MoDuL: Deep Modal and Dual Landmark-wise Gated Network for Facial Expression Recognition

Sacha Bernheim, Estephe Arnaud, Arnaud Dapogny, Kevin Bailly

MoDuL: Deep Modal and Dual Landmark-wise Gated Network for Facial Expression Recognition

Sacha Bernheim[1] and Estèphe Arnaud[1] and Arnaud Dapogny[2] and Kévin Bailly[1,2]
[1] Institut des systèmes intelligents et de robotique (ISIR), Sorbonne Université, Paris, France
[2] Datakalab, Paris, France

***Abstract*— Automatic facial expression recognition (FER) is a challenging computer vision problem that finds a number of applications in human-computer interaction. Most recent FER approaches are deep-learning based and involve the extraction of two types of features from face images: geometric features (e.g. distances between aligned facial landmarks) and appearance features extracted using convolutional neural networks applied on patches extracted around each landmark. In this paper, we explore the use of gating networks to learn an optimal combination of these two modalities (modal gate). Furthermore, we also design landmark-wise gates to adaptively weight each landmark as well as the corresponding patch contribution. The proposed MoDuL architecture achieves state-of-the-art results on several FER databases with negligible computational overhead.**

I. Introduction

Facial expression is a fundamental way for human beings to communicate their emotions and feelings. Automatic expression analysis is a challenging problem and impacts important applications such as human-computer interaction [1], health-care [12], surveillance [14], self-driving cars [18], etc. Nevertheless, recognizing facial expressions with a high accuracy is not an easy task due to the complexity and variability of facial expressions and their interpretation which can be quite subjective [30].

Most of classical and deep learning oriented approaches rely on two kinds of features. On the one hand, geometric features that are based on facial landmarks positions, provide information about head pose and facial expression variations. These features are thus essential to recognize facial expression, but they mainly depend on the reliability of the facial landmarks tracker. Moreover, they does not contain all the needed information. On the other hand, appearance features capture information relative to the image texture and can provide complementary information. For example, a frown is not only characterized by the displacement of the eyebrows landmarks with respect to the eyes position, but also by the appearance of glabella wrinkles.

Given the small number of available data to train models to recognize facial expressions, methods based on heterogeneous features still lead to the best results. This implies to carefully weight the contribution of each kind of features depending of the emotion to predict. Furthermore, each landmark do not have the same relevance in order to recognize a given expression (e.g. recognizing happiness will mostly rely on the occurrence of a smile in the lower part of the face).

This work was supported by the French National Agency (ANR) in the frame of its Technological Research JCJC program (FacIL, project ANR-17-CE33-0002).

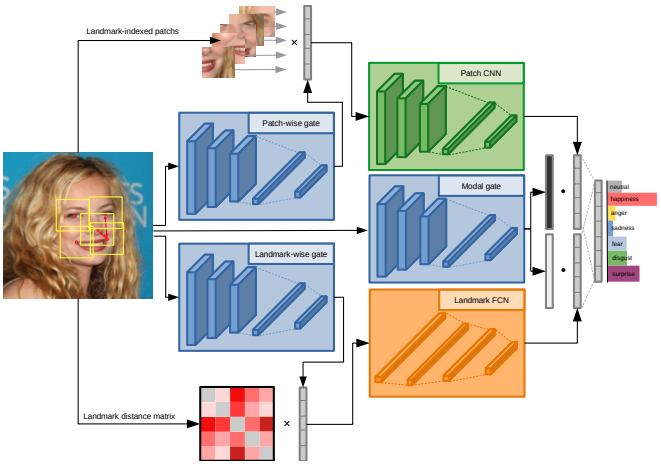

Fig. 1. Architecture of the proposed method. For each landmark point, a patch is extracted as long with the distances with the other landmark points. Those inputs are weighed differently by the outputs of a patch-wise gate and a landmark wise gate. Finally, before the concatenation of both subnets, we weigh another time with a modal gate.

In this paper, we introduce MoDuL, a deep neural network that uses both adaptive region-weighting *via* landmark/patch-wise gates and modality combination using a modal gating scheme, as illustrated on Figure 1. To sum it up, the contributions of this work are three-fold:

- A landmark-wise gate that allows to flexibly weight the contribution of each face region for sub-networks that processes geometric and appearance features.
- A modal gate that weights these two heterogeneous modalities, allowing to better capture the complementarity between appearance and texture information.
- We empirically show that MoDuL achieves state-of-the-art on several FER datasets with very little computational overhead.

II. Related work

One of the most challenging step for automatic facial expression recognition is to extract discriminative features that best describe the appearance and the geometry changes during the emotion production. Histogram of Oriented Gradients (HOG) [6], histograms of Local Binary Patterns (LBP) [27], Gabor wavelets [2], among others, are the most

efficient features designed by humans when it comes to appearance, and can deal with challenging issues of FER such as illumination changes and face occlusion [31]. They have been implemented a lot with classical machine learning frameworks such as SVM [22] or Random Forests [5]. Appearance features have often be combined with geometric features extracted from facial landmarks. Different fusion schemes have been proposed to combined these heterogeneous features, ranging from feature concatenation and late fusion to more elaborate strategies (e.g. 2K-SVM [23] or MK-SVM[26])

But most of current state of the art methods use deep learning in order to perform high accuracy facial expression recognition, as shown in the following survey [17]. Among the deep-based methods, convolutionnal neural networks (CNNs) are the most popular. Indeed, instead of handcrafting predetermined features, CNN can learn from the data collection which patterns are the most relevant for a specific task. Two different approaches exist: the hollistic methods and the patch-based ones. In the former, the faces are treated as a whole. Whereas in the latter, the face is divided into subregions [25]. In this case, the face can be divided into equal parts around facial landmarks, or it is possible to get those patches through a particular sampling strategy [31].

One of the major difficulty of the FER domain is related to the high intra-class (e.g. two different persons can express sadness very differently) and low inter-class (e.g. fear can be interpreted as disgust depending on the context) variances. This problem is especially worsened as the number of training samples is low with noisy labels. To address this issue, IL-CNN [3] substitutes traditional softmax loss for the handcrafted *Island Loss*. With a quite simple holistic approach composed of 3 convolutions, IL-CNN showed very high accuracy performance. Following the same purpose, Identity-Aware Convolutional Neural Network (IACNN [24]) uses a metric learning inspired approach based on pairwise images to reduce inter-subject variations. This kind of approach focus on one particular source of variation to the detriment of other sources such as illumination and pose.

Other approaches seek to be more robust to these variations either with data augmentation [20], or by combining model predictions [30], [15], [19]. For example, IPA2LT [30] trained multiple networks with heterogeneous images and labels. A last deep neural network is used to combine the decision arising from these heterogeneous networks.

Compared to these methods, [13] and [11] rely on different modalities, namely the raw image and the landmark positions, to infer the emotions. DTAGN [13], for instance, uses two deep networks. They each receive an image sequence along with the facial landmarks as inputs. The first network focuses over appearance and the second one over the geometry between the different facial landmarks. Finally, the output of these networks are integrated using a weighted summation.

In the same vein, our model rely on complementary modalities. In particular we introduce a gate-based end-to-end multimodal fusion scheme that is trained to combine geometric and appearance information according to the input image. It also uses a patch-based approach and weights the importance of each subregions, depending on its informativeness computed from the input face image. Those different weighting methods, implemented as gates, also allows to interpret easily the results by helping to visualize on which part the model focuses to perform the task.

Other existing approaches combine geometric and local appearance features [32], [33]. However, in these works, importance of each features is estimated during the training phase and remains constant for each prediction. On the contrary, in [6] weights are dynamically adjusted according to the current image. The weighting strategy is handcrafted (the weights are outputted by external auto-encoders that indicates whether the landmark is occluded or not, or at least, very different from its classical appearance). Weights are used to modulate the posterior prediction of each local prediction tree. Thus it gives no clues about the impact of each geometric and appearance features. In contrast, our approach learn to dynamically modulate the input features. The gating strategy is learned in a end to end manner and depends on both the appearance of the current image and the prediction of the network.

III. Overview of MoDuL

An overview of MoDuL is provided on Figure 1. Similar to what is done in the literature, MoDuL is composed of a joint network between a fully-connected network (FCN - Section III-A.1) based upon geometric features extracted from facial landmarks and appearance features provided by a CNN indexed by these landmarks (Section III-A.1). In Section III-B we describe our landmark-wise gate for both networks. Then, beyond a plain joint network, these primary networks can be combined using a third modal gate III-C. All the primary and gating networks are illustrated on 2.

A. Primary networks

As a preprocessing, we extract facial feature points using an off-the-shelf IntraFace [28] tracker that provides the location of N_l landmarks denoted as $\{f_i\}_{i=1,\dots,N_l}$. Also, in what follows, the whole face image is denoted as $\mathcal{I}$.

1) Distance FCN (Figure 2-left): The first part of the network consists of fully-connected layers whose inputs consist in the $N_l(N_l-1)/2$ distances between facial landmarks f_i and f_j, normalized by the inter-ocular distance (iod, *i.e.* the distance between the average of all landmarks belonging to left and right eyes). These features are intrinsically invariant to in-plane rotations and scaling. We denote $D_{i,j}$ the distance matrix defined as:

$$D_{i,j}(\mathcal{I}) = \frac{\|f_i - f_j\|_2}{iod(\mathcal{F})}, \forall i,j \in [\,0, N_l - 1\,]^2 \qquad (1)$$

Then Distance FCN is a function $F^{dist} : D(\mathcal{I}) \mapsto F^{dist}(D(\mathcal{I})) \in \mathbb{R}^{+n}$ where n is the number of output neurons. In practice, we model F^{dist} using a number of fully-connected layers with ReLU activation. Also, as matrix D is symetric we only extract the upper triangular elements

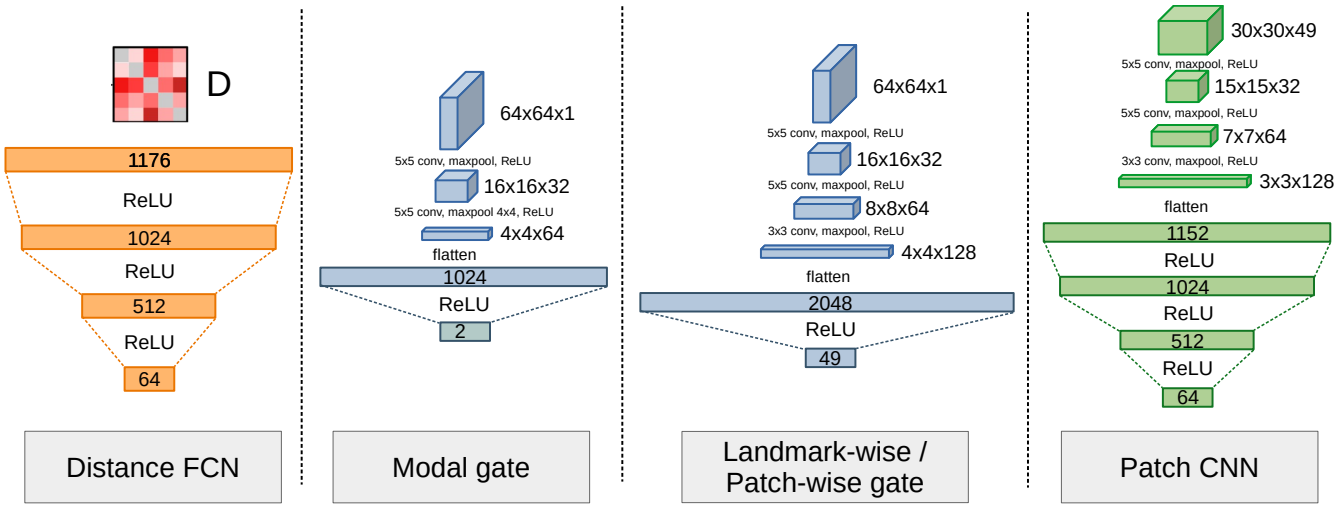

Fig. 2. Hyperparameters of each component of MoDuL. The **Distances FCN** applies a number of fully-connected layers with ReLU activation to extract geometric features. The **Patch FCN** processes the crops around each landmark with CNN and FCN to extract appearance features. The gates processes the whole face crop. The gate networks take as input the whole image and return weights for each modality (**modal gate**) and landmark (**landmark-wise/patch-wise gate**)

and reshape them into a vector to feed F^{dist}. Such model using only geometric information may lack the capacity to discriminate most subtle expressions, e.g. *anger* vs *sadness*, hence the need to incorporate appearance information.

2) Patch CNN (Figure 2-right): In order to do so, we crop the faces images around each landmark with a window size of 30% of the iod of the face image, then we rescale those crops to a constant size of 30×30 pixels. The patches are denoted as $\mathcal{P} = \{\mathcal{P}\}_i(\mathcal{I}), \forall i \in [0, N_l - 1]$ Thus, each input image consist of a 30x30 image with N_l channels, each channel corresponding to a specific landmark.

We then define our other primary network, namely the Patch CNN, as a function $F^{patch} : \mathcal{P}(\mathcal{I}) \mapsto F^{patch}(\mathcal{P}(\mathcal{I})) \in \mathbb{R}^{+n}$. F^{patch} is then modeled using a number of CNN followed by FC layers. Appearance information helps decipher subtle expressions, however, for both the primary networks, the information extracted within the features is highly redundant and these networks may struggle to extract relevant features. For that matter, we propose an adaptive region contribution weighting scheme.

B. Adaptive region contribution weighting

We propose to weight the input of each basic network using gating functions. Gating functions for the distance FCN and Patch CNN take the face image $\mathcal{I}$ as input and are defined as:

$$G^{dist} : \mathcal{I} \mapsto G^{dist}(\mathcal{I}) \in [0, 2]^{N_l} \tag{2}$$

and

$$G^{patch} : \mathcal{I} \mapsto G^{patch}(\mathcal{I}) \in [0, 2]^{N_l} \tag{3}$$

respectively.

These functions are illustrated on Figure 2 (2^{nd} from left), and output a weight for each landmark: in practice, they consist of CNN and FCN layers with ReLU activation, except for the last layer that has sigmoid activation to scale the gate outputs. Note that the gate sigmoid activation had mean 0.5, we multiply its output by 2 so that the expected value of the output remains the same with or without the gate. For the distance FCN, we multiply matrix $D_{i,j}(\mathcal{I})$ by the gate vector. The weighted distance FCN can thus be written as:

$$F^{dist}_w = F^{dist}(D(\mathcal{I})G^{dist}(\mathcal{I})) \tag{4}$$

For the Patch-CNN we multiply each input channel (*i.e.* the crop extracted around one landmark) by the gate output corresponding to this landmark:

$$F^{patch}_w = F^{patch}(\mathcal{P}(\mathcal{I}) \odot G^{patch}(\mathcal{I})) \tag{5}$$

Where $\odot$ stands as the replicated Hadamard product accross spatial dimensions of the crops. Note that the primary networks F^{dist} and F^{patch}, as well as the corresponding gating networks G^{dist} and G^{patch} are learned jointly in an end-to-end manner, allowing to find an optimal landmark weighting in both cases.

C. Joint network with adaptive modality weighting

Mixing the two primary networks gives a more accurate and robust network. A naive way to do so consists in concatenating the last hidden layer of both networks before the output layer:

$$F^{joint}(\mathcal{I}) = F^{dist}(D(\mathcal{I}))||F^{patch}(\mathcal{P}(\mathcal{I})) \tag{6}$$

We can also use the region-weighted networks by applying:

$$F^{joint}_w(\mathcal{I}) = F^{dist}_w(D(\mathcal{I}))||F^{patch}_w(\mathcal{P}(\mathcal{I})) \tag{7}$$

Where $||$ denotes the concatenation operator. The caveat of such naive design is that the two modalities are not equally

informative: hence, the joint network can have a tendency to rely too heavily on one type of feature, preventing to correctly learn the other sub-network. To avoid this, we advocate the use of another gating function that we call the modal gate. As before, the modal gate is defined as:

$$G^{mod} : \mathcal{I} \mapsto G^{mod}(\mathcal{I}) \in [0,1]^2 \tag{8}$$

G^{mod}, however, has a softmax activation in its last layer. The equation of the naive joint network where only the two final modalities are weighted is thus:

$$F^{mod}(\mathcal{I}) = G^{mod}(0)F^{dist}(D(\mathcal{I}))||G^{mod}(1)F^{patch}(\mathcal{P}(\mathcal{I})) \tag{9}$$

This so-called modal gate is illustrated on Figure 2 (3^{rd} from left. The final MoDuL equation that uses both modal gate and region contribution weighting can be written as:

$$F^{MoDuL}(\mathcal{I}) = G^{mod}(0)F_w^{dist}(D(\mathcal{I}))||G^{mod}(1)F_w^{patch}(\mathcal{P}(\mathcal{I})) \tag{10}$$

with F_w^{dist} and F_w^{patch} both defined in Section III-B. In what follow, we compare the accuracy of all these networks, namely the primary networks F^{dist} and F^{patch}, their region-weighted counterparts F_w^{dist} and F_w^{patch}, as well as the joint networks F^{joint}, F^{mod}, F_w^{joint} and F^{MoDuL}. For each of these networks, we append at softmax layer with as many output units as the number of FEs to predict after the last hidden layer.

IV. Experiments

In this Section, we introduce FER databases IV-A and experimental setup to ensure reproducibility of the results IV-B. We then perform ablation study in Section IV-C to highlight the contribution of each component of MoDuL. In Section IV-D we compare our method to existing approaches, showing that it provides state-of-the-art results. Finally, in Section IV-E we provide insight on the model behavior.

A. Databases

We validate our MoDuL approach on three databases from different applicative contexts and annotation.

The CK+ or **Extended Cohn-Kanade databse** [21] contains 123 subjects, each recorded producing various expressions. Those records contain an evolution from neutral to one of the 6 universal emotions described by Ekman [8]. From those records we extracted 309 sequences, each one corresponding to one of the six basic expressions, and use the three first and last frames from the records for training.

The BU-4DFE database [29] contains 101 subjects, each one recorded producing the six basic emotions with moderate head pose variations. Expressions usually have lower intensity and greater variability than in CK+. We manually selected neutral and apex of expression frames, for a total of 8219 examples for training.

The JEMImE-Paris and JEMImE-Nice databases [10] contains respectively 1458 and 2323 examples labeled with FE quality. The concatenated database is referred as **JEMImE-All** and contains 3781 examples. For the classification task, we decided to keep only the samples whose quality is higher than 7 over 10, making respectively 534 and 1312 examples from Nice and Paris.

B. Experimental setup

We trained 7-class networks (neutral and the six basic FEs) on CK+ and BU-4DFE databases and 4-classes networks on JEMImE-All with FEs neutral, happiness, anger and sadness.

Models are evaluated using the (unweighted) overall accuracy over the test set, along with the average per class accuracy (trace of the confusion matrix) to take into account the discrepancies in class repartition in the test sets. As it is classical in the literature, we implemented 10-fold subject-independent cross validation. Furthermore, on each dataset, we have more examples of FEs neutral and happiness than any other expressions. In order to handle class imbalance at train time, we trained our models using a rejection resampling method: at each beginning of an epoch during the train phase, we equilibrate each class by downsampling examples belonging to the majority classes and oversampling examples of the minority classes. As compared to alternative solutions, such as class weighting, this method leads to similar results, with reduced computation time, as explained in [4].

Optimization is applied with ADAM optimizer [16] with batch size 16 and $\beta_1 = 0.9$, the learning rate follows a polynomial decay with power 0.9. For MoDuL as well as the joint region-weighted networks we applied 25000 steps with a base learning rate of 0.001. For the other networks we applied 20000 updates with base learning rate 0.01, as it provided better results in practice.

Finally, it is important to notice that all our networks were trained from scratch without any pre-training.

C. Ablation study

a) Performance assessment:: in order to walidate the contribution of each separate component of MoDuL we conducted an ablation study on CK+ database. The results are summarized in table I.

TABLE I
UNWEIGHTED (UW) AND WEIGHTED (W) ACCURACY ON CK+ DATABASE. BEST RESULT IN BOLD, SECOND BEST UNDERLINED.

Model	uw. acc(%)	w. acc(%)
Distances	88.51	84.11
Patches	88.79	82.97
Joint	91.81	90.27
Joint, modal gate	93.25	91.25
Distances, Region-weighted	91.16	87.72
Patches, Region-weighted	88.56	86.31
Joint, Region-weighted	<u>93.76</u>	<u>93.08</u>
MoDuL	**94.09**	**93.22**

First, the region-weighted networks are more efficient than their unweighted counterparts. The region-weighted patch CNN, for instance, is equivalent to the basic patch CNN in terms of unweighted accuracy, however it is significantly

TABLE II

CONFUSION MATRIX ON CK+ DATABASE FOR JOINT NETWORK.

Joint	Ne	Ha	An	Sa	Fe	Di	Su
Ne	**91.6**	1	3.2	1.9	0	2.3	0
Ha	0.4	**99.6**	0	0	0	0	0
An	13.9	0	**79.9**	3.6	0	2.5	0
Sa	12.3	0	6.8	**78.6**	2.2	0	0
Fe	7.6	0	0	0	**91.7**	0	0.7
Di	4.1	0	3.1	0	0	**92.8**	0
Su	2.5	0	0	1.3	0	0	**96.2**

TABLE III

CONFUSION MATRIX ON CK+ DATABASE FOR MODUL.

MoDuL	Ne	Ha	An	Sa	Fe	Di	Su
Ne	**95.1**	0	1.9	1.6	0.7	0.7	0
Ha	0.9	**99.1**	0	0	0	0	0
An	11.6	0	**82.6**	4.5	0	1.3	0
Sa	9.3	0	3.1	**87.6**	0	0	0
Fe	3.1	3.3	0	0	**92.9**	0	0.7
Di	3.1	0	0	0	0	**96.9**	0
Su	2.5	0	0	0	0	0	**97.5**

better in terms of weighted accuracy. The region-weighted distance FCN is better that the distances FCN both in terms of unweighted and weighted accuracy. This holds for both the joint and joint modal networks. This indicates that the proposed region-weighting method helps to discriminate FEs by allowing to flexibly put emphasis on certain face regions.

Second, the Joint network is more accurate than the single modality (distance and patch) networks. The addition of the modal gate increases both accuracy metrics, showing that it allows to better capture the inter-modal complementarity. Finally, MoDuL is the best performing network in terms of both unweighted and weighted accuracy. Table II and III shows the confusion matrices for both the joint network and MoDuL, respectively. As one can see, the addition of both the landmark/patch-wise and modal gates allows to better discriminate the FEs, and particularly subtle FEs such as neutral, anger and sadness.

TABLE IV

NUMBER OF PARAMETERS FOR EACH MODEL.

Model	# parameters (M)
Distances	1.77
Patches	2.79
Joint	5.76
Joint, modal gate	5.82
Distances, Region-weighted	3.20
Patches, Region-weighted	3.02
Joint, Region-weighted	6.22
MoDuL	6.27

b) Number of parameters:: we also reported the number of parameters for each model in table IV. The modal gate only increases the number of parameters by 0.9% while the dual landmark-wise gate increases this number by 7.9%. We also have to take into consideration that this last augmentation is mostly due to the fact that the number of inputs for the distances subnetwork has doubled. Thus, this improvement of performance only represents an augmentation of 8.8% of the number of parameters between our baseline and our best model, which is quite reasonable. Last but not least, notice how our network has few parameters compared to state-of-the-art network [24] that employs bigger networks (e.g. ResNet): in what follows, we show that despite having fewer parameters, MoDuL provide satisfying accuracy on multiple datasets.

D. Comparison with state-of-the-art approaches

In this section, we evaluate our MoDuL network on several datasets and compare the performance with recent state-of-the-art facial expression recognition methods.

Compared to CK+, BU4-DFE images depict more subtle and challenging samples. Moreover, we test our model on the JEMImE dataset as it constitutes a testbed for our model on a different domain (children vs adult).

Table V summarizes the comparison between MoDuL and other state of the art approaches. We can notice that MoDuL outperforms approaches on CK+ such as IPA2LT (91.67%, overall accuracy) [30], which specifically address the inter-subject variance issue. The overall accuracy of IACNN is 95.35% but is hardly comparable with our approach as IACNN uses twice as much data (it pre-traines the CNN on FER-2013 dataset [9]).

TABLE V

COMPARISON WITH STATE-OF-THE-ART METHODS. REPORTED SCORE WITH (*) INDICATES UNWEIGHTED ACCURACY.

Method	CK+	BU-4DFE	JEMImE-All
IPA2LT [30]	91.67	-	-
WLS-RF [6]	94.3*	75.0	-
IACNN [24]	95.35*	-	-
RF [7]	-	-	81.9
Joint	91.81	70.95	80.55
MoDuL	**93.22**	**80.73**	**82.51**

Similarly, MoDuL clearly outperforms WLS-RF [6] on BU-4DFE by more than 5%, which was still the best score over this dataset, to the best of our knowledge. In addition, WLS-RF uses a leave-one-person-out methodology, which tends to give better results than our 10-fold subject independent cross validation.

We can observe more precisely the behavior of the model through the confusion matrices for BU-4DFE and JEMImE datasets in Table VI and Table VII respectively. It shows that our framework has way better results for subtle expressions: anger, fear and disgust.

Thus, by wisely weighting the different sub-regions and feature modalities, our framework easily deals with the inter and intra-class variance issue of FER, hence outperforming state of the art approaches over the most subtle FEs, which are the hardest to recognize with automatic methods. Furthermore, MoDuL substantially enhances the accuracy compared to the nave Joint approach in all cases, thus appears as an interesting solution to fuse heterogeneous modalities for FER, such as distance and patches.

TABLE VI

CONFUSION MATRIX ON BU-4DFE DATABASE FOR MODUL.

MoDuL	Ne	Ha	An	Sa	Fe	Di	Su
Ne	**87.4**	1.1	4.7	3.7	1.7	0.7	0.7
Ha	6.4	**87.6**	1.1	0.3	3.3	0.3	1
An	11.8	0.9	**74.9**	6.5	0.2	5.8	0
Sa	19	1.1	14.7	**62.5**	1.7	0.7	0.2
Fe	14.9	4.6	2.3	4.6	**52.8**	13	7.8
Di	4.7	0.9	7.7	2.5	7.8	**74.6**	1.7
Su	1.5	1.1	1.1	0	2.3	1.2	**92.8**

TABLE VII

CONFUSION MATRIX ON JEMIME DATABASE FOR MODUL.

MoDuL	Ne	Ha	An	Sa
Ne	**83.7**	2.3	6.5	7.6
Ha	3.3	**90.4**	1.9	4.4
An	8.9	4.8	**78**	8.3
Sa	12.2	6.4	10.7	**70.6**

E. Gate behavior introspection

In this section, we introspect the models to show the model behavior. To this end, we pass each test example through the models and record the average gate values (for modal gate and region weighting) for each expression class on CK+.

TABLE VIII

VALUES FOR THE MODAL GATE FOR EACH FE ON CK+ DATABASE.

	distance	patch
neutral	0.244	0.756
happiness	0.491	0.509
anger	0.312	0.688
sadness	0.231	0.769
fear	0.254	0.746
disgust	0.388	0.612
surprise	0.291	0.709

1) Modal gate introspection: Table VIII provides the average values outputted by the modal gate. As one can see, for FEs happiness and disgust, which involve large facial deformations, the distances FCN is given a lot of weight, whereas this is not the case for more subtle FEs, such as neutral, anger, sadness or fear. Overall, the distance FCN is given much less weight than the patch CNN: as such the relative performance of e.g. MoDuL as compared to the region-weighted joint model can be explained by the fact that the modal gate allows to give more weight to the patch CNN, whose extracted appearance features allows to more efficiently disentangle the more subtle FEs.

2) Region weighting introspection: Figure 3 provides a visualization of the facial landmarks importance, per type of feature and emotion, outputted by the region-weighted gates. Happiness is mostly characterized by distances related to the mouth. The interpretation is quite straightforward as happiness is mainly characterized by smiles. This observation also applies to surprise that his distinguishable by the raising of eyebrows and the mouth opening.

For more subtle and challenging facial expression such as anger, the information is more spread all over the face and across the geometry and the appearance.

V. CONCLUSION

In this work, we introduce MoDuL, a deep neural network that uses adaptive weighting scheme *via* multiplicative gating functions to more efficiently select relevant face region contributions and modalities from networks that extracts information from heterogeneous features, such as face landmarks and patches. MoDuL has a simple and straight-forward implementation and competes with other state of the art approaches and can be learned from scratch.

The different gating schemes that we presented could easily be used over other applications than FER. Especially, we propose an easy to implement way to merge heterogeneous modalities and better capture the complementarity thereof. These modalities can consist of images coming from heterogeneous sensors such as LIDAR or depth cameras, or data generated *via* high-level processing such as face landmarks, body pose estimation, or high-level contextual information such as image quality or environmental lighting. In the same vein, the idea to weight the input features could be applied in a more generic way to weight certain regions of an input image or feature map, which is also a direction that we would like to explore in future work.

REFERENCES

[1] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction, 2003. Conference on Computer Vision and Pattern Recognition Workshop.

[2] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. Fasel, and J. R. Movellan. Recognizing facial expression: Machine learning and application to spontaneous behavior, 2005. CVPR, pages 568573.

[3] J. Cai, Z. Meng, A. Shehab, K. Zhiyuan, L. J. OReilly, and Y. Tong. Island loss for learning discriminative features in facial expression recognition, 2017. 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018).

[4] C. Chen, A. Liaw, and L. Breiman. Using random forest to learn imbalanced data, 2004. Department of Statistics, UC Berkeley tech report.

[5] A. Dapogny, K. Bailly, and S. Dubuisson. Dynamic pose-robust facial expression recognition by multi-view pairwise conditional random forests, 2016. IEEE Transactions on Affective Computing.

[6] A. Dapogny, K. Bailly, and S. Dubuisson. Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection, 2018. International Journal of Computer Vision, Volume 126, Issue 24, pp 255271.

[7] A. Dapogny, C. Grossard, S. Hun, S. Serret, J. Bourgeois, H. Jean-Marie, P. Foulon, H. Ding, L. Chen, S. Dubuisson, O. Grynszpan, D. Cohen, and K. Bailly. Jemime: A serious game to teach children with asd how to adequately produce facial expressions, 2018. 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018).

[8] P. Ekman and W. Friesen. Constants across cultures in the face and emotion., 1971. Journal of personality and social psychology, vol. 17, no. 2, p.124.

[9] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, and D. L. et al. Challenges in representation learning: A report on three machine learning contests, 2013. ICML, pages 117124. Springer.

[10] C. Grossard, L. Chaby, S. Hun, H. Pellerin, J. Bourgeois, A. Dapogny, H. Ding, S. Serret, P. Foulon, M. Chetouani, L. Chen, K. Bailly, O. Grynszpan, , and D. Cohen. Children facial expression production: Influence of age, gender, emotion subtype, elicitation condition and culture, 2018. Frontiers in Psychology.

[11] B. Hasani and M. H. Mahoor. Facial expression recognition using enhanced deep 3d convolutional neural networks, 2017. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

[12] M. S. Hossain. Patient state recognition system for healthcare using speech and facial expressions, 2016. Journal of Medical Systems.

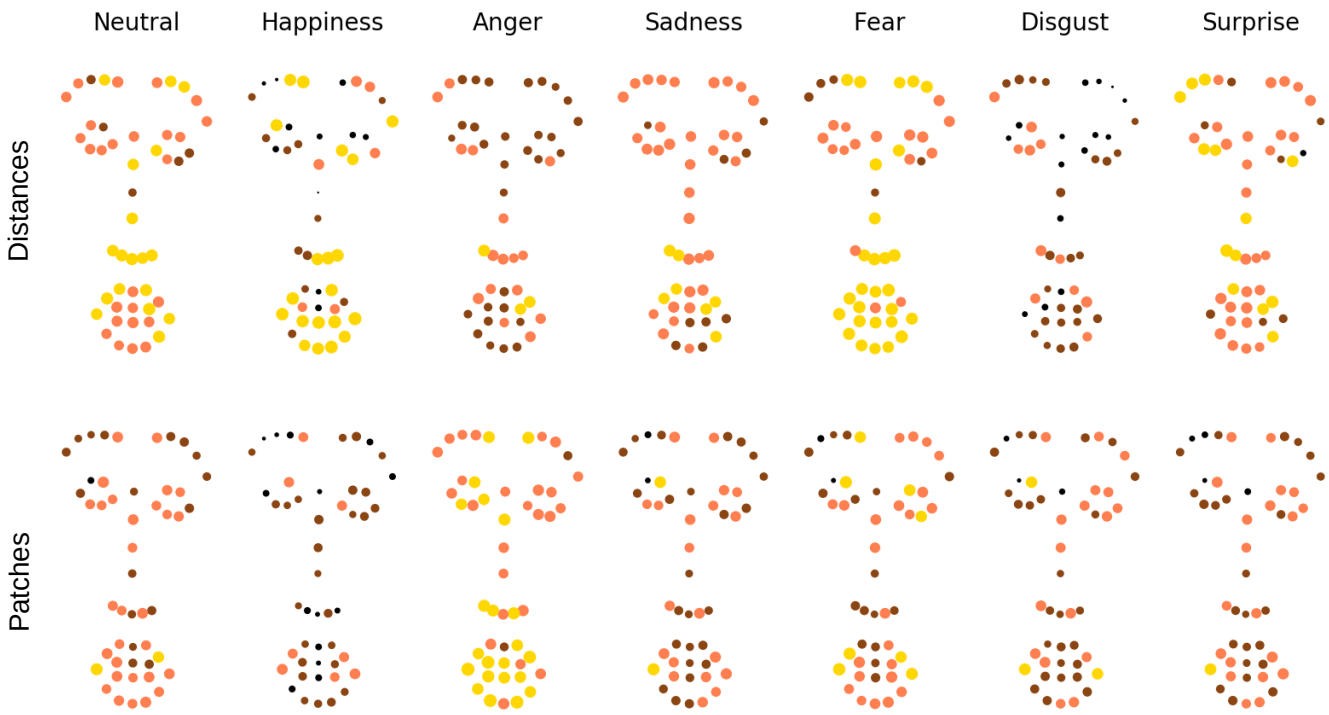

Fig. 3. Visualization of facial landmarks importance per type of feature and emotion. Big, yellow circles indicate a large importance of a feature and face area. Conversely, small and black circles suggest a low relevance. Best viewed in color.

[13] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[14] B. Kamgar-Parsi, W. Lawson, and B. Kamgar-Parsi. Toward development of a face recognition system for watchlist surveillance, 2011. IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 33 , Issue: 10.

[15] B.-K. Kim, H. Lee, J. Roh, and S.-Y. Lee. Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 427–434. ACM, 2015.

[16] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization, 2015. International Conference on Learning Representations.

[17] S. Li and W. Deng. Deep facial expression recognition: A survey, 2018.

[18] C. L. Lisetti and F. Nasoz. Affective intelligent car interfaces with emotion recognition, 2005. 11th International Conference on Human Computer Interaction.

[19] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812, 2014.

[20] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition*, 61:610–628, 2017.

[21] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, and Z. Ambadar. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, 2010. International Conference on Computer Vision and Pattern Recognition Workshops, pp. 94101.

[22] Y. Luoab, Cai-ming, and W. Y. Zhangb. Facial expression recognition based on fusion feature of pca and lbp with svm, 2013. Optik - International Journal for Light and Electron Optics Volume 124, Issue 17, Pages 2767-2770.

[23] H. Meng, B. Romera-Paredes, and N. Bianchi-Berthouze. Emotion recognition by two view svm 2k classifier on dynamic facial expression features, 2011. IEEE International Conference on Automatic Face Gesture Recognition (FG 2011).

[24] Z. Meng, P. Liu, J. Cai, S. Han, , and Y. Tong. Identity-aware convolutional neural network for facial expression recognition, 2017. FG, pages 558565.

[25] L. Nwosu, H. Wnag, J. Lu, I. Unwala, X. Yang, and T. Zhang. Deep convolutional neural network for facial expression recognition using facial parts, 2017. IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress.

[26] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. Facial action recognition combining heterogeneous features via multikernel learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4):993–1005, 2012.

[27] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study, 2009. Image and Vision Computing Volume 27, Issue 6, Pages 803-816.

[28] X. Xiong and F. D. la Torre. Supervised descent method and its applications to face alignment, 2013. International Conference on Computer Vision and Pattern Recognition, pp. 532539.

[29] L. Yin, X. Chen, T. W. Yi Sun, and M. Reale. A high-resolution 3d dynamic facial expression database, 2008. International Conference on Automatic Face and Gesture Recognition, pp. 16.

[30] J. Zeng, S. Shan, , and X. Chen. Facial expression recognition with inconsistently annotated datasets, 2018. The European Conference on Computer Vision (ECCV), pp. 222-237.

[31] L. Zhang, D. Tjondronegoro, and V. Chandran. Random gabor based templates for facial expression recognition in images with facial occlusion, 2014. Neurocomputing Volume 145, Pages 451-464.

[32] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on pattern analysis and machine intelligence*, 27(5):699–714, 2005.

[33] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang. Joint patch and multi-label learning for facial action unit and holistic expression recognition. *IEEE Transactions on Image Processing*, 25(8):3931–3946, 2016.