POLITEHNICA UNIVERSITY OF BUCHAREST

Doctoral School of Electronics, Telecommunications and Information Technology

Decision No. 1092 **from** 24-07-2023

Ph.D. THESIS SUM-MARY

Ing. Andrei-Mircea RACOVIŢEANU

ÎNVĂŢAREA CU MARJĂ LARGĂ PENTRU ANALIZA IMAGINILOR

LARGE MARGIN LEARNING FOR IMAGE ANALYSIS

THESIS COMMITTEE

Prof. Dr. Ing. Mihai Ciuc Politehnica Univ. of Bucharest	President
Prof. Dr. Ing. Corneliu Florea Politehnica Univ. of Bucharest	PhD Supervisor
Prof. Dr. Ing. Cătalin Căleanu Politehnica Univ. of Timisoara	Referee
Conf. Dr. Ing. Ioan Buciu University of Oradea	Referee
Prof. Dr. Ing. Constantin VERTAN Politehnica Univ. of Bucharest	Referee

BUCHAREST 2023

Cuprins

Listă de tabele

Listă de figuri

Capitolul 1

Introduction

1.1 Presentation of the field of the doctoral thesis

In recent years, the field of artificial intelligence and computer vision have experienced a very fast evolution. Easier access to a large volume of data together with algorithms of automatic learning have contributed to the development of systems that imitate human capabilities. Several industries benefit from the progress of the previously mentioned fields, such as: medicine, finance or the entertainment industry.

The work addresses the problems of recognizing facial expressions and finding images with similar content. The detection of facial expressions focuses on automatic recognition of people's emotions. This implies the development of some systems that can interpret and understand emotional states and reactions based on facial movements. Retrieving images with similar content involves the use of some techniques of searching and extracting images from a large volume of data according to visual content.

1.2 Scope of the doctoral thesis

The primary objective of the thesis is to discover more effective solutions for facial expression recognition and discovering similar images. The proposed techniques relied on convolutional network methods and the introduction of novel cost functions and augmentation techniques.

Convolutional networks need large amounts of data, which are difficult or impossible to acquire. To overcome this obstacle, they can employ semi-supervised learning or domain adaptation techniques. From this reason, the proposed solutions exceed the category of "supervised" algorithms and were included in the "non-supervised" category.

The first problem addressed is that of facial expressions. In this case, it were used both discrete facial expressions and facial movements known as "action units". Even though facial movements are more intuitive, action units are more objective and difficult to confuse. A new cost function for a better clusterization of the embedding space

was proposed to enhance the results. In addition, new augmentation and regularization methods were also tested.

The retrieval of images with similar content was the second topic studied. In that situation the visual content of the images can be very complex, which is why the descriptors obtained with a convolutional network can be quite easily confused with each other. Same cost function was evaluated to determine whether a better organization of the descriptors space increases the number of similar images returned for a reference image.

1.3 Content of the doctoral thesis

The work contained seven chapters. In chapters 2, 3, and 4, the theoretical concepts that served as the basis for the experimental results are presented. hese are highlighted in chapters 5 and 6, while the final chapter is reserved for the conclusions.

Chapter 2 provides information about convolutional networks. Here, the most common forms of layers, cost functions, optimization techniques, and architectures are described. In chapter 3, the various forms of automatic learning are discussed. The focus is on semi-supervised learning/transfer learning as well as the augmentation techniques employed. The fourth chapter concentrates on loss functions utilized for a better grouping of descriptors. Mathematical concepts, functionality and a short comparison of them are presented. Chapters 5 and 6 contain the results obtained within the thesis for the 2 problems addressed, while chapter 7 contains the conclusions.

Capitolul 2

Convolutional Networks

Convolutional neural networks are a type of deep learning models that have proven to be especially effective in computer vision tasks such as image classification, object detection, and image segmentation. These are designed to extract relevant features from input data, making them appropriate for image-related tasks.

2.1 Layers of Convolutional Neural Networks

The fundamental concept underlying convolutional networks is to use multiple types of layers and mathematical procedures to capture the most important features of input data. The first relevant layer is the convolutional layer, where the convolution operation is effectively carried out using filters with different weights. In addition, the concept of local connectivity appears, which indicates that not all the neurons in successive layers are connected.

Subsampling layers are the next important layer. They have the role of reducing the size of feature maps produced by convolutional operations. Thus, the computation effort is significantly reduced, and the system's capacity for generalization is enhanced. The most common variant of sub-sampling is *max-pooling*, which takes into consideration the maximal value in the neighborhood.

The phenomenon of overfitting can appear quite often in convolutional networks if regularization layers are not used. Such a layer is the *dropout* layer [1] which implies the random elimination of a certain percentage from neural connections between layers. The connections are only removed in the process of training, not in the testing one. The activations of each layer may also be normalized via the normalization of data collections (*batch normalization*) [2]. Therefore, training time is reduced, and instances in which the activations of successive layers are completely distinct are eliminated. The last layers are the fully connected ones that are often used as descriptors or decision layers.

2.2 Learning process

A machine learning system uses a series of mathematical functions entitled loss functions to measure how far its predictions are from the reference labels. Depending on the nature of the labels, there are two categories of problems: *classification* (discrete labels) and *regression* (continuous labels). For classification problems, the most common loss function is *cross entropy*, whereas the most common loss function for regression problems is *mean squared error*.

During training, a mathematical optimization algorithm (*optimizer*) helps the system to reduce errors. It modifies the weights after each iteration based on the impact of each weight on the total loss function. The process of adjusting weights based on the total error, is known as *backpropagation*.

2.3 Convolutional Architectures

As the use of convolutional networks became more widespread, a number of standard architectures that can be applied in diverse tasks were sought. AlexNet [3] was the first architecture to produce remarkable results. A new activation function that cancel negative weights was also introduced.

VGG [4] is a well-known architecture that succeeded to increase the depth of networks by adding new layers. Contrary to ALexNet, the convolution filters have a smaller size and sub-sampling windows do not overlap. However, the increasing number of layers favored the appearance of another concerning phenomenon called *vanishing gradients*

The ResNet architecture [5] was proposed as a countermeasure to the previously mentioned phenomenon. The residual block was the primary innovative component. In contrast to other remembered architectures, in residual architectures the input of a layer is transferred to the input of upper layers. In this manner, the depth of a network could be increases without a negative effect on performance. After it was established that the excessive increase in depth no longer brings significant benefits other variants of ResNet were also proposed. Wideresnet [6] proposes an increase in feature maps by enlarging the convolutional layer width.

Capitolul 3

Machine Learning Concepts

3.1 Types of learning

According to the type of the input data labels, automatic learning algorithms can be classified into three large categories. If data contains labels, the algorithms are supervised; if not, they are unsupervised. The third category consists of semi-supervised algorithms, which typically utilize a lesser quantity of annotated data and a substantial portion of annotated data. The latter were predominantly used in the experimental part.

3.2 Transfer Learning/ Domain Adaptation

As a result of insufficient data and incorrect labeling, transfer learning techniques and domain adaptation gained more popularity. Learning by transfer involves using a pre-trained system. The system can be used further for tasks that are different with the one it was originally trained. In other words, this technique is trying to improve performance for a new task using previously learned features.

Domain adaptation, on the other hand, involves establishing a connection between two domains with distinct data distributions. ypically, a system is trained for an initial domain before being adapted to a target domain using techniques such as loss function weighting or feature alignment. By reducing the structural differences between the two domains, the ultimate result is an increase in the power of generalization.

3.3 Semi-supervised algorithms

3.3.1 Pseudo-Labels

When a system trained on labeled data is applied to a series of unlabeled samples, pseudo-labeling [7] results. Basically, the labels for the unannotated dataset are obtained

using a system that has already been trained. Then, the data with the newly generated labels are added to the original ones and the entire ensemble is retrained.

3.3.2 Mean-Teacher

Mean-Teacher is a semi-supervised algorithm that uses two similar networks. The first network is considered a "student" and has a classification role. The second one is the "teacher" network and must replicate as accurately as possiblethe output of the "student" network. During the training process, one of the primary objectives is to minimize the distribution differences between the "student" and "teacher" networks.

3.3.3 MixMatch

MixMatch [8] is an algorithm that maximizes performance by combining several semi-supervised paradigms. It employs the MixUp method [9] to generate new samples and labels. The total loss is computed fot both original and generated data.

3.4 Augmentations methods

Considering that convolutional networks require significant quantities of data, many solution to resolve this problem were sought. Methods of augmentation are used to artificially increase the number of training samples.

The most common augmentation methods are related to the spatial distribution of the elements that form an image. Consequently, new samples can be generated via rotation, mirroring, or resizing. It is also possible to do this at the pixel level using filtering or contrast techniques .

The *MixUp* augmentation technique [9] helps generate new data by linearly combining a random pair of existing samples from the training database. Labels are created for this data using the same method. Due to the difficulty of interpreting as a singular value the linear combination of 2 discrete labels, a probabilistic distribution was used.

Capitolul 4

Methods of structuring the descriptive space

For complex problems, it is possible that the descriptive space provided by convolutional networks to contain examples with substantial overlap. For this reason, research was carried out in the direction of new loss functions to organize more efficiently the characteristics obtained.

4.1 Center Loss

Among the most well-known clustering losses is *Center Loss* [10]. The subsequent guiding principle is the minimization of the distances between objects that belong to the same category.The center losst function is mathematically represented by Equation 4.1, where e_i is the descriptor obtained from the fully connected layer preceding the decision one and c_i is the associated centroid for current sample.

$$L_C = \frac{1}{2}\sum_{i=1}^{N} \|e_i - c_{l_i}\|_2 \tag{4.1}$$

According to 4.2, the position of the centroids is recalculated after each iteration in order to create a more open space. Δc_k^i s the mean of the data belonging to class i în lotul de date curent, and α is a subunit parameter that tempers a potential negative impact on the new position of the centroids of some incorrectly labeled samples. Unlike other cost functions,center loss cannot be used independently because does not have a decision role.

$$c_k^{i+1} = c_k^i - \alpha \Delta c_k^i \tag{4.2}$$

4.2 Island Loss

In contrast to the cost function discussed in the preceding section, the *Island Loss* [11] provides an additional benefit. Besides minimizing the variance in the same class, it tries to maximize the distance between the centroids associated with each label. Equation 4.3 describes the mathematic formula, where L_C is the center loss. The second term accumulates the angular distances between the centroid of current sample and the centroid of all other classes.

$$L_{IL} = L_C + \lambda_1 \sum_{c_i \in C} \sum_{c_j \in C, c_i \neq c_j} \left(\frac{c_i \cdot c_j}{\|c_i\|_2 \|c_j\|_2} + 1 \right) \tag{4.3}$$

4.3 Ring Loss

The *Ring Loss* [12] necessitates the efficient normalization of the descriptors so they can be interpreted geometrically as a circle. In equation 4.4, F_{x_i} is the embedding corresponding to the penultimate fully connected layer, and R is the targeted normed value. From a geometric point of view, it is associated with the radius of a circle.

$$L_R = \frac{\lambda}{2m} \sum_{i=1}^{m} (\|F_{x_i}\|_2 - R)^2 \tag{4.4}$$

4.4 Large Margin Loss

Center loss is based only on the distances between embeddings of the same class. The island loss combats this shortcoming, but it has a number of limitations due to the angular distance. If the angle between two groups of samples is extremely narrow, they cannot be adequately separated. The circular function imposes a circular geometric representation of the descriptors, thereby optimizing angular distances. iHowever, if two classes overlap in the initial representation, there is a high probability that they will also overlap in the circular representation.

Taking into account the limitations mentioned in the preceding paragraph, a new loss function was proposed to satisfy current requirements. It was entitled *Large Margin Loss* [13] and requires the use of a Euclidean distance between each sample and the centroids associated with the other classes. Thus, limitations imposed by the previous losses are avoided.

This function is mathematically defined by ecuation 4.5. e_i is the embedding represented by the penultimate fully connected layer, c_j is the associated centroid for e_i, and c_k represents all the other centroids except the centroid of e_i.

$$\mathbf{L_{LM}} = \sum_{i=1}^{N} \left(\left\| \frac{\mathbf{e}_i}{\|\mathbf{e}_i\|_2} - \frac{\mathbf{c}_j}{\|\mathbf{c}_j\|_2} \right\|_2 - \frac{1}{C-1} \sum_{k=1, k \neq j}^{C} \left\| \frac{\mathbf{e}_i}{\|\mathbf{e}_i\|_2} - \frac{\mathbf{c}_k}{\|\mathbf{c}_k\|_2} \right\|_2 \right) \tag{4.5}$$

Figure 4.1 shows the functionality of the large margin function. With 3 samples available (X_A;X_B;X_C) with labels A, B and C, the feature–space will be updated such that the distance to the appropriate class is reduced (shown by continuous arrows) while distances to other centroids is increased (illustrated by dashed lines).

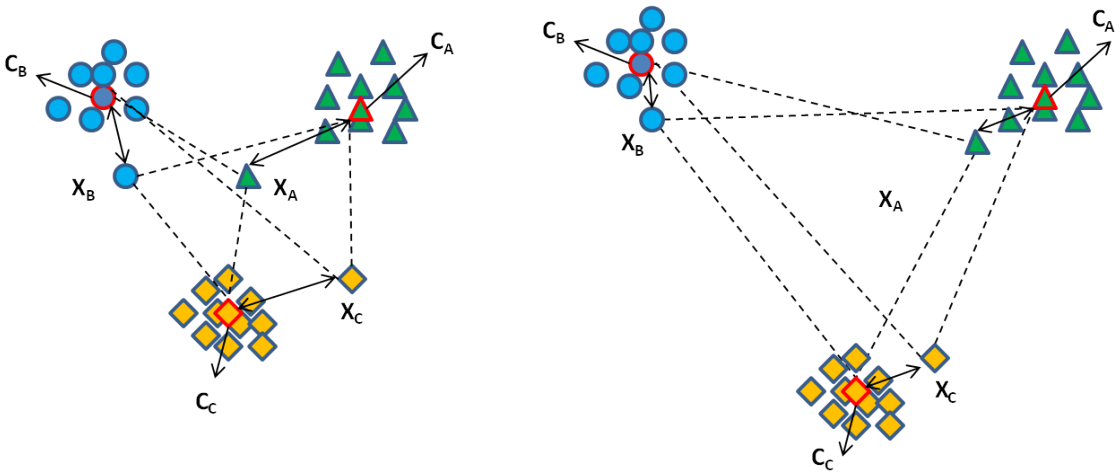

Figura 4.1 Large Margin Loss functionality. Left: Before Large Margin. Right: After Large Margin

Capitolul 5

Facial Expression Analysis

This chapter is the most comprehensive of the entire thesis and represents the outcome of years of research. Among the tasks addressed are the recognition of discrete facial expressions and facial movement detection. To improve the performance were used new cost functions and innovative augmentation and regularization methods.

5.1 Facial Expresssions Quantization

Recognizing facial expressions is not an easy task even for humans. This problem is even more challenging for a machine learning system. Over time, several models that would define this issue more conveniently were sought .

The first extensively used model in this discipline is Ekman's definition of discrete expressions [14]. It is based on 6 fundamental expressions: fear, disgust, happiness, anger, surprise, sadness. Usually, the neutral expression is also added. Even though this variant is straightforward to use, other models have been developed to more objectively define emotions. Thus, the Action Unit Coding System was developed, which implies that each facial expression is composed of a series of activations of specified facial muscles (*Action Units*).

The system comprises 43 action elements that have been subdivided based on their position on the face. There are facial movements associated with the most essential facial features, including the eyes, mouth, and eyebrows. Even though it is a model that leaves little room for interpretation, a sufficiently qualified staff is required to identify even the most subtle facial movements. There are also more complex models that consider the intensity of expression and whether it is positive or negative. This paradigm was proposed by Russel [15] and takes compound emotions into account.

5.2 Challenges in Facial Expression Analysis task

As demonstrated in section 5.1, the problem of facial expressions is one very difficult to solve. Moreover, there are also a multitude of challenges that can affect recognition performance. An initial factor is inconsistent data volume. Even though it may be simple to obtain many images with facial expressions, the annotation process can be time-consuming and costly. Finding qualified human annotators is not a simple task. Consequently, there are still numerous images with incorrect labels.

Another reason is the short appearance time of facial expressions. The majority, if not all, of the facial movements that form the emotions are extremely subtle, even for computers. Also, there are several expressions that are easily confused with wach other such as fear and surprise.The reasos in that oth emotions entail mouth opening and eyebrow elevation.

5.3 Related Work

Being a difficult problem to solve, the recognition of facial expressions has attracted a large number of researchers in the recent past; consequently, the number of articles on this subject has increased significantly. Fundamental expression recognition with convolutional networks has been discussed in works like [16, 17] (section 5.1).

Du [18] and Zhang [19] observed the inconsistency of the training data and migrated to semi-supervised learning solutions. In addition, the recognition of the action units was addressed. Corneanu [20], Zhao [21], and Benitez [22] are among the one who detected facial movements and their intensity. Domain adaptation and transfer learning were discussed in [23–25].

5.4 Facial Detection Solutions

The majority of datasets contain images with faces or facial expressions and their backgorund. Background information found in most images is redundant and not useful for convolutional networks. In addition, cropping reduces the dimensionality of the faces in the initial pictures , which contributes to a faster training.

Among the numerous available solutions, the Viola-Jones [26] and MTCNN [27] algorithms stand out. The Viola-Jones method captures faces in images, regardless of their size, using a series of features at different scales. In addition, the integral image is utilized to obtain the facial recognition procedure in real time.

In contrast, MTCNN [27] is a convolutional network-based technique organized in 3 stages. Faces are identified first, followed by the matching of detection boxes. Last stage is the recognition of the facial elements necessary for the maximum alignment of the face.

5.5 Databases

o satisfy the requirements imposed by the semi-supervised methods evaluated in this section, multiple experimental data sets were employed. The recognition of fundamental facial expressions and detection of facial movements (action units) has been addressed. More details about each dataset can be found in the following sections.

5.5.1 Facial Expression Recognition Databases

FER/FER+. FER2013 [28] and FER+ [29] represent 2 data sets that provide images with basic facial expressions. FER+ is an extension of FER2013 in which several incorrectly assigned expressions have been corrected. It consists of 35,000 faces in the wild images.

Megaface. Megaface [30] is a much more comprehensive database that contains approximately 1 million faces in the wild images. It has no labels, which is why it was used the unlabeled portion of the data for the semi-supervised experiments.

RAF-DB. RAF-DB [31] is similar to FER because it contains discrete expressions. However, it contains fewer images acquired under laboratory conditions. Forty individuals were responsible for annotating the images.

Facial expression in children. Within this work, facial expressions in children recognition was also studied. CAFE [32] is one of the most well-known data sets on this subject. LRIS [33]is an additional set that compensates for CAFE's limitations by increasing the numbner of images and emotional diversity .

5.5.2 Action Unit Recognition Datasets

CK+. CK+ [34] is a data set containing both discrete emotions and action units. It is included in this section because it was only used to identify action units. The images are organized by every subject, with each sequence containing frames varying from neutral expression to maximal expression intensity. The final images in each sequence are also annotated at the action unit level.

Emotionet. Emotionet [22] is a data set that contains approximately 1 million images. Unlike Megaface [30], it has 50000 images with action units labels. The images contains faces in the wild and the labels for facial movement are binary.

DISFA. Compared to Emotionet [22], DISFA [35] also provides annotations for the intensity of action units. Thus, the labels are between 0 (action unit is not active) and 5 (action unit has maximum intensity). It contains 130 000 images divided into 27 subjects, obtained in a laboratory setting.

5.6 Facial Expression Recognition

This chapter focused on enhancing the outcomes for facial expression detection using several novel or previously published techniques. Among these concepts are semi-supervised learning/transfer learning, loss functions for a better structure of the descriptor space (Center Loss - Section 4.1 , Ring Loss - Section 4.3 , and Large Margin - Section 4.4) and augmentation methods such as MixUp (Section 3.4) .

5.6.1 Large Margin Loss for Learning Facial Movements from Pseudo-Emotions

Having as motivation the success of the center and island loss functions, in this section, the potential of large margin loss discussed in section 4.4 was explored. For this scenario, an Alexnet architecture (section 2.3) was used to detect discrete facial expressions and action units simultaneously. The architecture contained 2 output layers connected in parallel: one for facial expression recognition and one for action unit detection.

Attempts to train multiple data sets with different label types were made. Because of this, it was necessary to modify the domain from action units to discrete facial expressions. A set of equations describing the fundamental expressions as a sum of concurrently active action units was used to establish the link between the two categories. The result consisted in obtaining some pseudo-expressions.

Because the large margin loss requires the concept of centroids (discrete classes), these pseudo-expressions were necessary. Multiple facial movements can occur simultaneously, making the detection of action units a multi-class problem. Even though action units woud have a unique label, it could result in an excessive number of underrepresented classes.

The results for the recognition of facial expressions can be seen in table 5.1, while the performance for the action units is shown in table 5.2. Someone may notice that the large margin loss function performs better than center and island loss.

Method	Framework	Avg. Acc.	Acc.
AlexNet - [31]	Superv	55.60	68.90
AlexNet + Feat.Sel.Net [16]	Superv	72.46	81.10
AlexNet + Island loss [11]	Superv	57.1	75.08
AlexNet + Center loss [10]	SSL	63.15	78.81
AlexNet + Island loss [11]	SSL	64.53	78.81
AlexNet + LM loss	SSL	67.26	79.85

Tabela 5.1 Acc obtained on the classification problem on the RAF-DB database. Prior work: Feat.Sel.Net - feature selection network, Our proposal uses Large Margin (LM)[13]

Method	FW	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU17	AU20	AU25	AU26	AU43	Avg. small	Avg. full
AlexNet [21]	Sv	24.2	n/a	34.7	**39.5**	73.1	n/a	86.8	n/a	n/a	88.5	45.6	n/a	56.1	n/a
AlexNet cen. loss [10]	Sv	**34.4**	30.3	55.3	33.3	69.10	**46.1**	79.3	27.8	**32.3**	84.4	43.2	48.8	57.9	48.8
AlexNet +WSC [21]	SSL	25.3	n/a	34.5	39.3	**75.6**	n/a	**87.4**	n/a	n/a	**88.8**	47.4	n/a	57.0	n/a
AlexNet + Isl. loss [11]	T.	30.4	29.5	**56.7**	30.6	66.7	44.1	77.3	26.7	23.8	83.9	47.3	43.9	56.14	46.7
AlexNet + LM loss [13]	T.	34.1	**31.1**	56.6	33.9	71.0	45.1	78.1	**30.9**	25.3	83.8	**50.9**	47.2	58.33	49.0

Tabela 5.2 F1 score (%) while detecting action units on the EmotioNet database. The framework (FW) is either supervised (Sv), semi-supervised (SSL) or transfer (T). "Avg small" is the average over the reduced set of AU:1,4,5,6,12,25,26, Avg full is over the entire set Table from [13]

5.6.2 Margin-Mix

The *Margin-Mix* algorithm [36] combines the large margin loss function with a few augmentation techniques, including MixUp [9]. As stated in section 3.4, the MixUp method is used for purely supervised learning. Although the linear combination between 2 images from the dataset is possible, the result will not be able to be assigned to a class in the absence of initial labels.

This is where the large margin concept comes into play; it is used to label the new formed examples with MixUp using the descriptors formed by a convolutional network. To minimize the effect of overlapping data in the descriptive space the label assignment was achieved using a Fuzzy technique [37]. In this way, a sample was assigned a probability distribution for each class, not just a unique label. In addition, these new samples along with their predicted labels were used during training.

A WIde-ResNet architecture [6] was used throughout the experiments. Table 5.3 displays the data obtained for the standard data sets STL-10 and SVHN. The RAF-DB [31] results were organized in Table 5.4. The results in the first table are comparable to those in the specialized literature. The second table's results may be more suggestive. In comparison to other purely supervised methods, Margin-Mix has a significantly higher performance. Importantly, as the amount of labeled data decreases, the proposed method produces better results. However, when the entire data set is utilized (the last column in table 5.4), the values are comparable, proving that Margin-Mix is an option to consider when the data set lacks sufficient annotations.

	SVHN		STL	
Methods/Labels	1000	4000	1000	5000
Supervised [6]	–	12.84	–	–
Π-Model [38]	8.06	5.57	17.41	39.19
VAT [39]	5.63	18.68	11.05	–
MeanTeacher [40]	5.65	3.39	10.36	–
ICT [41, 6]	3.53	–	7.66	–
MixMatch [8]	3.27	2.89	10.18	5.59
MarginMix [36]	3.35	8.33	**9.85**	5.80

Tabela 5.3 Comparative error (smaller is better) on SVHN and STL datasets obtained with WideResNet-28-2.[36]

Methods/Labels	320	400	1000	4000	
Supervised	nc	26.75	35.25	55.66	85.58
Supervised [31]	–	–	–	–	84.13
MeanTeacher [40]	nc	28.83	36.53	60.36	–
MixMatch [8]	35.60	42.25	60.37	65.24	–
MarginMix [36]	**40.55**	**45.75**	**66.47**	**70.68**	85.36

Tabela 5.4 Comparative accuracy (larger is better) on RAFDB dataset obtained with WideResNet-28-2 . Top row lists the number of examples with labels (over all classes) considered. nc - not converged [36]

5.6.3 Randomization Injection for Efficient Transfer in Face Expression Recognition

The technique proposed in [42, 43] is composed of semi-supervised learning and the technique described in section 3.3.1(Pseudo-Labels). It is easy to use pseudo-labeling, which consists of applying a pre-trained system to a supervised problem to label a succession of unlabeled samples.

However, Pseudo-Label starts from the premise that the labeled data has a similar distribution to the labeled one. This claim is false on multiple occasions and can negatively impact performance. In Figure 5.1 is exposed schematically the functionality of the proposed method. or the supervised case (a), the decision line is based on the existing data, so everything is evident. When unlabeled data are also present (blue circles), pseudo-labeling causes the system to become overconfident on the provided labels, even if they are incorrect. In this case, the border can look like a system prone to overfitting (b). The random injection in the gradient (AIR) induces a random degree of uncertainty in the prediction, which reduces the likelihood of the border to be focused too much on uncertain points.

The mathematical function that altered the gradients is defined as::

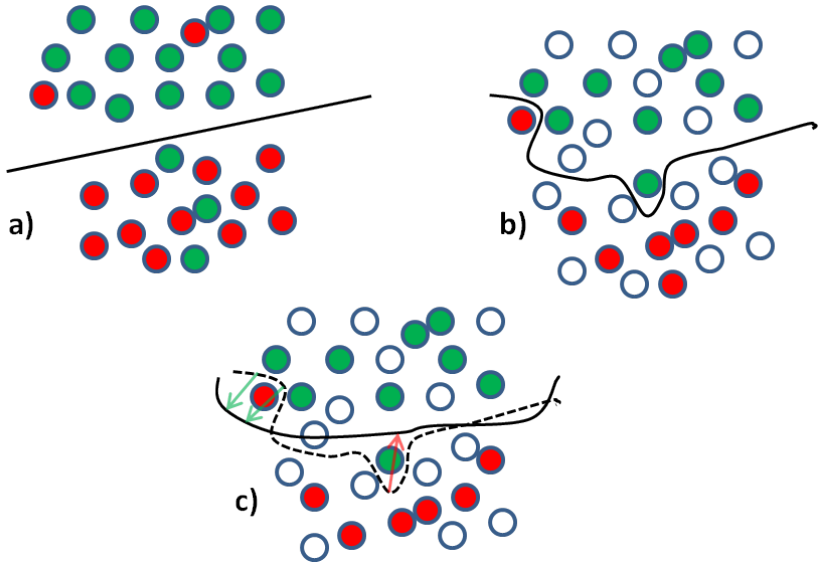

Figura 5.1 Separation boundary for a) Ideal separation -supervised b) Pseudolabel- semi-supervised c) AIR - Transfer learning

$$f(n,\lambda) = \begin{cases} \frac{\lambda n}{50}, & n < 50 \\ 0, & n \geq 50 \end{cases} \tag{5.1}$$

where $g : \{1, N_{epochs}\} \times [-1,1] \rightarrow [-1,1]$, λ is a uniformly distributed radom variable in [-1,1] and n is the number of epochs. This quantity is added to the current cost, and the weights only change if the performance improvement is substantial. Thus, the negative potential generated by the dissimilar distributions between data sets was minimized.

The results for the RAF-DB database are shown in Table 5.5. It can be noted that the numbers obtained with the random injection in the gradient are predominantly greater than the methods chosen for comparison. Notably, the performance is approximately 2% to 3% better than pseudo-labeling, indicating a more efficient tranfer of information.

A series of experiments were also conducted on the LRIS database [33] containing images of children's facial expressions. The results are shown in table 5.6. The outcomes were still significantly better in comparison with the supervised baseline.

	Method / Metric	Avg. Acc.	Acc.
SUPERV	AlexNet [31]	55.60	68.90
	VGG-16 [31]	58.22	70.53
	DLP-CNN [31]	74.20	84.13
	ResNet-18 [44]	–	80.00
	FSN [16]	72.46	81.10
	gCNN [45] - VGG16	–	85.07
	ensCNN [46]	75.14	86.31
TRANSFER	AlexNet + PL	69.5	78.5
	AlexNet + ALT [42]	72.3	81.50
	AlexNet + AIR [43]	72.5	82.1
	VGG–16 + PL	74.6	83.25
	VGG–16 + ALT [42]	76.50	84.5
	VGG–16 + AIR - [43]	76.82	85.15
	ResNet-50 + PL	77.12	84.8
	ResNet-50 + AIR - [43]	**78.22**	**86.67**

Tabela 5.5 Performance within 7-class problem on the RAF-DB database. FSN - feature selection network, FSM - frame-to-sequence method, PL - standard Pseudo-Label, ALT (Annealing Label Transfer). The current proposal is marked by AIR. With bold are marked the best results. Table from [42]

Method	Accuracy
VGG-16 [33] - supervised	67.2
VGG-16 - +AIR [43]	68.5
ResNet-50 - supervised	72.3
ResNet-50 + AIR [43]	76.6

Tabela 5.6 Performance (recognition rate) within 5-class problem on the LIRIS database containing expression of children. Table from [43]

5.6.4 Action Units Detection with Large Margin Loss

The information in the current section uses the principles of wide margin, pseudo-expressions and domain adaptation exactly as in 5.6.1, but experiments and tested scenarios were extended. In [47] all efforts were concentrated on detecting action units(AUs).

The architectures used in this case were mainly those from the ResNet family [5]. As in section 5.6.1, action units were identified using the information provided by the pseudoexpression-related decision layer. Consequently, three distinct associated loss functions were utilized for each problem: binary cross entropy for predicting facial movements, cross entropy for predicting pseudo-expressions, and large margin for clustering the descriptors space. The total cost function is depicted in Equation 5.2. The constants λ_1, λ_2, and λ_3 are used to numerically balance the three terms of the final cost.

$$L_T = \lambda_1 L_{BCE} + \lambda_2 L_{SE} + \lambda_3 L_{LM} \tag{5.2}$$

The fact that some action units occur much less frequently than others was one of the issues that emerged in the databases used in the experiments. As shown in Figure 5.2, this factor contributes to a reduced detection, as the action units that appear less frequently will overlap with the other ones in the embeddings space. On the left side of the figure are the pseudo-expressions represented by the action units that appear less frequently, while on the right are the pseudo-expressions representing all action units. The pseudo-expressions described by less frequent AUs (coloured) are almost completely covered by the other ones (gray color).

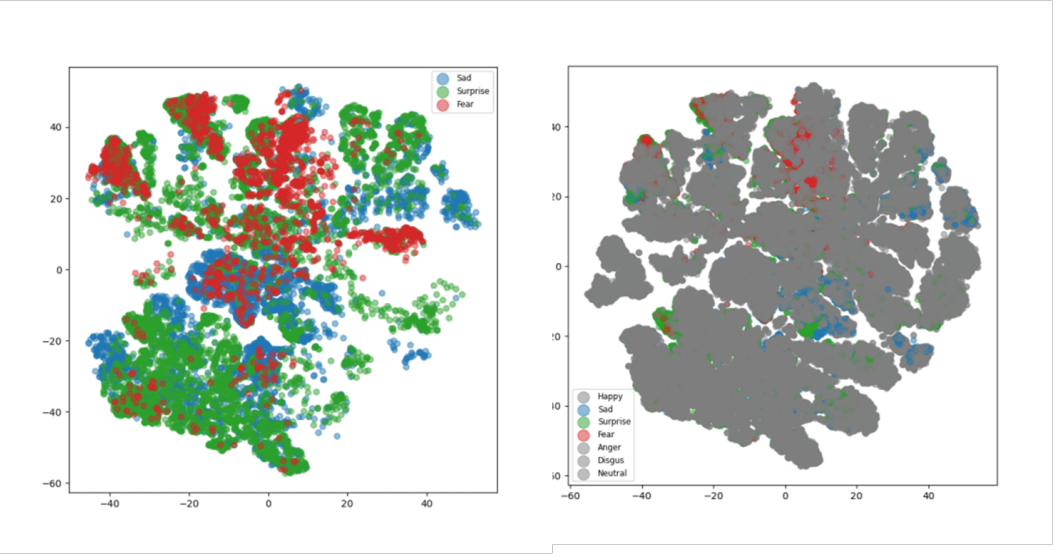

Figura 5.2 Embeddings space representation for pseudo-expressions. Left – Features for the pseudo-expressions represented by sparse AUs. Right - Features for the pseudo-expressions represented by all AUs . Figure from [47]

Tables 5.7 and 5.8 show the results obtained on the DISFA [35] and Emotionet [22] data set in comparison with other similar techniques from the literature. For DISFA, it can be observed that the average results obtained with the large margin loss are better for action units with a lower frequency of occurrence. The overall average is not necessarily superior. On the Emotionet data set, the differences are preserved.

The large margin concept is able to better distinguish on a descriptive level the pseudo-expressions that are represented by the less frequents action units, which may explain why the performance is higher. This is evident in Figure 5.3, where the expressions formed from facial movements that occur less frequently(the coloured dots) are much more compact and slightly overlapped with the others (gray dots - action units that occur more frequently) when large margin is used.

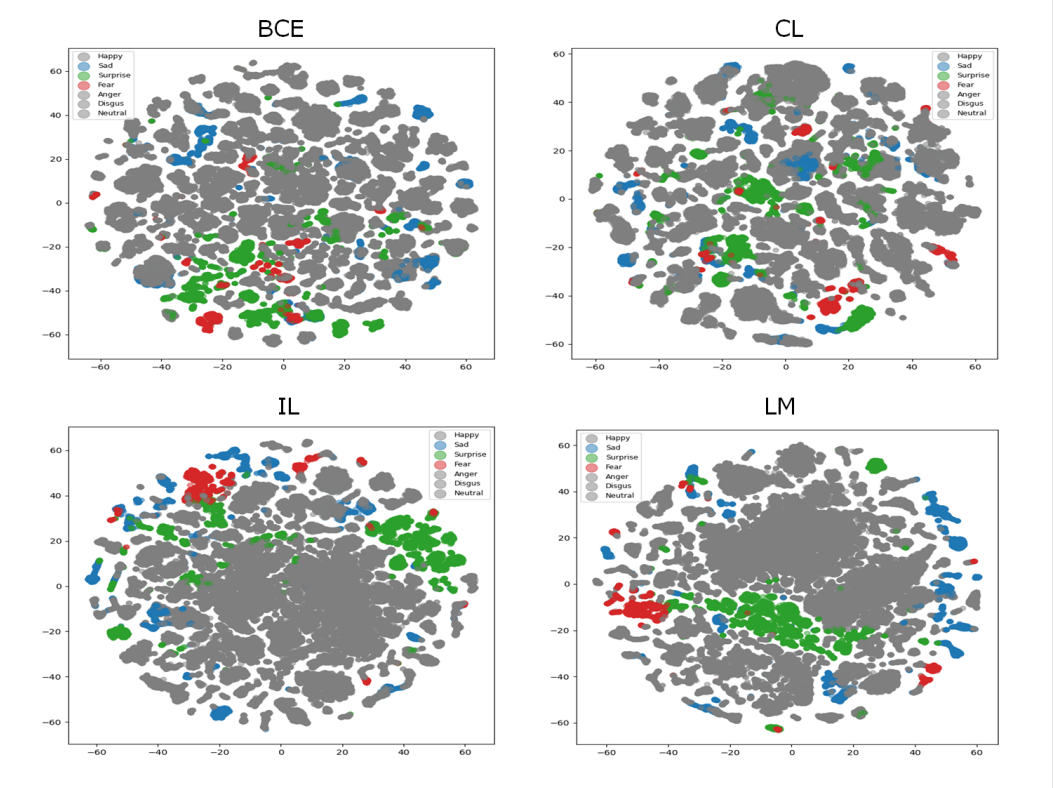

Figura 5.3 Embeddings space evolution at epoch 40 for all synthetic emotions. BCE (Binary Cross Entropy) , CL (Center Loss), IL (Island Loss) and LM (Large Margin). Figure from [47]

Tabela 5.7 Comparison with state of the art for DISFA. DA stands for domain adaptation in the current proposal. LM stands for Large Margin. F1-all is the average of all AUs. F1-sparse is the average of sparse AUs. The best results are represented in bold. The resullts obtained with LM are highlighted with gray color. Table from [47]

Method	F1-all	F1-sparse
DA (DISFA unlabeled) PreActRes18 - - pretrained LM Loss	55.4	**51.2**
DA (DISFA unlabeled) PreActRes18 - - not pretrained LMS Loss	50.4	44.7
DA (DISFA unlabeled) PreActRes18 Center Loss [10]	46.6	38.9
DA (DISFA unlabeled) PreActRes18 Island loss [11]	47.3	40.0
DA (DISFA unlabeled) PreActRes18 Ring Loss [12]	46.0	39.4
DRML [48]	26.7	14.2
ROI [49]	48.4	23
DSIN [20]	53.6	42.6
JAA [50]	56.0	48.2
SRERL [51]	55.9	45.6
MLT-RM [52]	**60.1**	46.6
UGN-B [53]	60.0	49.65

Tabela 5.8 Results for EmotioNet dataset. Comparison with other clustering losses and state-of-the-art. F1-all is the average of all AUs. F1-sparse is the average of sparse AUs . DA = Domain Adaptation; BCE = Binary Cross Entropy; LM = Large Margin; CL = Center Loss; IL = Island Loss; RL = Ring Loss. The best results are represented in bold. The best resullts obtained with LM are highlighted with gray color. Table from [47]

Method	**F1-all**	**F1-sparse**
SV-BCE +PrActRes18	47.08	35.19
SV-BCE +PrActRes18 - CE(SynExpr)	48.48	36.75
SV-LM	50.16	38.90
DA - LM -Alexnet	49.04	35.96
DA - LM +PrActRes18	52.12	40.74
DA - LM +PrActRes18 Imagenet pretrain	54.31	43.25
DA - NM+PrActRes18 DISFA pretrain	**55.89**	**45.58**
DA - CL[10] +PrActRes18	48.92	36.83
DA - RL[12] +PrActRes18	49.14	36.34
DA - IL[11] +PrActRes18	50.38	38.20

Capitolul 6

Image Retrieval

Retrieving similar images is an increasingly common task in various fields such as: image search engines or medical applications. Over time, a variety of descriptors have been used to efficiently describe the visual information from images. In this chapter, the possibility of using a large margin loss to acquire a set of improved descriptors via convolutional networks was investigated.

6.1 Database

The database used in the experiments is known as Places365 [54] and contains approximately 1.8 million images organized into 365 categories. Images contains various scenes all over the world. The training set has a number between 3000 and 5000 images. The test set has 900 pictures for each class.

6.2 Related Work

Before convolutional networks became popular, different variants of descriptors for image retrieval were sought. Among those who appeared before therise of deep learning are local binary pattern [55], histogram od oriented gradients [56] or color histograms. Then, descriptors that identify similar points such as SIFT [57] and SURF [58] were used for retrieval tasks.

In recent years, improved information-highlighting descriptors have emerged, such as the histogram of visual words [59]. Obviously with the development convolutional networks, more and more individuals desired to use densely connected layers as embeddings [60, 61] .

Scenario/Metric [%]	mAP-5-query	mAP-8-query	Top-1-err-rate	Top-5-err-rate	Accuracy	AUC – PR curve
CE-pre-trained-baseline	29.93	28.35	66.74	37.53	53.10	9.17
LM-fine-tuned	29.89	28.56	66.99	37.95	54.69	9.23
LM-from-scratch	**31.35**	**29.98**	66.18	37.79	53.51	11.18
Resnet-152 [19]	-	-	-	-	54.74	-

Tabela 6.1 Experimental results on Places365 (mAP-mean average precision; AUC – area under curve; CE- cross entropy, LM- large margin).

Table from [62]

6.3 Proposed method

A ResNet convolutional network was used for the problem of finding similar images and it was trained in Places365 database [54]. After this phase, the penultimate fully connected layer was extracted ans it was used as a descriptor for the retrieval part.

Using the large margin concept (section 4.4), the network was trained to increase the spatial density of the descriptive space. The outcomes are shown in table 6.1. Among the performance metric used are: mean averaged precision(mAP), Top error rate and are under precision-reacall curve. It can be seen that the results are only slightly in favor of the presented method.

To determine the utility of the large margin function, three new scenarios were developed containing separable data, non-separable data, and data from both categories at the embeddings level. Below is an illustrative figure of the data space before training (Figure 6.1).

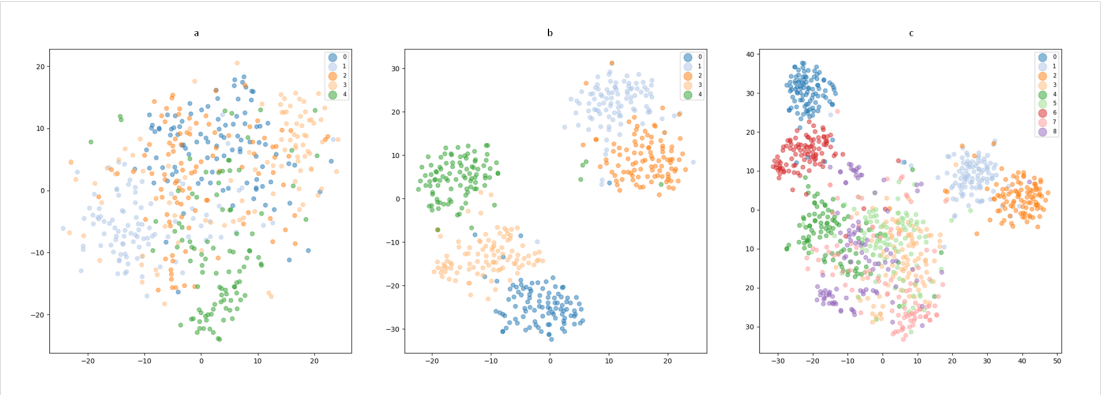

Figura 6.1 t-SNE representation of the 3 different situations. a – Non separable data scenario; b- Separable data scenario; c- Both separable and non separable data scenario
Figure from [62]

Table 6.2 shows the results for the 3 analyzed scenarios. It should be noted that the large margin function produces superior results when the data space is densely populated. For the worst scenario with barely separable data, the performance metrics

Scenario/Metric [%]	mAP-5-query	mAP-10-query	Top-1-err-rate	Top-5-err-rate	Acc	AUC – PR curve
CE-separable data	92.82	92.93	6.85	2.46	94.82	82.81
LM-separable data	92.82	92.93	6.85	2.46	94.82	**85.16**
CE-non-separable data	55.83	56.00	43.61	12.69	67.20	37.25
LM-non-separable data	59.75	58.97	40.60	13.45	70.08	**44.75**
CE-both separable and non-separable data	72.78	71.50	25.82	7.13	78.30	52.12
LM-both separable and non-separable data	72.11	72.43	22.15	9.32	79.85	**58.96**

Tabela 6.2 Experimental results on Places365 for the new scenarios(mAP-mean average precision; AUC – area under curve; CE- cross entropy, LM- large margin).

Table from [62]

are clearly better than in the case of using descriptors from a network trained with classical cross-entropy.

To illustrate the importance of the large margin concept, refer to figure 6.2. Here, it is presented the evolution of the descriptive space for cross entropy and large margin during the training process. As can be seen, the space is significantly more compact with the large margin, indicating that the network will provide more useful descriptors for the image retrieval problem.

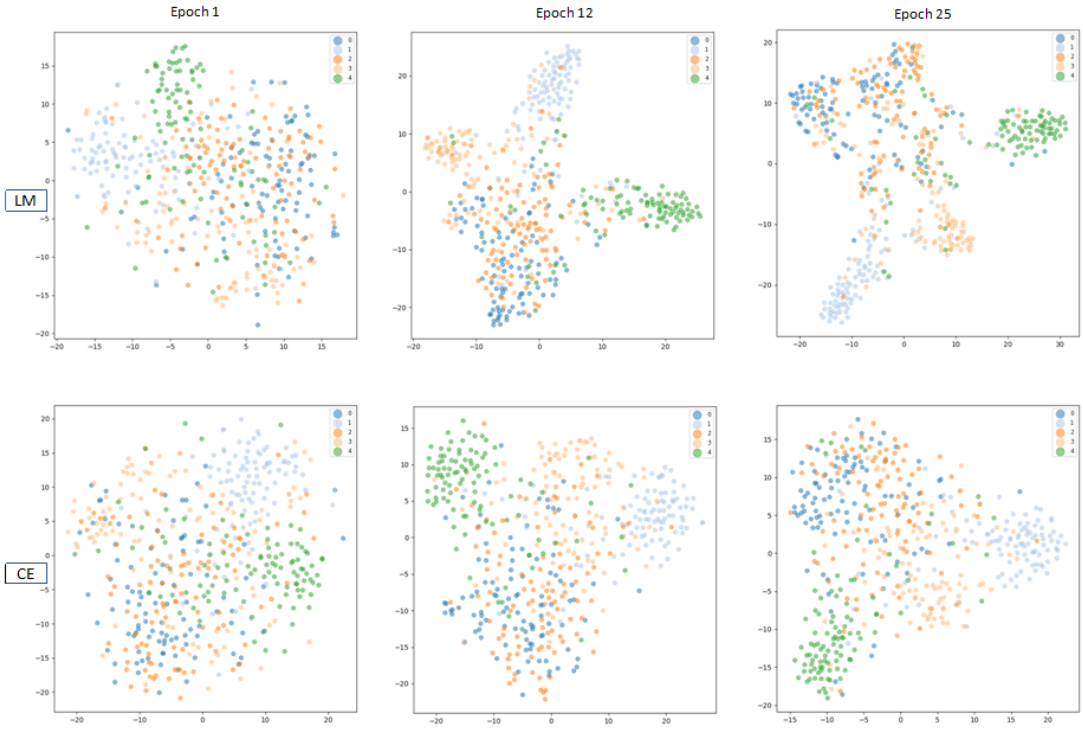

Figura 6.2 The modification of the descriptive space during the training process for the case with non-separable data (LM –up, CE- bottom)
Figure from [62]

The previously mentioned idea can be verified in figure **??**. It can be seen the first five images retrieved for a given query image. Using the trained network's embeddings in conjunction with the large margin loss increases the number of returned images that belong to the real class. However, there are instances when cross entropy provides superior results. This aspect proves that for tangled data even the concept of large margin is not enough.

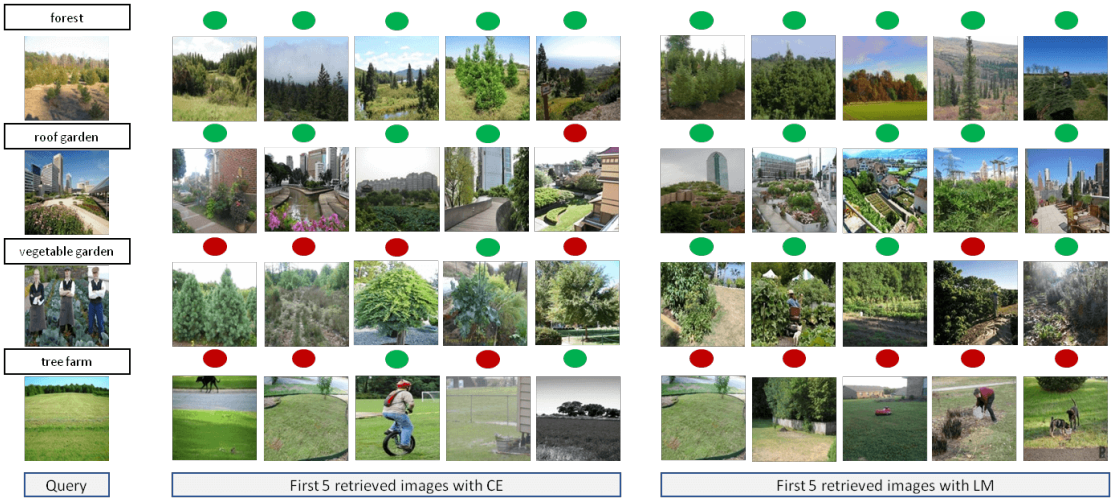

Figura 6.3 Examples of first 5 images retrieved with CE and LM for inseparable data scenario. Red bullets mark retrieved images with a different class compared to query image. Green bullets describe correctly retrieved images
Figure from [62]

Even though the large margin has its limitations, it achieves a better compaction of the embeddings space. The reference set used for the retrieval problem is depicted at a descriptive level in figure 6.4. When the network is trained with large margin loss, the data is more compact and less overlapping, facilitating the system's retrieval decisions for images with similar content. Figure 6.5 shows an increase in the area under the precision-recall curve, which confirms the utility of this technique for problems with tangled data.

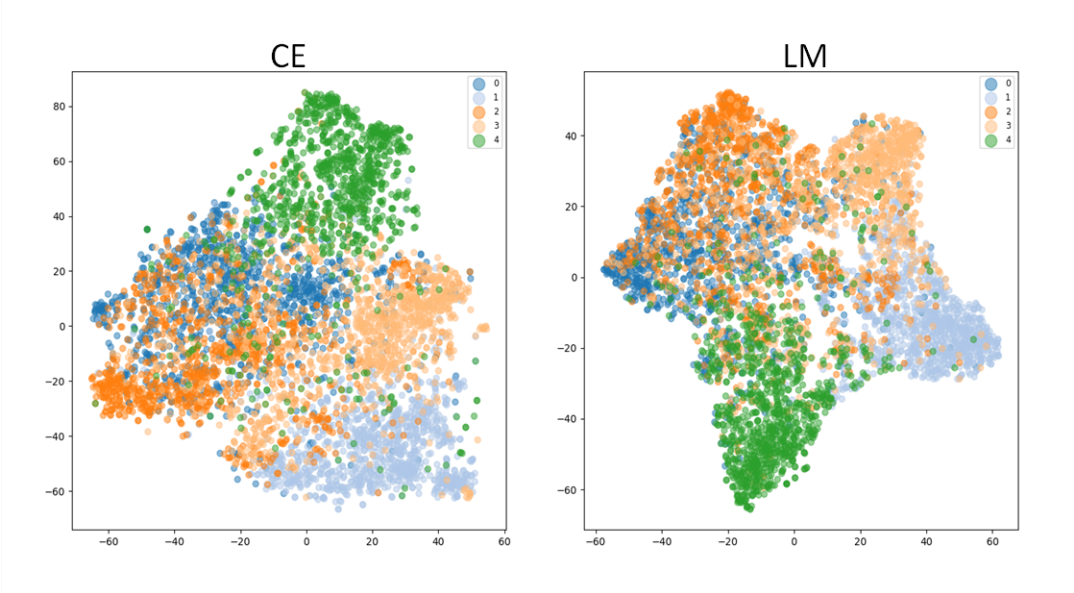

Figura 6.4 The descriptors representation associated with the reference set with CE (left) and LM (right) for non-separable data scenario
Figure from [62]

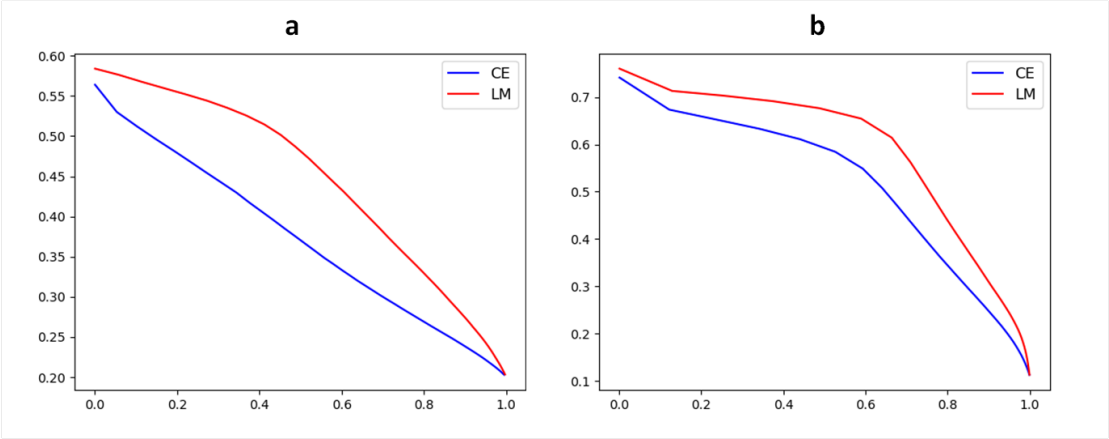

Figura 6.5 PR- curves for LM and CE. a- non separable data scenario; b- both separable and non-separable data scenario)
Figure from [62]

Capitolul 7

Conclusions

7.1 Obtained results

In this thesis two essential research topics were addressed: facial expression recognition and image retrieval. The first also focused on the detection of facial movements, known as action units. Methods based on semi-supervised learning/domain adaptation, loss functions for a better discrimination in the embeddings space, and novel augmentation/regularization techniques were developed. The successful outcomes validated the potential of the proposed methods.

The image retrieval was the second area of interest. In this instance, it was determined whether a new loss function (large margin) can more effectively organize the descriptors used to discover similar images. It has been demonstrated that this concept is particularly beneficial when descriptive-level data are highly overlapping.

7.2 Contributions

- A new technique for recognizing facial expressions and action units was approached, which has several original components. A domain adaptation solution to connect facial movements and discrete expressions was used to benefit from the potential of semi-supervised learning. In addition, a clustering loss of the descriptive space was used for increased performance. [13]

- A solution for the recognition of facial expressions was proposed , focusing on the use of a new regularization method based on randomization injection in the gradient. [42]

- A new way that combines annotated data and unannotated data with a technique to increase the number of samples was proposed. The algorithm was tested not only in the case of facial expressions, but also on standard benchmarks [36]

- The effectiveness of a clusterization loss of the embedddings' space was tested for image retrieval task. In this case, a more effective grouping of descriptors significantly increased the retrieval performance for the scenarios with tangled data. [62]

- The method from [13] was extended to several data sets containing action units. Here, the proposed loss function was studied in relation to other similar functions from the specialized literature and it proved to be more efficient. It was also demonstrated that the proposed loss contributes to a better recognition of the action units that appear less often in the data sets, confirming its ability to decipher the data. [47] [63]

- For all the proposed algorithms, an extensive comparison with the literature was carried out. Methods similar to the contextual approach were discussed in order to gain a more objective idea about the effectiveness of the proposed algorithms.

7.3 Publications

- Andrei Racoviteanu, 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 Racoviteanu, 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 Badea, Constantin Vertan, Corneliu Florea, Laura Florea, and Andrei Racoviteanu. Improving small convolutional neural networks with semi-supervised learning. UPB Scientific Bulletin, Series C: Electrical Engineering, pg Series C, Vol. 84, Iss. 3, 2022, pp 107-119

- Andrei Racoviteanu, 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

- M. Boeru, A. Racoviţeanu and C. Florea, "Facial Expressions Recognition by Structuring the Embeddings Space," 2021 International Conference on e-Health and Bioengineering (EHB), Iasi, Romania, 2021, pp. 1-4

- B. Stoica, L. Florea, A. Bădeanu, A. Racoviţeanu, I. Felea and C. Florea, "Visual saliency analysis in paintings," 2017 International Symposium on Signals, Circuits and Systems (ISSCS), Iasi, Romania, 2017, pp. 1-4
- Badea, M., Florea, C., Racoviţeanu, A., Florea, L., Vertan, C. (2023). Timid semi–supervised learning for face expression analysis. Pattern Recognition, 138, 109417.
- Florea, Corneliu, et al. "Automatic Real-Estate Image Analysis for Retrieval and Classification." Bulletin of the Polytechnic Institute of Iași. Electrical Engineering, Power Engineering, Electronics Section 68.2 (2022): 35-45.
- 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 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
- Andrei Racoviteanu, Corneliu Florea, Mihai-Sorin Badea. Large margin loss for Image Retrieval. Accepted to UPB Scientific Bulletin, Series C: Electrical Engineering
- Andrei Racoviteanu, Corneliu Florea, Laura Florea, and Constantin Vertan. Normalized Margin Loss for Action Unit Detection. Submitted to MVAP

- Project "Technologies and innovative video/audio systems for the recognition/identification of people and simulated behavior" - SPIA-VA, PN-III-P2-2.1- SOL-2016-02-0002
- Project "TRANSLATE" , TE 66/2020, PN-III-P1-1.1-TE-2019-0543.
- Project “Innovative Artificial Intelligence systems in the field of real estate portals” - online number 137-221-A2, MySMIS number: 129132
- Project "OPTIM research" through Human Capital Sectoral Operational Program 2014-2020 - nr. 62461/03.06.2022, SMIS code 153735.

7.4 Future work

In the field of semi-supervised learning, the development possibilities are numerous, regardless of the chosen topic (easy expressions or image retrieval). Although the potential is huge, this approach also has a number of limitations. The most important one is that it fails to outclass the supervised algorithms for datasets with sufficient labels. However, for the datasets with few annotations, the situation changes radically for the better.

The domain adaptation technique used for facial expressions can also be used in other contexts. Face landmark localization and head pose classification are practical instances, as the three angles of the head can be conveyed relative to face landmarks; another example is image captioning and object detection, where the captions are derived from a particular set of objects.

In the case of image retrieval, the clustering loss of the embeddings' space proved to be very effective. In this context, it could also be extended to databases with facial expressions. Given that many expressions are similar at the descriptive level, this solution has some potential.

Bibliografie

[1] 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.

[2] 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.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

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

[5] 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.

[6] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[7] 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, page 896, 2013.

[8] 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.

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

[10] 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.

[11] 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.

[12] Y. Zheng, D. Pal, and M Savvides. Ring loss: Convex feature normalization for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5089–5097, 2018.

[13] 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.

[14] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[15] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[16] 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.

[17] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 435–442, 2015.

[18] Changde Du, Changying Du, Hao Wang, Jinpeng Li, Wei-Long Zheng, Bao-Liang Lu, and Huiguang He. Semi-supervised deep generative modelling of incomplete multi-modality emotional data. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 108–116, 2018.

[19] Zixing Zhang, Jing Han, Jun Deng, Xinzhou Xu, Fabien Ringeval, and Björn Schuller. Leveraging unlabeled data for emotion recognition with enhanced collaborative semi-supervised learning. *IEEE Access*, 6:22196–22209, 2018.

[20] Ciprian Corneanu, Meysam Madadi, and Sergio Escalera. Deep structure inference network for facial action unit recognition. In *Proceedings of the european conference on computer vision (ECCV)*, pages 298–313, 2018.

[21] 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.

[22] 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.

[23] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. *Advances in neural information processing systems*, 28, 2015.

[24] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27, 2014.

[25] Augustus Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.

[35] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[36] 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.

[37] James C Bezdek, Robert Ehrlich, and William Full. Fcm: The fuzzy c-means clustering algorithm. *Computers & geosciences*, 10(2-3):191–203, 1984.

[38] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[39] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

[40] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[41] Vikas Verma, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Arno Solin, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *Neural Networks*, 145:90–106, 2022.

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

[43] Andrei Racoviteanu, Corneliu Florea, Laura Florea, and Constantin Vertan. Randomization injection for efficient transfer in face expression recognition. In *Submitted to Applied Intelligence*, 2023.

[44] Valentin Vielzeuf, Corentin Kervadec, Stéphane Pateux, Alexis Lechervy, and Frédéric Jurie. An occam's razor view on learning audiovisual emotion recognition with small training sets. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 589–593, 2018.

[45] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2018.

[46] Yanling Gan, Jingying Chen, and Luhui Xu. Facial expression recognition boosted by soft label with a diverse ensemble. *Pattern Recognition Letters*, 125:105–112, 2019.

[47] Andrei Racoviteanu, Corneliu Florea, Laura Florea, and Constantin Vertan. Normalize margin loss for action unists detection. In *Submitted to MVAP*, 2023.

[48] 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.

[49] W. Li, F. Abtahi, and Z. Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6766–6775, 2017.

[50] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European conference on computer vision (ECCV)*, pages 705–720, 2018.

[51] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8594–8601, 2019.

[52] Jiyuan Cao, Zhilei Liu, and Yong Zhang. Cross-subject action unit detection with meta learning and transformer-based relation modeling. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

[53] Tengfei Song, Lisha Chen, Wenming Zheng, and Qiang Ji. Uncertain graph neural networks for facial action unit detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5993–6001, 2021.

[54] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[55] 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.

[56] 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.

[57] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

[58] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. *Lecture notes in computer science*, 3951:404–417, 2006.

[59] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, IEEE International Conference on*, volume 3, pages 1470–1470. IEEE Computer Society, 2003.

[60] Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval. *Remote Sensing*, 9(5):489, 2017.

[61] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3304–3311. IEEE, 2010.

[62] Andrei Racoviteanu, Corneliu Florea, and Mihai Badea. Large margin loss for image retrieval. In *Accepted to UPB Scientific Bulletin, Series C: Electrical Engineering*, 2023.

[63] 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.

[64] 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.