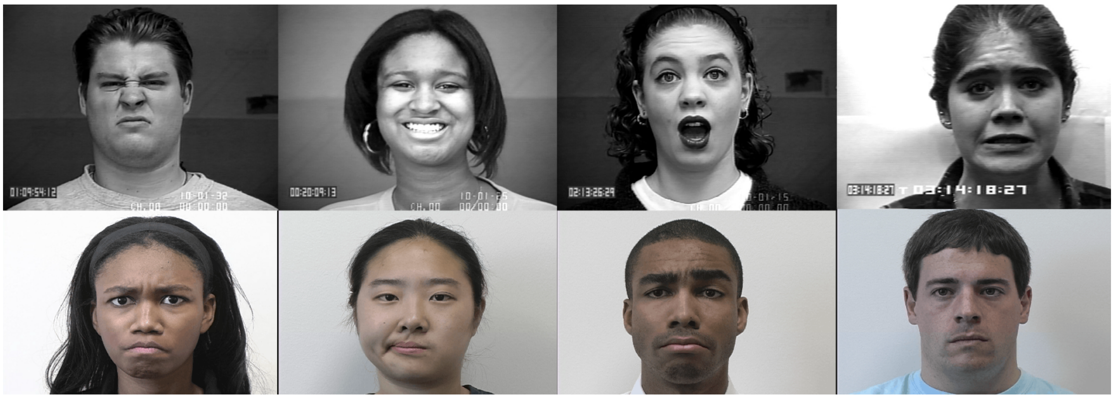

Fig. 6.1 The 7 universal expressions and the neutral one. From left to right, on the first row: *disgust*, *happiness*, *surprise*, *fear*. The second row: *anger*, *contempt*, *sadness*, *neutral*. Image taken from [49].

6.3 Particularities of facial expression analysis in Machine Learning

Any field of research in which ML algorithms can be used will bring a series of difficulties related to the problem's particularities. Facial expression analysis can be considered as being more difficult than other tasks, such as general object detection, because of the nature of the data. First of all, the composition of the images showcases a relevant aspect. If for datasets such as CIFAR-10 [50] each class is significantly different from the rest, facial expressions datasets have the same main subject: the human face. Moreover, if for the generic dataset the correct label can be confirmed in an objective manner, there will always be some degree of subjectivity in identifying difficult expressions.

When looking at another task related to facial analysis, face recognition, it was noticed that the usual Cross Entropy loss is outperformed by solutions which encourage generating more discriminative features [51]. This observation has led to introducing new, better performing, loss functions, two such functions being especially relevant to the proposed implementations.

6.4 Methods of preprocessing face images

The preprocessing stage is of special significance when considering facial analysis. Besides operations such as normalization, there are some domain specific procedures which can be noted. Face detection probably has the largest impact in the preprocessing pipeline, allowing for a better positioning of the subject in the frame. Additionally, some more adjustments can be made, using information about face orientation or the location of certain landmarks.

Face detection is one of the most studied subjects in Computer Vision. Because of this, many solutions have been proposed throughout time, along with various improvements to the base algorithms. From the multitude of available choices, the algorithm proposed by Paul Viola and Michael Jones [52], and the MTCNN architecture (*Multi-Task Cascaded Convolutional Neural Networks*) proposed by Zhang et al. [53] should be mentioned. The first method has used ideas such as the *integral image* to ensure the low processing times of the proposed cascade classifier. Even though the Viola-Jones algorithm continues to be a reference in the field, CNN-based algorithms have become, as in other cases, the preferred option. The MTCNN architecture is made up of three networks: P-Net (*Proposal Network*), R-Net (*Refine Network*) and O-Net (*Output Network*). Both algorithms have been used in the experimental stage.

Besides face detection, another important step in facial image analysis is *face alignment*. This step consists of extracting information related to the positioning of facial landmarks, in order to adjust the position and orientation of the face. The images used in the experiments which will be presented later have been processed using the algorithm proposed by Kazemi and Sullivan in [54], implemented in the DLIB library[1].

6.5 Center Loss

In order to encourage the generation of discriminative features, Wen et al. have created the function known as *Center Loss*, which can be used alongside the Cross Entropy function [55]. The new function minimizes the intra-class distances, by computing the distance between every sample x_i and the centroids of their respective classes c_{y_i} (6.1).

$$L_C = \frac{1}{2} \sum_{i=1}^{N} ||x_i - c_{y_i}||_2^2 \tag{6.1}$$

6.6 Island Loss

Center Loss has proven to be useful in an experimental setting, but some possible improvements have been identified. Proposed by Cai et al. [56], *Island Loss* extends the function, by adding an extra term which penalizes centroid proximity (6.2).

$$L_{IL} = L_C + \lambda_1 \sum_{c_m \in M} \sum_{\substack{c_n \in M \\ c_m \neq c_n}} \left(\frac{c_m \cdot c_n}{||c_m||_2 ||c_n||_2} + 1 \right) \tag{6.2}$$

6.7 Datasets

The variety of experiments which will be presented required using a large number of datasets with various labels related to facial expressions. Besides the multiple ways in which the labels can be used, choosing multiple datasets was necessary because in

[1] http://dlib.net

cases such as AU detection, data is considerably harder to gather, since annotators need to have a FACS (*Facial Action Coding System*) certification to ensure correct labeling. Additionally, images without labels related to facial expressions have also been used.

6.7.1 FER+

The FER+ dataset [57] consists of a subset of FER2013 [58], and it aimed to address some of the issues encountered with the original collection. In the initial version, the training set was made up of 28709 images downloaded from the Internet, all images presenting labels related to the universal set of expressions. Besides removing some of the images from the original dataset, FER+ also brought a complete relabelling of the images.

6.7.2 RAF-DB

Introduced by Li and Deng in [59], RAF-DB also features images from the Internet with universal facial expressions labels. This time, the 29,672 images were directly annotated by a larger number of users, each example been evaluated by 40 persons.

6.7.3 MegaFace

The purpose of the MegaFace dataset [60] was to provide to the scientific community a large collection of images (>1,000,000) containing a large number of unique identities, to be used for facial recognition tasks. Even though it does not feature labels related to facial expressions, the large number of available samples have made this dataset appealing for experiments which require an unlabeled portion.

6.7.4 CK+

Unlike the first two datasets, CK+ (*Extended Cohn-Kanade*) [49] is an important reference for algorithms intended to detect AUs. The 593 image sequences feature 123 subjects who were instructed to display various facial expressions, in laboratory conditions. Each sequence starts from the neutral pose, while the last frame displays an expression at maximum intensity.

6.7.5 EmotioNet

If, in previous cases, each set presented only one type of labels, EmotioNet [61] contains information about AUs and the discrete expressions shown in its images. The large number of samples (approximately 1,000,000) downloaded from the Internet would usually be an impediment for the labelling stage, which is why most images were annotated using powerful automatic systems. However, in order to ensure a solid reference, 25,000 images were manually labeled, to be used for the EmotioNet Challenge [62].

6.7.6 UNBC-McMaster Shoulder Pain Expression Archive Database

An important aspect which needs to be taken into account when analyzing facial expressions is the method used to create the dataset. If the images were found on the Internet, there is no guarantee that the expressions are genuine/spontaneous. This issue was tackled in the *UNBC-McMaster Shoulder Pain Expression Archive Database* (UNBC-McMaster) dataset, created by Lucey et al. [63]. The 48,398 images come from 129 subjects who had shoulder pains. They were requested to make certain motions with their arms, the images being afterwards annotated with AU labels.

6.7.7 Facial expressions of children

During the experiments some of the considered datasets featured children as the main subjects. A first dataset, CAFE (*Child Affective Facial Expression*) [64] has 1192 images, in which 154 children with ages between 2 and 8 years displayed various expressions. Additionally, the LIRIS dataset [65] was used, which contains 208 videos in which the 12 subjects (aged between 6 and 12 years) display spontaneous expressions.

6.8 Experiments

In the previous sections some usual difficulties of working with facial expressions were presented. The experiments which will soon be discussed have sought, for the most part, to bring improvements to the way in which CNNs perform by using unlabeled images.

6.8.1 The facial analysis loss function using pseudo-expressions

Loss functions such as Center Loss and Island Loss have proven their usefulness in generating discriminative features. In [66] the concept was extended, by encouraging a behavior similar to Center Loss in a scenario which allows for the use of semi-supervised learning.

The AlexNet architecture was trained for both the recognition of universal expression and detection of AUs in an image. In order to use loss functions such as Center Loss, the target classes need to be mutually exclusive, a condition which is not satisfied when working with AUs. Because of this, a set of pseudo-expressions were defined, by using the correspondences between AUs and fundamental expressions. In this way, for samples which only present this kind of labels a supplemental expression class was determined by combining the detected facial movements. The total cost is defined as:

$$L = \alpha_1 L_S + \alpha_2 L_M \tag{6.3}$$

Alongside the regular supervised loss L_S, an additional L_M loss is used, described by (6.4) in [66], where x_i represents features in fully connected layers. Images with AU labels were taken EmotioNet, while RAF-DB was chosen as the source for discrete expression labels. All unlabeled samples, approximately 311,000, where taken from MegaFace. The results on the EmotioNet dataset are found in Table 6.1.

$$L_M = \sum_{i=1}^{N} \left(\left\| \frac{x_i}{\|x_i\|_2} - \frac{c^j}{\|c^j\|_2} \right\|_2 - \frac{1}{C-1} \sum_{\substack{k=1 \\ k \neq j}}^{C} \left\| \frac{x_i}{\|x_i\|_2} - \frac{c^k}{\|c^k\|_2} \right\|_2 \right) \tag{6.4}$$

Method	Type	AU_1	AU_2	AU_4	AU_5	AU_6	AU_9	AU_{12}	AU_{17}	AU_{20}	AU_{25}	AU_{26}	AU_{43}	Subset avg.	Avg.
AlexNet [67]	S	24.2	-	34.7	**39.5**	73.1	-	86.8	-	-	88.5	45.6	-	56.1	-
AlexNet *Center Loss*[55]		**34.4**	30.3	55.3	33.3	69.1	**46.1**	79.3	27.8	**32.3**	84.4	43.2	**48.8**	57.9	48.8
AlexNet WSC[67]	SSL	25.3	-	34.5	39.3	**75.6**	-	**87.4**	-	-	**88.8**	47.4	-	57.0	-
AlexNet *Island Loss*[56]	T	30.4	29.5	**56.7**	30.6	66.7	44.1	77.3	26.7	23.8	83.9	47.3	43.9	56.14	46.7
AlexNet *Large Margin*[66]		34.1	**31.1**	56.6	33.9	71.0	45.1	78.1	**30.9**	25.3	83.8	**50.9**	47.2	**58.33**	**49.0**

Table 6.1 The F1 scores for detecting manually annotated AUs from EmotioNet. There are three considered training types: *Supervised (S)*, *Semi-supervised (SSL)* and *Transfer (T)*. *Subset avg.* refers to the average score reported for the subset of the most common AUs: 1, 4, 5, 6, 12, 25 and 26. Results also reported in [66].

6.8.2 Improving small networks using SSL

Another stage of the experiments consisted in running a series of tests, starting from the idea of improving the classification performance of small CNNs [68]. For this purpose, three methods were considered, using architectures with few convolutional layers (2 or 3). The first method was a reimplementation of a modified version of the Π-*model* [69]. The second one added an extra reconstruction loss term, focused on a separate autoencoder branch defined in the fully connected area of the network, while the last approach sought to minimize the distance between the features of an unlabeled sample and its closest labeled neighbor. All three methods were trained and tested on the CIFAR-10 and FER+ (Table 6.2) datasets.

6.8.3 Margin-mix

While the first two approaches used unlabeled images separately, the Margin-mix algorithm, described in [70], suggests a different course of action, in which these samples are mixed with images from the labeled dataset, using MixUp [20]. This method requires labels for data generation, which is why a solution similar to *Pseudo-Labels* was

No. labels	S +CE	Π-*model*		*Autoencoder loss*		*NN distance minimization*
		SSL	SSL+ 3 aug.	S	SSL	SSL
320	50.44	51.88	51.31	52.38	52.44	52.92
400	49.16	50.38	49.7	51.49	51.34	53.58
2000	57.91	60.77	60.17	63.67	63.46	64.29
4000	66.2	65.04	64.85	67.25	68.53	70.47
10000	71.82	72.92	72.89	74.11	74.26	76.59

Table 6.2 The classification accuracy in percentages reported on the FER+ dataset for the experiments focused on small networks and the use of SSL. The columns which contain *S* represent trainings where no additional (unlabeled) images were used. The columns marked with *SSL* have used the remaining images in the dataset, but disregarded their labels. The baseline with no extra loss terms is denoted with *CE*. Results also reported in [68].

employed, but operating in the feature space, combined with a loss similar to *Large Margin Loss*. In the end, the additional losses, used alongside the Cross Entropy loss, were similar to the *Large Margin Loss*, and were defined separately based on the type of the currently processed sample. From the perspective of facial expressions, the results on the FER+ dataset prove the method's usefulness (Table 6.3).

No. labeled samples Algorithm	320	400	2000	4000	10000	All
WideResNet-28-2	nc	37.92	50.29	56.78	63.56	84.88
Supervised [57]	-	-	-	-	-	84.99
MeanTeacher [18]	-	45.56	50.84	58.28	68.36	-
MixMatch [19]	45.60	50.25	58.35	70.91	71.24	-
Margin-mix [70]	50.76	56.75	60.83	75.18	81.25	85.36

Table 6.3 A comparison between various approaches used with the WideResNet-28-2 architecture, reported on the FER+ dataset. The cases in which the algorithms did not converge are denoted with "nc". Results also reported in [70].

In order to prove the method's usefulness in a generic setting, it was first tested on general datasets such as CIFAR-10, CIFAR-100 [50], SVHN [71] and STL-10 [72].

6.8.4 Other experiments

Besides the methods presented above, the solution proposed in [73] tackles an important issue regarding methods such as *Pseudo-Labels* used in semi-supervised learning scenarios. Although simple, the labeling algorithm can prove to be inefficient if the distributions of the used datasets are strongly dissimilar. In [73] an extra regularizing technique was used to counter this issue.

The unlabeled data is initially automatically labeled using *Pseudo-Labels*, but the network weights are adjusted only if the performance gain surpasses a minimum threshold, determined by a function applied to a temperature parameter. Additionally, random disturbances are added to the loss, when computed for unlabeled samples. All the unlabeled samples originate from the MegaFace dataset. The results on the FER+ datasets are reported in Table 6.4, additional trainings targeting RAF-DB. Moreover, two extra scenarios were considered. The first one involves testing the method on datasets with children as subjects, while the second one introduced a new expression of *anxiety* in RAF-DB, alongside relevant samples.

Method	Type	FER2013	FER+
AlexNet [74]	S	61.10	-
AlexNet [73]		68.2	78.08
FSN [74]		67.60	-
VGG-13 (majority voting) [57]		-	83.85
VGG-13 (probabilistic label drawing) [57]		-	84.99
FUS [75]		-	67.03
AlexNet [73] + *Pseudo-Labels*	T	69.12	80.05
AlexNet + *Pseudo-Labels* + ALRAO [73]		68.65	80.60
AlexNet + ALT [73]		69.62	82.38
VGG-16 + *Pseudo-Labels* [73]		69.27	84.35
VGG-16 + *Pseudo-Labels* + ALRAO [73]		69.27	82.15
VGG-16 + ALT [73]		69.85	85.20

Table 6.4 The classification accuracy for discrete facial expression recognition using the original FER and FER+ datasets. The *Annealed Label Transfer* algorithm from [73] is highlighted by the *ALT* notation in the table. The two methods from [57] are differentiated by the way the label is obtained. Results also reported in [73].

6.9 Conclusions

This chapter focused on introducing various methods in which CNN performance can be improved, in the scenario of facial expression recognition. Considering the variety of typical approaches, both the case of universal facial expressions and Action Units were considered. An important aspect which needs to be highlighted is that the methods stand out by the fact that they integrate ideas specific to semi-supervised learning.

Chapter 7

Conclusions

Even though almost 10 years have passed since CNNs started getting attention, the pace with which various innovations appear has remained high even lately. Progress is visible both in the general context of convolutional architectures, but also for more specific applications. In the current paper, two main themes were selected on which to discuss possible improvements: painting analysis and facial expression analysis.

Machine Learning has grown in many directions ever since it was established, each with its own unique vision of solving specific tasks. Many of the previously top-tier methods, which were presented in Chapter 2, have lost ground to the current neural approaches. They can still present some competitive results and can be viewed as a good baseline in many cases. Even though some of the general aspects related to neural networks are found here, because of the importance of CNNs in the presented experiments, a separate section was reserved for them, completing some of the aspects which were left out in the previous section.

Besides the exact themes which employed the use of neural networks, a particular interest was shown for semi-supervised learning. In Chapter 3, the necessary basics were presented, since they would be used in the rest of the paper. Some algorithms were proven to be effective in increasing network performance with the aid of unlabeled samples. The future potential is clear, and SSL algorithms could be used to address more general scenarios, once they are developed even further.

7.1 Obtained results

The results presented throughout the paper are the result of multiple years of research. Initially focused on medical applications ([76], [77]), the research efforts were then followed by experiments using a larger available dataset, as will be presented in the following paragraphs.

Chapter 5 introduced the discussion related to the classification of a painting's genre. The completely different nature of this part, when compared to the previous sections, required a separate introduction, where the field of research was presented, before ML elements would be brought into discussion. Besides using CNNs, this chapter was focused on the idea of style transfer. Two relevant algorithms were presented, which could generate new images in order to be used alongside the paintings. From the perspective of the used data, besides the ones with paintings, some which contained photographs were employed, one being comprised of artistic photographs. It was noted that using style transfer methods could increase classification accuracy, but the issue of highly abstract paintings is still a difficult challenge.

The other large area of applications which was explored in detail was the analysis of facial expressions, using Convolutional Neural Networks. In the Chapter 6 focused on this second direction of research, some aspects related to psychology were discussed, in order to explain the chosen representation systems which were used afterwards. A strong emphasis was put on universal facial expressions and Action Units, which were used both separately, and together, in the experimental stage. Also, because of the specific content of face images, some dedicated loss functions needed to be presented. Starting from these, new loss functions were created and used in the presented experiments, also using elements of semi-supervised learning. Some improvements of the results were noted, confirming the usefulness of the proposed solutions, along with the potential of SSL approaches in this area.

7.2 Original contributions

- An extended analysis of the literature was done, both for neural and non-neural solutions for the research areas of genre classification and facial expression analysis. To complete the comparison, additional experiments were conducted, based on necessity([35], [34]).

- From available sources, two auxiliary datasets containing relevant images were created to augment the collection of paintings which was being used, in order to approach genre classification. The two datasets are differentiated by the initial scope of the data sources, one of them being specifically created to contain artistically relevant photographs ([34]).

- After a conceptual comparison of two style transfer methods, both were used to generate new samples to be added to the training process. Besides the raw increase of the dataset, an increase of diversity was also sought, by transferring modern styles ([34]).

- In order to address the long time necessary to generate new samples, some solutions to accelerate the process were analyzed ([33]).

- The impact of a painting's style on genre classification was analyzed, underlining the difficulty of adding highly abstract styles ([34]).

- Multiple methods of improving classification accuracy were proposed, in the context of facial expression recognition. A special aspect of this part of the paper is represented by the integration of elements specific to semi-supervised learning. A first set of experiments targeted the use of the connection between the codification of facial muscle movements and basic expressions. The result was a neural network which was trained to predict labels for both ways of analyzing expressions, also using unlabeled samples in an additional loss function ([66]).

- A second facial expression recognition solution was proposed, this one focusing on using a new regularization technique, based on simulated annealing. Additionally, two more specific aspects were investigated, namely the facial expressions of children and also the expression of *anxiety* ([73]).

- In the third set of experiments a new algorithm was proposed, which merges the use of labeled and unlabeled samples, as is specific for semi-supervised learning, with an additional method of increasing the number of training examples. This algorithm was tested for both facial expression recognition and in the context of a series of popular general use datasets ([70]).

- Besides the three algorithms mentioned above, another series of smaller scale experiments were conducted, by using smaller architectures and integrating semi-supervised learning concepts, in order to enhance universal facial expression classification. In order to verify their usefulness in other tasks, the methods were also tested on a general classification dataset.

7.3 List of original publications

- Mihai Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Efficient domain adaptation for painting theme recognition. In *2017 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4. IEEE, 2017

- Corneliu Florea, Mihai Badea, Laura Florea, and Constantin Vertan. Domain transfer for delving into deep networks capacity to de-abstract art. In *Scandinavian Conference on Image Analysis*, pages 337–349. Springer, 2017

- Mihai Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Can we teach computers to understand art? domain adaptation for enhancing deep networks capacity to de-abstract art. *Image and Vision Computing*, 77:21–32, 2018

- Andrei Racoviteanu, Mihai-Sorin Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Large margin loss for learning facial movements from pseudo-emotions. In *BMVC*, page 108, 2019

- Corneliu Florea, Mihai Badea, Laura Florea, Andrei Racoviteanu, and Constantin Vertan. Margin-mix: Semi-supervised learning for face expression recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 1–17. Springer, 2020

- Corneliu Florea, Laura Florea, Mihai-Sorin Badea, Constantin Vertan, and Andrei Racoviteanu. Annealed label transfer for face expression recognition. In *BMVC*, page 104, 2019

- Andrei Racoviţeanu, Corneliu Florea, Mihai Badea, and Constantin Vertan. Spontaneous emotion detection by combined learned and fixed descriptors. In *2019 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4. IEEE, 2019

- Andrei Racovițeanu, Mihai Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Dual task training for face expression recognition. In *2020 12th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–4. IEEE, 2020

- Andrei Racoviţeanu, Iulian Felea, Laura Florea, Mihai Badea, and Corneliu Florea. Clustering based reference normal pose for improved expression recognition. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 51–61. Springer, 2018

- Mihai-Sorin Badea, Constantin Vertan, Corneliu Florea, Laura Florea, and Silviu Bădoiu. Automatic burn area identification in color images. In *2016 International Conference on Communications (COMM)*, pages 65–68. IEEE, 2016

- Mihai-Sorin Badea, Constantin Vertan, Corneliu Florea, Laura Florea, and Silviu Bădoiu. Severe burns assessment by joint color-thermal imagery and ensemble methods. In *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–5. IEEE, 2016

- Mihai Badea, Constantin Vertan, Corneliu Florea, Laura Florea, and Andrei Racoviteanu. Improving small convolutional neural networks with semi-supervised learning. Submitted to UPB Scientific Bulletin, Series C: Electrical Engineering

- *Burn Assessment by MultiSpectral Imaging* - *BAMSI* project, PN-II-PT-PCCA-2013-4-0357

- *Perceptual Analysis and Description of Romanian visual Art - PANDORA* project, PN-II-RU-TE-2014-4-0733
- *Technologies and Innovative Video Systems for Person Re-Identification and Analysis of Dissimulated Behavior - SPIA-VA* project, PN-III-P2-2.1-SOL-2016-02-0002

7.4 Perspectives for further developments

Regardless of the chosen perspective, either the specialized applications (painting analysis and facial expression recognition) or the training augmentation techniques, there are many opportunities for further developments. Probably the most important direction to focus on is semi-supervised learning. Even though multiple original solutions were approached, alongside some methods from the literature, some general issues were noticed.

Most of the semi-supervised techniques show a significant impact for smaller datasets, but in the context of regular tasks with large datasets, their usefulness is not guaranteed. A possible explanation for this behavior is the ratio between labeled and unlabeled samples. If we take this aspect into account, semi-supervised algorithms could greatly benefit from using generative methods or solutions such as style transfer, even outside of the intended artistic context.

Starting from the intuition mentioned above, the interaction between the semi-supervised loss term and the generative process can be developed. The generative process could be regulated during training by the semi-supervised loss, based on the current state of the network. Taking into account the clear potential of semi-supervised learning, we can consider that the research area of CNNs still has room for growth for applications in which data availability is still an issue.

References

[1] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[2] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[3] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[4] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning*, pages 647–655, 2014.

[5] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[6] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[7] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[8] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. PMLR, 2015.

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[13] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[14] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D Cubuk, and Ian J Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3239–3250, 2018.

[15] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-supervised learning*. The MIT Press, 2006.

[16] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

[17] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, 2013.

[18] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1195–1204, 2017.

[19] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[20] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[21] A Bentkowska-Kafel and J Coddington. Computer vision and image analysis of art. In *Proceedings of the SPIE Electronic Imaging Symposium*, 2010.

[22] Corneliu Florea, Cosmin Toca, and Fabian Gieseke. Artistic movement recognition by boosted fusion of color structure and topographic description. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 569–577. IEEE, 2017.

[23] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. In *Proceedings of the British Machine Vision Conference. BMVA Press*, 2014.

[24] Siddharth Agarwal, Harish Karnick, Nirmal Pant, and Urvesh Patel. Genre and style based painting classification. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 588–594. IEEE, 2015.

[25] Razvan George Condorovici, Corneliu Florea, and Constantin Vertan. Painting scene recognition using homogenous shapes. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 262–273. Springer, 2013.

[26] Wei Ren Tan, Chee Seng Chan, Hernán E Aguirre, and Kiyoshi Tanaka. Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3703–3707. IEEE, 2016.

[27] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.

[28] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

[29] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

[30] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340, 2001.

[31] Mathieu Aubry, Sylvain Paris, Samuel W Hasinoff, Jan Kautz, and Frédo Durand. Fast local laplacian filters: Theory and applications. *ACM Transactions on Graphics (TOG)*, 33(5):1–14, 2014.

[32] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

[33] Mihai Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Efficient domain adaptation for painting theme recognition. In *2017 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4. IEEE, 2017.

[34] Mihai Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Can we teach computers to understand art? domain adaptation for enhancing deep networks capacity to de-abstract art. *Image and Vision Computing*, 77:21–32, 2018.

[35] Corneliu Florea, Mihai Badea, Laura Florea, and Constantin Vertan. Domain transfer for delving into deep networks capacity to de-abstract art. In *Scandinavian Conference on Image Analysis*, pages 337–349. Springer, 2017.

[36] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010.

[37] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[38] Christopher Thomas and Adriana Kovashka. Seeing behind the camera: Identifying the authorship of a photograph. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3494–3502, 2016.

[39] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *International Journal for Digital Art History*, (2), 2016.

[40] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98:45–60, 1999.

[41] Cătălin-Daniel Căleanu. Face expression recognition: A brief overview of the last decade. In *2013 IEEE 8th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pages 157–161. IEEE, 2013.

[42] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1113–1133, 2014.

[43] Chieh-Ming Kuo, Shang-Hong Lai, and Michel Sarkis. A compact deep learning model for robust facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2121–2129, 2018.

[44] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020.

[45] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016.

[46] Paul Ekman and Wallace V Friesen. *Manual for the facial action coding system*. Consulting Psychologists Press, 1978.

[47] Emily B Prince, Katherine B Martin, Daniel S Messinger, and M Allen. Facial action coding system, 2015.

[48] James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161, 1980.

[49] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 94–101. IEEE, 2010.

[50] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[51] Yash Srivastava, Vaishnav Murali, and Shiv Ram Dubey. A performance evaluation of loss functions for deep face recognition. In *National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics*, pages 322–332. Springer, 2019.

[52] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[53] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[54] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[55] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.

[56] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 302–309. IEEE, 2018.

[57] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283, 2016.

[58] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer, 2013.

[59] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2018.

[60] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016.

[61] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5562–5570, 2016.

[62] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, Qianli Feng, Yan Wang, and Aleix M Martinez. Emotionet challenge: Recognition of facial expressions of emotion in the wild. *arXiv preprint arXiv:1703.01210*, 2017.

[63] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, and Iain Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 57–64. IEEE, 2011.

[64] Vanessa LoBue and Cat Thrasher. The child affective facial expression (cafe) set: Validity and reliability from untrained adults. *Frontiers in Psychology*, 5:1532, 2015.

[65] Rizwan Ahmed Khan, Arthur Crenn, Alexandre Meyer, and Saida Bouakaz. A novel database of children's spontaneous facial expressions (liris-cse). *Image and Vision Computing*, 83:61–69, 2019.

[66] Andrei Racoviteanu, Mihai-Sorin Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Large margin loss for learning facial movements from pseudo-emotions. In *BMVC*, page 108, 2019.

[67] Kaili Zhao, Wen-Sheng Chu, and Aleix M Martinez. Learning facial action units from web images with scalable weakly supervised clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2090–2099, 2018.

[68] Mihai Badea, Constantin Vertan, Corneliu Florea, Laura Florea, and Andrei Racoviteanu. Improving small convolutional neural networks with semi-supervised learning. Submitted to UPB Scientific Bulletin, Series C: Electrical Engineering.

[69] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. 2017.

[70] Corneliu Florea, Mihai Badea, Laura Florea, Andrei Racoviteanu, and Constantin Vertan. Margin-mix: Semi-supervised learning for face expression recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 1–17. Springer, 2020.

[71] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[72] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[73] Corneliu Florea, Laura Florea, Mihai-Sorin Badea, Constantin Vertan, and Andrei Racoviteanu. Annealed label transfer for face expression recognition. In *BMVC*, page 104, 2019.

[74] Shuwen Zhao, Haibin Cai, Honghai Liu, Jianhua Zhang, and Shengyong Chen. Feature selection mechanism in cnns for facial expression recognition. In *BMVC*, page 317, 2018.

[75] Elizabeth Tran, Michael B Mayhew, Hyojin Kim, Piyush Karande, and Alan D Kaplan. Facial expression recognition using a large out-of-context dataset. In *2018 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 52–59. IEEE, 2018.

[76] Mihai-Sorin Badea, Constantin Vertan, Corneliu Florea, Laura Florea, and Silviu Bădoiu. Automatic burn area identification in color images. In *2016 International Conference on Communications (COMM)*, pages 65–68. IEEE, 2016.

[77] Mihai-Sorin Badea, Constantin Vertan, Corneliu Florea, Laura Florea, and Silviu Bădoiu. Severe burns assessment by joint color-thermal imagery and ensemble methods. In *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–5. IEEE, 2016.

[78] Andrei Racoviţeanu, Corneliu Florea, Mihai Badea, and Constantin Vertan. Spontaneous emotion detection by combined learned and fixed descriptors. In *2019 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4. IEEE, 2019.

[79] Andrei Racovițeanu, Mihai Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Dual task training for face expression recognition. In *2020 12th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–4. IEEE, 2020.

[80] Andrei Racoviţeanu, Iulian Felea, Laura Florea, Mihai Badea, and Corneliu Florea. Clustering based reference normal pose for improved expression recognition. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 51–61. Springer, 2018.