Facial Expression Recognition Under Partial Occlusion via Confidence-weighted Local Subspace Random Forests

Arnaud Dapogny[1]
Kévin Bailly[1]
Séverine Dubuisson[1]

[1] Sorbonne Universités, UPMC Univ Paris 06, CNRS, ISIR UMR 7222, 4 place Jussieu 75005 Paris

arnaud.dapogny@gmail.com

Résumé

La reconnaissance automatique des expressions faciales constitue une tâche difficile, impliquant la prise en compte d'une grande variabilité dans les différences morphologiques inter-individus, ainsi que la survenue éventuelle d'occultations partielles du visage. Dans ce cas, l'apparence de certaines régions du visage diffère drastiquement, ce qui est source d'erreurs de classification pour un classifieur Random Forest global. A contrario, dans cette étude, nous proposons d'apprendre des Random Forest définies sur des sous-espaces locaux représentant des régions du visage. De plus, ces prédictions locales peuvent être pondérées par des scores de confiance fournies par un réseau autoassociatif, modélisant l'apparence des données non occultées de manière hierarchique. Nous démontrons via de un certain nombre d'expérimentations, couvrant divers scénarios de reconnaissance d'expressions faciales, que la méthode proposée améliore les performances de l'état de l'art dans un cadre non occulté, en plus de constituer une solution élégante pour la prise en compte des occultations.

Mots Clef

Expressions faciales, occultations, random forests locales.

Abstract

Fully-Automatic Facial Expression Recognition (FER) from still images is a challenging task as it involves handling large interpersonal morphological differences, and as partial occlusions can occasionally happen. In such a case, the appearance of the face locally differ from the non-occluded pattern, causing a global Random Forest model trained on the whole face to misclassify the expression. In this work, we propose to train Random Forests upon spatially defined local subspaces of the face. Additionnaly, those local predictions can be weighted by confidence scores provided by an autoencoder network. This network is trained to capture the manifold of the non-occluded training data in a hierarchical way. Extensive experiments on multiple FER benchmarks show that the proposed approach improves the recognition accuracy compared to a global model as well as state-of-the-art methods in the non-occluded case. It also leads to interesting perspectives towards the design of occlusion-robust FER systems.

Keywords

Facial Expressions, Occlusions, Local Random Forests.

1 Introduction

Automatic Facial Expression Recognition (FER) is a key to many human-computer applications such as consumer robotics or social monitoring. As stated in [25], FER from still images is a challenging task as there may exist large variability in the morphology or in the expressiveness of different persons. Furthermore, countless configurations of partial occlusion can occasionally happen (e.g. hand or accessories). In parallel, annotation of expressive images is a time-consuming process, limiting the amount of training data available to cope with this variability.

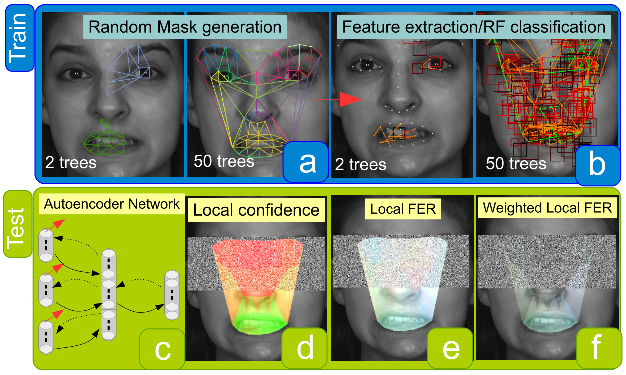

FIGURE 1 – Flowchart of our WLS-RF framework. Randomized trees are trained upon local subspaces corresponding to facial masks, on which binary features candidates are generated. When testing on an occluded image (bottom row), local trees are weighted by confidence scores given by an autoencoder network to provide a robust prediction.

Most Recent approaches covering FER from still images work in controlled conditions on frontal, lab-recorded environments [15, 24]. Shan *et al.* [19] evaluated the recognition accuracy of Local Binary Pattern features. Zhong *et al.*

[28] proposed to learn active facial patches that are relevant for FER. Zhao *et al.* [27] design a multitask framework for simultaneously performing facial alignment, head pose estimation and FER. Such approaches showed satisfying results in constrained scenarios, but they likely face difficulties on more challenging benchmarks [6]. Furthermore, none of them addresses the problem of facial occlusions that are likely to happen in such unconstrained cases.

Kotsia *et al.* [12] studied the impact of human perception of facial expressions under partial occlusions, and the predictive capacities of automated systems thereof. Cotter *et al.* [4] used sparse decomposition-based classification to perform FER on corrupted images. Ghiasi *et al.* [8] uses a discriminative approach for facial feature point alignment under partial occlusion. Those approaches relies on explicitly incorporating synthetic occluded data in the training process, and thus struggles to deal with realistic, unpredicted occluding patterns. Zhang *et al.* [26] trained classifiers upon random Gabor-based templates of non-occluded data. They evaluated their algorithms on synthetically occluded face images and showed that their approach leads to a better recognition rate when the same occluded examples are used for training and testing. Should this not be the case, unpredicted mouth/eye occlusions still lead to a significant loss of performance. Huang *et al.* [10] proposed to automatically detect the occluded regions using sparse decomposition residuals. However, the proposed approach may not be flexible enough, as the occlusion detection only outputs binary decisions, and as the face is divided into only three subparts (eyes, nose and mouth). This limits the capacities of the method to deal with unpredicted occlusions. Finally, another approach consists in learning generative models of non-occluded faces, as it was done by Ranzato *et al.* [16]. When testing on a partially occluded face image, the occluded parts can be generated back and expression recognition can be performed. The pitfall of such approach is that training can be computationally expensive and does not allow the use of heterogeneous features (e.g. geometric/appearance descriptors).

In this work, we propose to address the problem of FER under partial occlusion by using a Weighted Local Subspace Random Forest (WLS-RF) framework described in Figure 1. During training, local subspaces are generated under the form of random facial masks (a), onto which binary candidate features can be selected (b) to train randomized trees. When testing on a potentially occluded image, a hierarchical autoencoder network (c) is used to capture the local manifold of non-occluded faces around separate aligned feature points. The reconstruction error outputted by such network provides a confidence measurement of how close a face region lies from the training data manifold (d), with high and low confidences depicted in green and red respectively). The output local predictions (e) of local randomized trees are thus weighted by the confidence scores (f). The contributions of this work are thus three-fold :

1. A hierarchical autoencoder network for learning local non-occluded face manifolds, which can be used to provide local confidence measurements.
2. A method for training random trees upon local subspaces of the face, which consists in generating random masks covering a specified fraction of the face.
3. A real-time Weighted Local Subspace Random Forests framework that improves the state-of-the-art and is robust to unpredicted occlusions.

The rest of the paper is organized as follows : in Section 2 we describe the proposed autoencoder network architecture and how it is trained to capture the local manifold around facial feature points. Section 3 describes how we train our Local Subspace Random Forest classifiers by generating heterogeneous binary feature candidates. We also describe how those local models can be effectively combined and weighted by the confidence measurement to produce robust expression estimates. Finally, in Section 4 we show that our approach significantly improves the state-of-the-art on multiple datasets, both on the non-occluded and occluded cases. Finally, Section 5 provides a conclusion as well as a few perspectives raised in the paper.

2 Manifold Learning of non-occluded faces with hierarchical autoencoder network

2.1 Network architecture

In this section, we present how we can use an autoencoder network to model local face pattern manifold, and to output a local confidence measurement that will be used for FER. Autoencoders are a particular type of neural network that can be used for manifold learning. Compared with other approaches such as PCA [11], autoencoders offer the advantage to theoretically be able to model complex manifolds using non-linear encoding and regularization criteria such as denoising [21] or contractive penalties [17]. Furthermore, they benefit from efficient training using stochastic gradient descent, as well as the possibility of online fine-tuning for subject-specific calibration.

As shown in Figure 2, we use a 2-layer architecture, with a first encoding at the feature point level and a second one at the face subpart level. Indeed, the occlusions of neighbouring points are closely related. Thus, by encoding the local texture in a hierarchical way, we can more efficiently capture the relationships between such correlated occlusions. To do so, we first extract HOGs within the neighbourhood of each feature point aligned on the face image $\mathcal{I}$. The choice of modeling a manifold of HOG patterns rather than gray levels stems from the fact that HOGs are used for both the alignment of facial feature points, as well as for RF classification. Thus, the reconstruction error of these patterns provides a confidence measurement that is relevant for both tasks. Additionally, in order to ensure fast processing, we use integral feature channels to

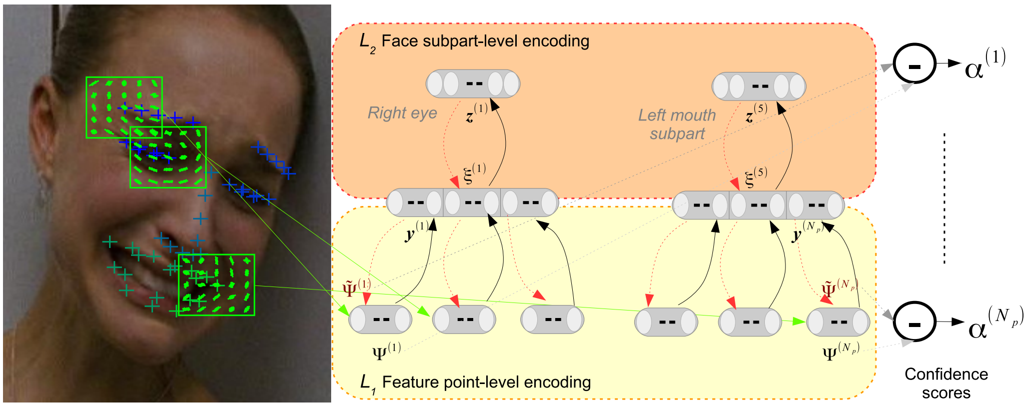

FIGURE 2 – Architecture of our hierarchical autoencoder network. The network is composed of 2 layers : the first one (L_1) captures the texture variations around the feature points. The second one (L_2) is defined over 5 face subparts, each of which embraces points whose appearances are closely related. The network outputs a confidence $\alpha^{(p)}$ for each feature point.

extract the HOGs. The local descriptor $\mathbf{\Psi}^{(k)}$ for a specific feature point k consists in the concatenation of gradient magnitudes and quantized orientation values in 5×5 cells around this feature point, with a total window size equal to a third of the inter-ocular distance. This descriptor of dimension 225 then feeds the N_p autoencoders (one per feature point) of the first layer (L_1) which are trained to reconstruct non-occluded patterns. Because occlusion of local patterns extracted at the feature point level are not independent (*i.e.* a feature point close to an occluded area is more likely to be occluded itself), we employ a second layer (L_2) of autoencoders, that are trained to reconstruct non-occluded patterns of groups of encoded feature point descriptors. Those groups represent five face subparts (left and right eyes, nose, left and right parts of the mouth) from which the local patterns are closely related. Specifically, L_1 is composed of 125 units for each landmark. L_2 layer for a feature point group contains $65 \times N$ units ($\frac{1}{2}$ compression), where N is 12, 12, 8, 11 and 11 respectively for left/right eye, nose and left/right mouth areas.

2.2 Training the network

Autoencoders are trained in an unsupervised way, one layer at a time, by optimizing a reconstruction criterion. The input descriptor $\mathbf{\Psi}^{(k)}$ at feature point k is first encoded via the L_1 encoding layer into $h^1(\mathbf{\Psi}^{(k)})$, which is the output of a first neuron layer with a sigmoid activation. This intermediate reconstruction can thus be reconstructed by applying an affine decoder with tied input weights : $\tilde{\mathbf{\Psi}}^{(k)} = g^1 \circ h^1(\mathbf{\Psi}^{(k)})$. The set of K encoded descriptors $\{h^1(\mathbf{\Psi}^{(k)})\}_{k=1...K}$ associated to feature points $k = 1...K$ that belong to the face subpart m are concatenated to form the input $\xi^{(m)}$ of the layer L_2 for that subpart. Once again, the input of the L_2 layer is successively encoded into an intermediate representation $h^2(\xi^{(m)})$ and decoded in the same way into a reconstructed version $\tilde{\xi}^{(m)} = g^2 \circ h^2(\xi^{(m)})$.
Each layer is trained separately using stochastic gradient descent and backpropagation, by optimizing the squared $\mathcal{L}_2$-loss between an input and its reconstruction through the network. We tried various combinations of training hyperparameters and the best reconstruction results were obtained by applying 15000 stochastic gradient updates with alternating sampling between the expression classes in the databases. Indeed, we want the network to be able to reconstruct local variations of all possible expressive patterns on an equal foot. We also use a constant learning rate of 0.01 as well as a weight decay of 0.001, which provides good results in testing. Finally, we found that adding 25% random masking noise provided satisfying results. From a manifold learning perspective, the goal of using such denoising criterion is to learn to project corrupted examples (e.g. partially occluded ones, which lie further from the manifold) back on the training data manifold. Such example will be reconstructed closer to the training data and its confidence shall be smaller.

2.3 Local confidence measurement

Given a face image $\mathcal{I}$, we define the confidence $\alpha^{(k)}(\mathcal{I})$ for point k as a function of the $\mathcal{L}_2$-loss (*i.e.* the reconstruction error) between the HOG pattern $\mathbf{\Psi}(\mathcal{I})$ extracted from this point, and its reconstruction $\tilde{\mathbf{\Psi}}$ outputted by the network, after successively encoding by layers L_1 then L_2, and decoding in the opposite order. By abuse of notation :

$$\alpha^{(k)}(\mathcal{I}) = 1 - \frac{||\mathbf{\Psi}^{(k)} - g^1 \circ g^2 \circ h^2 \circ h^1(\mathbf{\Psi}^{(k)})||^2}{(||\mathbf{\Psi}^{(k)}|| + ||g^1 \circ g^2 \circ h^2 \circ h^1(\mathbf{\Psi}^{(k)})||)^2} \tag{1}$$

We used the normalized Euclidean distance as a confidence score as it was directly optimized during training. However, we experimented with other metrics such as RBF, which provided similar results. We introduce a confidence $\alpha^{(\tau)}(\mathcal{I})$ defined over triangles $\tau = \{k_1, k_2, k_3\}$ as :

$$\alpha^{(\tau)}(\mathcal{I}) = \min(\alpha^{(k_1)}(\mathcal{I}), \alpha^{(k_2)}(\mathcal{I}), \alpha^{(k_3)}(\mathcal{I})) \tag{2}$$

As highlighted in the following experiments, this triangle-wise confidence measurement can be used to weight LEPs to enhance the robustness to partial occlusions.

3 Local Subspace Random Forest

3.1 Learning local trees with random facial mask generation

Random Forests (RF) is a popular learning framework introduced in the seminal work of Breiman [1]. They have been used to a significant extent in computer vision and for FER tasks in particular [27, 5] due to their ability to nicely handle high-dimensional data such as images as well as being naturally suited for multiclass classification tasks. In the classical RF framework, each tree of the forest is grown using a subset of training examples (bagging) and a subset of the input dimension (random subspace). Individual trees are then grown using a greedy procedure that involves, for each node, the generation of a number of binary split candidates that consist in features associated with a threshold. Each candidate thus defines a partition of the labelled training data. The "best" binary feature is chosen among all features as the one that minimizes an impurity criterion (which is generally defined as either the Shannon entropy or the Gini impurity). Then, the above steps are recursively applied for the left and right subtrees with accordingly rooted data until the label distribution at each node becomes homogeneous, where a leaf node can be set. As stated in [1], the rationale behind training each tree on a random subspace of the input dimension is that the prediction accuracy of the whole forest depends on both the strength of individual trees and on the independence of the predictions. Thus, by growing individually weaker (e.g. as compared to C4.5) but more decorrelated trees, we can combine these into a more accurate tree collection.
Following this idea, we propose an adaptation of the RF framework that uses Local Subspaces (LS) instead of the traditional Random Subspaces (RS). Each tree is trained using a restricted subspace corresponding to a specific part of the face. Then a global prediction is computed as an average of those local predictions. Using a combination of local classifiers offers multiple advantages over using trees defined over the whole face :

1. Local models (LS-RF) capture more diverse information by forcing the trees to use less informative features, that can still hold some predictive power.
2. We use the confidence outputted by the autoencoder network in Section 2 to weight the local responses for which the pattern lies further from the training data manifold (WLS-RF). This way, in case of occlusion or illumination changes, we can still use the information from the other face subparts.

In order to train the local trees, we first compute the mean shape $\bar{f}$ and the surface $s(\tau(\bar{f}))$ covered by each triangle τ on the mean shape, normalized by the total surface. For each tree t in the forest, we generate a face mask M_t defined over triangles τ. This mask is initialized with a single triangle τ_i randomly selected from the mesh. Then, neighbouring triangles are added until the total surface covered by the selected triangles w.r.t. $\bar{f}$ becomes superior to hyper-parameter R, that represents the surface that shall be covered by each tree. Figure 3 provides an illustration of such masks generated on the face mesh. As shown on Figure 3, R controls the locality of the trees. From a RF perspective, it allows to find a compromise between the strength of individual tree predictors, and the decorrelation between them. Thus, it plays a similar role as the number of features used to set each split node in a traditional RF. We show experimentally that setting $R = 0.1$ or 0.2 is a good tradeoff in the non-occluded case, in addition to bringing substantial improvements in the occluded case.

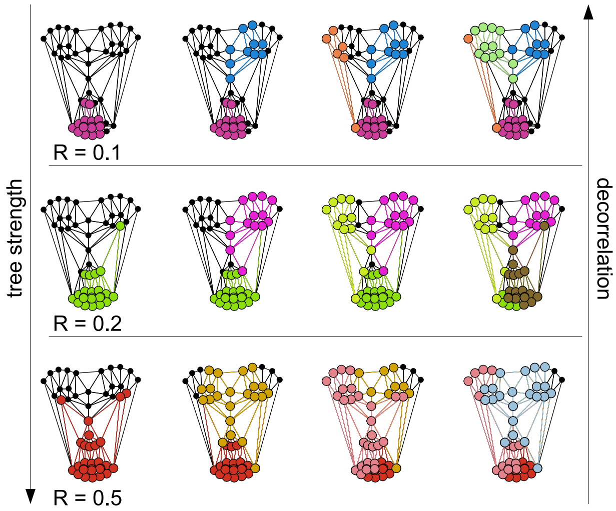

FIGURE 3 – Masks generated with $R = 0.1$, 0.2 and 0.5 for 1,2,3, 4 trees. Notice how the face is covered by independent masks upon which local trees can be trained. The setting of R allows to find a compromise between tree strength and decorrelation. Best viewed in color.

Then, as in the traditional RF induction procedure, we generate a bootstrap by randomly picking $2/3$ of the subjects for training tree t. In order to enforce class balance within the bootstraps, we downsample the majority classes. As compared to other methods for balancing RF classifiers (*i.e.* class weighting and upsampling the minority classes), downsampling leads to similar results while reducing the computational cost, as it is explained in [3]. Finally, tree t is grown on a subspace corresponding to the mask M_t, with the exact same on-the-fly heterogeneous feature generation scheme defined in [5].

3.2 Combination of local models

When testing, a face image $\mathcal{I}$ is successively rooted left or right for each tree t depending of the outputs of the binary tests stored in the tree nodes, until it reaches a leaf. The tree t thus outputs a probability vector $p_t(l|\mathcal{I})$ whose components are either 1 for the represented class, or 0 otherwise. Prediction probabilities are then averaged among the T trees of the forest (Equation (3)).

$$p(l|\mathcal{I}) = \frac{1}{T}\sum_{t=1}^{T} p_t(l|\mathcal{I}) \quad (3)$$

Those prediction probabilities are computed similarly for the global RF (RS-RF) and the LS-RF. However, for LS-RF the output probabilities of the trees have some degrees of locality and we can write the above formula as a sum over local probabilities :

$$p(l|\mathcal{I}) = p(l|\mathcal{I}) = \frac{\sum_{\tau} \alpha^{(\tau)} Z_\tau p(l|\mathcal{I},\tau)}{\sum_{\tau} \alpha^{(\tau)} Z_\tau} \quad (4)$$

With $\alpha^{(\tau)}$ being the confidence measurement outputted by the autoencoder network for triangle τ, for the confidence-weighted model WLS-RF, and Z_τ is the sum of prediction values for all expression classes l. For the LS-RF model, $\alpha^{(\tau)}$ is 1. Furthermore, we have :

$$p(l|\mathcal{I},\tau) = \frac{1}{Z_\tau}\sum_{t=1}^{T} \frac{\delta(\tau \in M_t) p_t(l|\mathcal{I})}{|M_t|} \quad (5)$$

Where $\delta(\tau \in M_t)$ is a function that returns 1 if triangle τ belongs to mask M_t, and $|M_t|$ is the number of times tree t is used in Equation (4). Note that the local responses $p(l|\mathcal{I},\tau)$ are not strictly limited to triangle τ but defined within its neighbourhood, with a radius that depends on hyperparameter R. The setting of R thus controls the locality of the trees, as it will be discussed in the experiments.

4 Experiments

In this section, we evaluate our approach on several FER benchmarks. In Section 4.2, we show results for FER on non-occluded data on three publicly available FER benchmarks that exhibit various degrees of difficulty, showing that our approach improves the state-of-the-art. Then, in Section 4.3, we report results on synthetically occluded images. Thus, we can precisely measure the robustness of our approach to occlusions, as well as the relevance of the confidence scores outputted by the autoencoder network.

For the tests on the CK+ and BU-4DFE databases, the autoencoder networks are trained in a cross-database fashion (*i.e.* training on CK+ and testing and BU-4DFE and vice versa). The RF classifiers are evaluated with Out-Of-Bag (OOB) error estimate which, according to [1], is an unbiased estimate of the true generalization error. Moreover, as stated in [2] this estimate is generally more pessimistic than traditional (e.g. 10-fold) cross-validation, further reflecting the quality of the results. In order to decrease the variance of the error we train large collection of trees ($T = 1000$). For the test on SFEW database we align 49 feature points with SDM [22] using the locations of the 5 provided facial landmarks. Also, we use the autoencoder network trained on CK+, as SFEW has several examples of occluded faces.

4.1 Datasets

The CK+ or **Extended Cohn-Kanade database [15]** contains 123 subjects, each one displaying some of the 6 universal expressions (*anger*, *happiness*, *sadness*, *fear*, *digust* and *surprise*) plus the non-basic expression *contempt*. Expressions are prototypical and performed in a controlled environment with no head pose variation. As it is done in other approaches, we use the first (*neutral*) and three apex frames for each of the 327 sequences for 8-class FER. As some approaches discard the frames labelled as *contempt*, we also report 7-class accuracy from 309 sequences.

The BU-4DFE database [24] contains 101 subjects, each one displaying 6 acted facial expressions with moderate head pose variations. Expressions are still prototypical but they are performed with lower intensity and greater variability than in CK+, hence the lower baseline accuracy. Sequence duration is about 100 frames. As the database does not contain frame-wise expression, we manually select neutral and apex frames for each sequence.

The SFEW or **Static Facial Expression in the Wild database [6]** contains 700 images from 95 subjects displaying 7 facial expressions in a real-world environment. Data was gathered from video clips using a semi-automatic labelling process. The strictly person-independent evaluation (SPI) benchmark is composed of two folds of (roughly) same size. As done in other approaches, we report cross-validation results averaged over the two folds.

4.2 FER on non-occluded images

In Tables 1, 2, 3 we report the average accuracy obtained by our Local Subspace Random Forest (LS-RF) and the confidence-weighted version (WLS-RF). We also compare with standard RF (RS-RF). For the local models, we set the locality parameter R to 0.2 (which means that each local model uses $1/5$ of the face total surface) which provided good classification results and robustness to occlusions.

Generally speaking, classification results of LS-RF are a little better than those of the RS-RF. Indeed, forcing the trees to be local allows to capture more diverse information. RS-RF relies quite heavily on the mouth region, but other areas (e.g. around the eyes, eyebrows and nose regions) may also convey information that can be captured by local models. Figure 4 displays the proportion of top-level features (*i.e.* the most “critical” features, selected at the root of the trees) over all triangles of the face area.

While more than 90% of the features extracted by RS-RF are concentrated around the mouth, the repartition for LS-RF is more homogeneous. Hence, LS-RF is less prone to misalignment of mouth feature points. Furthermore, weighting the local predictions (WLS-RF) using the confidence score from the autoencoder network allows to enhance the results on BU-4DFE and SFEW. The reason is that subjects from those datasets exhibit uncommon morphological traits, occlusion or lighting patterns. As such, more emphasis is put on reliable local patterns, resulting in a better overall accuracy. It also explains why the accuracy

TABLE 1 – CK+ database. † : CK database

CK+	7em	8em
LBP [19]	88.9†	-
CSPL [28]	89.9†	-
iMORF [27]	-	90.0
AUDN [14]	93.7	92.0
RS-RF	92.6	91.5
LS-RF	94.1	**93.4**
WLS-RF	**94.3**	**93.4**

TABLE 2 – BU-4DFE database

BU-4DFE	% Acc
BoMW [23]	63.8
Geometric [20]	68.3
LBP-TOP [9]	71.6
2D FFDs [18]	73.4
RS-RF	73.1
LS-RF	74.3
WLS-RF	**75.0**

TABLE 3 – SFEW database

SFEW	% Acc
PHOG-LPQ [6]	19.0
DS-GPLVM [7]	24.7
AUDN [14]	30.1
Semi-Supervised [13]	34.9
RS-RF	35.7
LS-RF	35.6
WLS-RF	**37.1**

TABLE 4 – Conf. matrix (CK+-8em)

	ne	ha	an	sa	fe	di	co	su
ne	92.4	0.3	0.9	0.6	1.2	0.6	3.97	0
ha	0	100	0	0	0	0	0	0
an	4.4	0	91.1	0	0	2.3	2.3	0
sa	22.6	0	0	77.4	0	0	0	0
fe	1.3	4	0	0	90.7	0	0	4
di	3.4	0	0.6	0	0	96.1	0	0
co	11.1	0	0	3.7	0	0	85.2	0
su	1.6	0	0	0	0.4	0	1.2	96.8

TABLE 5 – Conf. matrix (BU-4DFE)

	ne	ha	an	sa	fe	di	su
ne	89.5	0	1.8	4.4	0.9	0.9	2.6
ha	2	89.9	0	0	5	2	1
an	10.1	0	70.7	7.1	2	9.1	1
sa	11	0	15	71	3	0	0
fe	9.8	17.6	2.9	5.9	38.3	11.8	13.7
di	3	4	6.9	1	7.9	73.3	4
su	0	1	0	1	6.2	0	91.8

TABLE 6 – Conf. matrix (SFEW)

	ne	ha	an	sa	fe	di	su
ne	50.2	8.8	9.0	10.0	2.0	16.9	3.1
ha	10.6	67.5	6.2	6.9	2.6	3.5	2.6
an	25.4	16.1	31.3	10.1	3.7	0.9	12.5
sa	21.2	21.2	8.1	22.2	7.1	9.1	11.1
fe	14.2	16.2	13.0	5.0	23.1	7.1	21.3
di	31.3	23.7	10.4	7.1	3.7	15.6	8.2
su	15.4	11.0	12.1	3.3	7.7	6.6	44.0

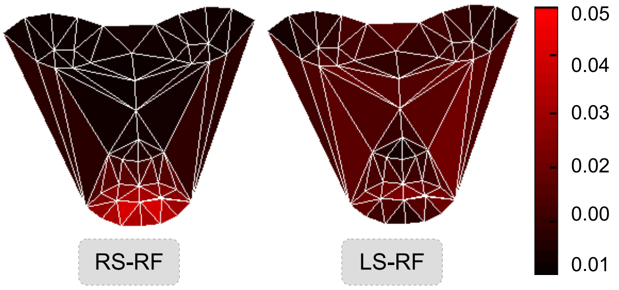

FIGURE 4 – Proportion of top-level (tree root) features per triangle. Best viewed in Color.

is equivalent for LS-RF and WLS-RF on CK+ database, where there is less variability. On the three databases, LS-RF and WLS-RF models provide better results compared to state-of-the-art approaches, even though some of these use complex FFD or spatio-temporal (LBP-TOP) features, or use additional unlabeled data for regularization [13]. Note however that the evaluation protocols are different for some of these approaches. For example, authors in [7] use only the texture information and not the provided landmarks.
Tables 4, 5, 6 show the confusion matrices of WLS-RF on CK+, BU-4DFE and SFEW respectively. Generally speaking, expressions *neutral*, *happy* and *surprise* are mostly correctly recognized, as they involve the most recognizable patterns (smile or eyebrow raise). *Anger* and *disgust* are also accurately recognized on CK+ and BU-4DFE but not so much on SFEW. *Sadness* and *fear* seems to be the most subtle ones, particularly on BU-4DFE and SFEW where it is often misclassified as either *surprise* or *happy*.

4.3 FER on occluded face images

In order to assess the robustness of our system to partial face occlusion, we measured the average accuracy outputted by RS-RF, LS-RF and WLS-RF on CK+ (8 expressions) and BU-4DFE (7 expressions) databases with synthetic occlusions. More precisely, for each image we use the feature points tracked on non-occluded images to highlight the eyes and mouth regions. We then overlay a noisy pattern (see Figure 5), which is a more challenging setup than black boxes used in [26, 10]. We add margins of 20 pixels to the bounding boxes to make sure we cover the whole eyes (with eyebrows, as it represent the most valuable source of information from the eye region) and mouth region. Finally, we align the feature points on the occluded sequences.

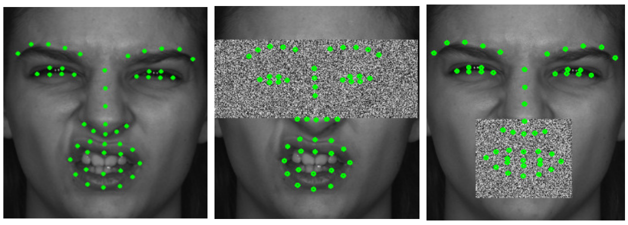

FIGURE 5 – Examples of occluded faces from the BU-4DFE database with aligned feature points.

Figure 6 displays point-wise confidence score outputted by the autoencoder network, averaged over all feature points for groups corresponding to *left* and *right eyes* as well as *left* and *right mouth* part, respectively in the non-occluded, eyes occluded and mouth occluded cases. Overall, the confidence is lower on BU-4DFE, due to more diverse facial morphologies. On the occluded cases, the confidence values outputted for the occluded parts are significantly lower than for the non-occluded ones, indicating that the autoencoder network succeeds in discriminating the occluded patterns. Graphs of Figure 7 show the variation of average accuracy *vs.* hyperparameter R that controls the locality of the trees, respectively under eyes and mouth occlusion on CK+ database. Performances of RS-RF fall heavily when

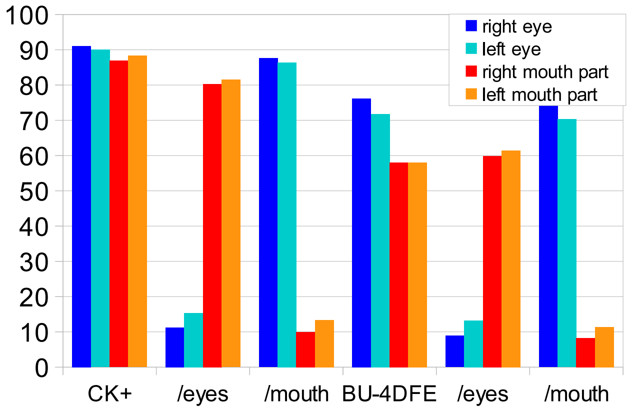

FIGURE 6 – Confidence scores

the mouth is occluded (from 91.5% down to 25.4%), as observed in [26]. This further proves that the global model relies essentially on mouth features to decipher facial expressions. Forcing the trees to be more local (e.g. setting R to 0.1 or 0.2) allows to capture more diverse cues from multiple facial areas, ensuring more robustness to mouth occlusions. It also explains why LS-RF models with $R = 0.8 - 0.5$ can already be quite robust to eyes occlusions, as the majority of the information used on such models likely comes from mouth area. Nevertheless, on those two occlusion scenarios, WLS-RF displays substantially better accuracy than the unweighted local models. Figure 7 also shows the accuracy comparison for both eyes and mouth occlusion scenarios on CK+ and BU-4DFE, with $R = 0.2$. On the two databases, LS-RF is more robust to partial occlusions than RS-RF. Furthermore, WLS-RF also provides better accuracy than both LS-RF and RS-RF. Overall, WLS-RF accuracy under mouth occlusion is 67.1% against 30.3% for [26] in the case where classifiers are trained on non-occluded faces and tested on occluded ones.

5 Conclusion and perspectives

In this paper, we proposed to train Random Forests upon spatially defined local subspaces of the face in order to more efficiently capture the local non-occluded expression patterns, as well as to provide robustness to occlusions. Furthermore, we introduced a hierarchical autoencoder network to model the manifold around specific facial feature points. We showed that the outputted reconstruction error could effectively be used as a confidence measurement to weight the prediction outputted by the local trees. The proposed framework improves state-of-the-art results on several standards FER benchmarks, in addition of significantly adding robustness to partial occlusions.

The ideas introduced in this work lead to interesting directions for future works on face analysis. First, note that the confidence weights outputted by the hierarchical autoencoder network are representative of the spatially defined local manifold of the training data. Thus, these confidence values can be used to guess which parts of the face are the most reliable in a general way (e.g. to address unpredicted illumination patterns or head pose variations), and are not limited to occlusion handling. Furthermore, we could inject these confidence weights into the feature point alignment framework [22] to enhance the robustness of the feature point alignment w.r.t. partial occlusions. Compared to a discriminative approach trained on synthetic examples [8], the proposed manifold learning approach could in theory more efficiently deal with realistic occlusions. Last but not least, we also would like to use the local responses outputted by LS-RF to predict activation of facial Action Units. The idea is that categorical expression labels are easier to collect than FACS coding, thus it can be used to learn high-level representations that are relevant for AU prediction.

Acknowledgements

This work was partially supported by the French National Agency (ANR) in the frame of its Technological Research CONTINT program (JEMImE, project number ANR-13-CORD-0004), and by the Labex SMART (ANR-11-LABX-65) within the Investissements dâĂŹAvenir program, under reference ANR-11-IDEX-0004-02.

Références

[1] L. Breiman. Random forests. *Machine learning*, 45(1) :5–32, 2001. 4, 5

[2] T. Bylander. Estimating generalization error on two-class datasets using out-of-bag estimates. *Machine Learning*, 48(1-3) :287–297, 2002. 5

[3] C. Chen, A. Liaw, and L. Breiman. Using random forest to learn imbalanced data. 2004. 4

[4] S. F. Cotter. Sparse representation for accurate classification of corrupted and occluded facial expressions. In *International Conference on Acoustics, Speech and Signal Processing*, pages 838–841, 2010. 2

[5] A. Dapogny, K. Bailly, and S. Dubuisson. Pairwise conditional random forests for facial expression recognition. In *International Conference on Computer Vision*, 2015. 4

[6] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Static facial expression analysis in tough conditions : Data, evaluation protocol and benchmark. In *International Conference on Computer Vision Workshops*, pages 2106–2112, 2011. 2, 5, 6

[7] S. Eleftheriadis, O. Rudovic, and M. Pantic. Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *Transactions on Image Processing*, 24(1) :189–204, 2015. 6

[8] G. Ghiasi and C. C. Fowlkes. Occlusion coherence : Localizing occluded faces with a hierarchical deformable part model. In *International Conference on Computer Vision and Pattern Recognition*, pages 1899–1906, 2014. 2, 7

[9] M. Hayat, M. Bennamoun, and A. El-Sallam. Evaluation of spatiotemporal detectors and descriptors for facial expression recognition. In *International Conference on Human-System Interaction*, pages 43–47, 2012. 6

[10] X. Huang, G. Zhao, W. Zheng, and M. Pietikäinen. Towards a dynamic expression recognition system under facial occlusion. *Pattern Recognition Letters*, 33(16) :2181–2191, 2012. 2, 6

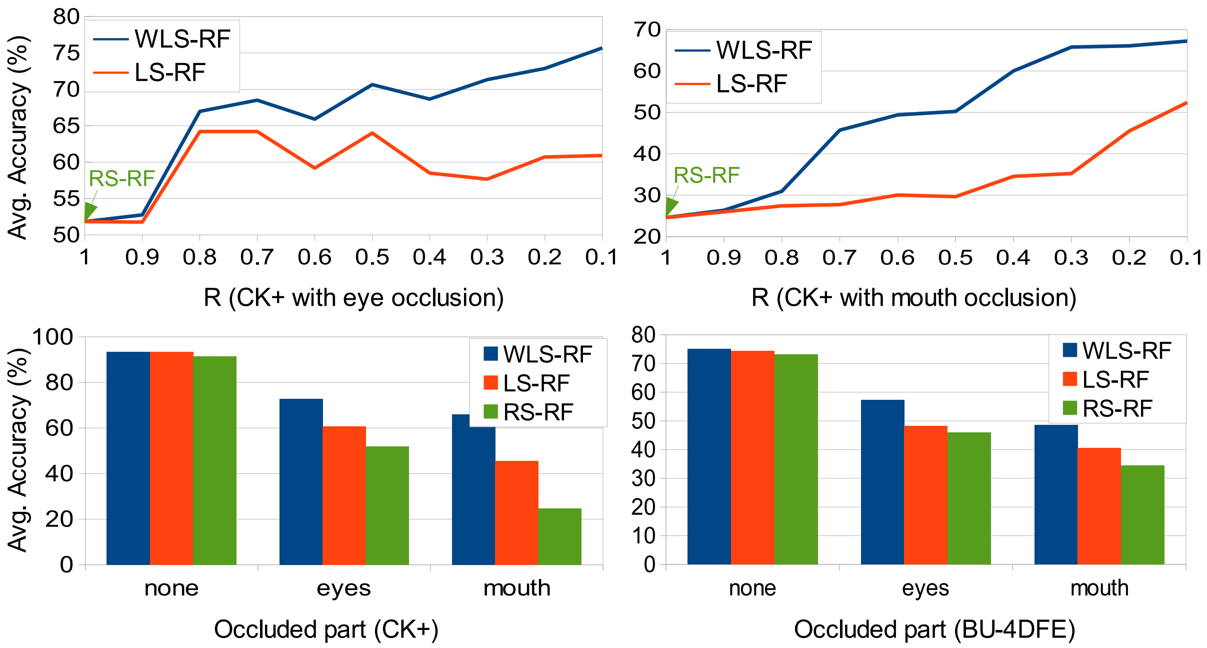

FIGURE 7 – Accuracy outputted on occluded CK+ and BU-4DFE databases

[11] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002. 2

[12] I. Kotsia, I. Buciu, and I. Pitas. An analysis of facial expression recognition under partial facial image occlusion. *Image and Vision Computing*, 26(7) :1052–1067, 2008. 2

[13] M. Liu, S. Li, S. Shan, and X. Chen. Enhancing expression recognition in the wild with unlabeled reference data. In *Asian Conference on Computer Vision*, pages 577–588. 2013. 6

[14] M. Liu, S. Li, S. Shan, and X. Chen. Au-inspired deep networks for facial expression feature learning. *Neurocomputing*, 159 :126–136, 2015. 6

[15] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (CK+) : A complete dataset for action unit and emotion-specified expression. In *International Conference on Computer Vision and Pattern Recognition Workshops*, pages 94–101, 2010. 1, 5

[16] M. A. Ranzato, J. Susskind, V. Mnih, and G. Hinton. On deep generative models with applications to recognition. In *International Conference on Computer Vision and Pattern Recognition*, pages 2857–2864, 2011. 2

[17] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio. Contractive auto-encoders : Explicit invariance during feature extraction. In *International Conference on Machine Learning*, pages 833–840, 2011. 2

[18] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueck. A dynamic approach to the recognition of 3D facial expressions and their temporal models. In *International Conference on Automatic Face and Gesture Recognition*, pages 406–413, 2011. 6

[19] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns : A comprehensive study. *Image and Vision Computing*, 27(6) :803–816, 2009. 1, 6

[20] Y. Sun and L. Yin. Facial expression recognition based on 3D dynamic range model sequences. In *European Conference on Computer Vision*, pages 58–71. 2008. 6

[21] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders : Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11 :3371–3408, 2010. 2

[22] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *International Conference on Computer Vision and Pattern Recognition*, pages 532–539, 2013. 5, 7

[23] L. Xu and P. Mordohai. Automatic facial expression recognition using bags of motion words. In *British Machine Vision Conference*, pages 1–13, 2010. 6

[24] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3D dynamic facial expression database. In *International Conference on Automatic Face and Gesture Recognition*, pages 1–6, 2008. 1, 5

[25] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods : Audio, visual, and spontaneous expressions. *Transactions on Pattern Analysis and Machine Intelligence*, 31(1) :39–58, 2009. 1

[26] L. Zhang, D. Tjondronegoro, and V. Chandran. Random gabor based templates for facial expression recognition in images with facial occlusion. *Neurocomputing*, 145 :451–464, 2014. 2, 6, 7

[27] X. Zhao, T.-K. Kim, and W. Luo. Unified face analysis by iterative multi-output random forests. In *International Conference on Computer Vision and Pattern Recognition*, pages 1765–1772, 2014. 2, 4, 6

[28] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas. Learning active facial patches for expression analysis. In *International Conference on Computer Vision and Pattern Recognition*, pages 2562–2569, 2012. 2, 6