Using Normal/Abnormal Video Sequence Categorization to Efficient Facial Expression Recognition in the Wild

Taoufik Ben Abdallah[1(✉)], Radhouane Guermazi[2(✉)], and Mohamed Hammami[3(✉)]

[1] Faculty of Economics and Management, MIR@CL, University of Sfax, Sfax, Tunisia
taoufik.benabdallah@fsegs.rnu.tn
[2] Saudi Electronic University, Riyadh, Kingdom of Saudi Arabia
r.guermazi@seu.edu.sa
[3] Faculty of Sciences, MIR@CL, University of Sfax, Sfax, Tunisia
mohamed.hammami@fss.rnu.tn

Abstract. The facial expression recognition in real-world conditions, with a large variety of illumination, pose, resolution, and occlusions, is a very challenging task. The majority of the literature approaches, which deal with these challenges, do not take into account the variation of the quality of the different videos. Unlike these approaches, this paper suggests treating the video sequences according to their quality. Using Isolation Forests (IF) algorithm, the video sequences are categorized into two categories: normal videos that visibly express clear illumination and frontal pose of face, and abnormal videos that present poor illumination, different poses of face, occulted face. Two independent facial expression classifiers for the normal and abnormal videos are built using Random Forests (RF) algorithm. The experiments have demonstrated that processing independently normal and abnormal videos can be used to improve the efficiency of the facial expression recognition in the Wild.

Keywords: Facial expression · Isolation Forests · Normal video sequences · Abnormal video sequences · Random Forests · Uncontrolled environment

1 Introduction

Facial expressions are a practical and important means of human communication. They help express the feelings, emotions, attitudes and behavior of humans [21]. Thus, research on facial expressions is a fundamental issue, affecting many areas of science such as psychology, behavioral science, medicine and computer science. One of the increasingly important research fields is the Automatic Facial-Expression Recognition (AFER). It can be useful in a wide range of applications [13] such as intelligent human-computer interfaces, educational software,

J. Blanc-Talon et al. (Eds.): ACIVS 2020, LNCS 12002, pp. 504–516, 2020.
https://doi.org/10.1007/978-3-030-40605-9_43

etc. Considerable attention has been dedicated to AFER from videos. The literature on this field shows a variety of approaches. However, the majority of these approaches consider the facial expression recognition under a lab-controlled environment where the faces are captured in frontal pose with fairly clear illumination. As a matter of fact, the facial expressions are artificial with almost the same degree of intensity.

Recently, Dhall *et al.* have created the Acted Facial Expressions in the Wild (AFEW) dataset [6–8] to promote the transition from lab-controlled to uncontrolled settings [18]. To the authors' best knowledge, AFEW is the largest and the most famous public video dataset proposed to train facial expression recognition models in real-world conditions. In contrast to the lab-controlled environment video datasets like Cohn-Kanade (CK+) [14] and MMI [22], AFEW shows different intensities of expressions due to the high variations of subjects with different ages, gender, and ethnic backgrounds. In addition to the variation of degree of expression, AFEW presents many other challenges like the low resolution of the videos, the face occlusions, the variation of poses, etc. To deal with these challenges, many researchers have proposed various approaches, which can be divided into two main categories: the traditional-based approaches and the deep learning-based approaches.

The traditional-based approaches are based on low-level features and shallow learning as SVM [27] and decision trees [4,24]. The low-level features have intensively studied over the past few decades, and a variety of methods have been proposed based on appearance and/or motion. Several operators have been proposed in the literature to define features that take into consideration not only the spatial information but also the temporal one as well. For example, in [6], the authors handled some AFER's challenges as the variation of illumination, the variation of face poses, the variation of face scales, and the occlusions, using the spatio-temporal extended-variant of LBP, called LBP on Three Orthogonal Planes (LBP-TOP) [34]. Compared to the use of samples captured under a lab-controlled environment, the use of samples in the Wild decreases the performance of AFER. To go further, Huang *et al.* [12] proposed the Spatio-Temporal Local Monogenic Binary Patterns (STLMBP) operator in order to analyze the magnitude, the orientation, and the phase information for facial macro-expression recognition in the Wild. Kaya *et al.* [15] proposed the Local Gabor Binary Patterns on Three Orthogonal Planes (LGBP-TOP) operator for facial macro-expression recognition in the Wild. The authors applied a set of Gabor Filters on the frames of a video sequence in which they used LBP-TOP to detect features. They observed that the use of the LGBP-TOP operator slightly improves the performance of AFER compared to the use of the basic LBP-TOP operator. Guo *et al.* [11] explored other mechanisms so as to describe facial appearance for facial macro-expression recognition in the Wild. They constructed the longitudinal facial expression atlases to obtain salient facial feature changes during an expression process. In [5] and [30], the authors used multiple features to describe facial appearance and motion information. The appearance features were detected through the Histogram of Oriented Gradients from Three

Orthogonal Planes (HOG-TOP) operator. The motion features were derived from the warp transformation of facial points that captures facial configuration changes. Experiments have demonstrated that the second way of combination enhances the discriminative power of features used for AFER in the Wild. The low-level features have been reported to be incapable of addressing the challenges of uncontrolled environment. Indeed, the performance of the majority of these approaches does not exceed 50% [11]. Other modalities such as acoustic (*i.e.,* voice) features have also been used in multimodal systems to improve the facial expression recognition [5], but the enhancement is not important.

The deep learning-based approaches are based on artificial neural networks of multiple layers. Each layer transforms its input data into a slightly more abstract and composite representation [20,31]. These neural networks make it possible to detect high-level features and classify the facial expressions into a unified process. The Convolutional Neutral Network (CNN) and Recurrent Neural Network (RNN) are the two most fundamental network architectures which have shown great superiority in extracting discriminating features and modeling temporal relationship within sequences [18]. Yao *et al.* [32] designed a new CNN architecture, referred to as HoloNet, that reduces redundant filters. It is applied simultaneously to detect high-level features from images and classify facial expressions in the Wild. Sun *et al.* [25] and Li *et al.* [17] proposed the Region-based CNN (R-CNN) to learn features for AFER. Several other CNN-derived architectures have implemented to detect the temporal information from video sequences. In [9,28] and [33], the improved version of the Recurrent Neural Network (RNN), called Long Short Term Memory (LSTM), is applied to link the detected features from each frame to those of the others. Recently, Sun *et al.* [26] have proposed the Multi-channel Deep Spatial-Temporal feature Fusion neural Network (MDSTFN) to exploit the spatial and temporal features detected from an image for AFER. Fan *et al.* [9] used LSTM to model the temporal relation within feature sequences that are extracted by a fine-tuned CNN. The major drawback of the deep learning-based approaches is the need for a large amount of data to avoid the overfitting of the classifier. The related works cited above do not achieve promising results especially for facial expression recognition in the Wild. Contrary to the above-mentioned approaches, some works as [15] and [29] are based on a deep neural architecture to detect features and a shallow learning method to build the classifier.

In our study, through a deep examination of the AFEW dataset, it was observed that the quality of some videos is poor. Thereby, the detection and tracking of faces in these videos is wrong, which influences the efficiency of the model. One possible solution to this problem is to treat this kind of videos separately. Based on the Isolation Forest (IF) method presented in [19], the purpose of this paper is to study if a separate treatment of normal video sequences that visibly express clear illumination and frontal pose of face, and abnormal videos that present poor illumination, different poses of face, occulted face, and poor face detection and tracking, can improve the ability of our low-dimensional feature space PCA[PTLBP^{u2}] [1] to automatically recognize facial expression in real-world conditions.

The remainder of this paper is organized as follows. Section 2 describes the proposed facial expression recognition approach and focuses on the step of normal/abnormal video sequence categorization. Section 3 shows experimental results and discusses the effectiveness of our proposal. Section 4 draws a conclusion and some perspectives.

2 Proposed Approach

We propose to build an approach for an automatic video facial expression recognition. It is based on a low-dimensional temporal feature space in order to produce a classifier, making it possible to recognize facial expression represented by a video sequence that captured under lab-controlled environment or in the Wild. The proposed feature space is called Pyramid of uniform Temporal Local Binary Patterns (PTLBP^{u2}) [1]. Figure 1 shows a flowchart of the proposed approach. It is performed in four main steps.

Step 1: Preprocessing the input video sequences through face detection and tracking.

Step 2: Expressing the video sequences through PCA[PTLBP^{u2}] features [1]. It represents a temporal low-dimensional space through the second and the third levels of a pyramid representation, using the 33 discriminating cuboids selected by applying the method proposed in [1] (8 from level 2 and 25 from level 3).

Step 3: Classifying the video sequences into two categories: normal and abnormal on the basis of the Isolation Forests (IF) algorithm [19]. This classification is particularly valuable for facial expression recognition in the Wild where the quality of video sequences is invariant. Indeed, a good or "normal" video sequence shows a clear illumination and a frontal pose without occlusion of the eyes, nose, and mouth. Contrariwise, a poor or "abnormal" video sequence presents poor illumination, different face poses and scales, and occlusions. Figure 2 shows some examples of abnormal video sequences of cropped faces.

Normal/Abnormal Video Sequence Categorization. To automatically classify video sequences as normal or abnormal, we rely on the Isolation Forests algorithm (IF) [19]. This algorithm separates an instance from other instances. Since the anomalies are few and different, they are more likely to be separated. Unlike other existing anomaly-detection algorithms, the IF takes advantage of the process of scaling up to handle extremely large data size and high-dimensional problems with a large number of irrelevant attributes without defining distances or density measures [19].

Given O the vector of training video sequences, firstly, we calculate the PCA[PTLBP^{u2}] feature matrix $trainV$. Further, we apply IF to build the model that separates normal from abnormal video sequences. This model is formed by nbT binary trees whose leaf-nodes contain only one sample. Each internal-node has exactly two child nodes, $node.left$ and $node.right$.

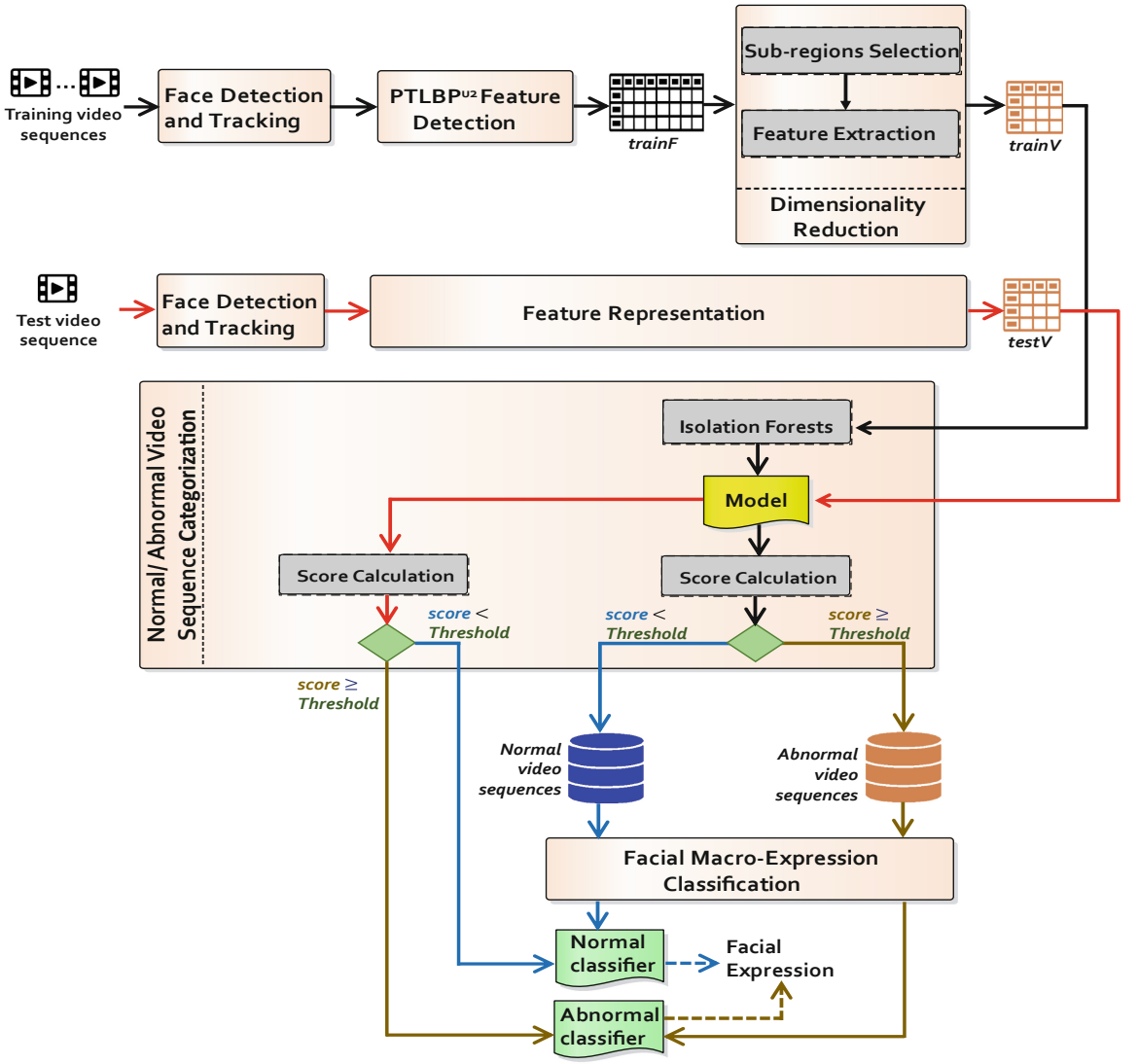

Fig. 1. An overview of the proposed facial expression recognition approach

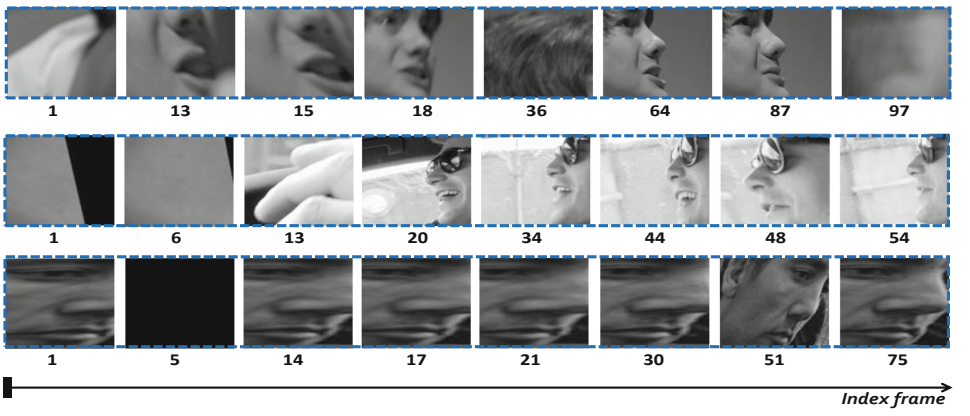

Fig. 2. An example of abnormal video sequences

For each building of a tree t_j ($j \in [1..nbT]$), we proceed as follows:

1. Selecting randomly ψ samples from $trainV$. The result is registered in a ψ-by-nbE matrix, referred to as $subV$.
2. Initializing t_j, by the root node that contains all samples of $subV$.
3. Selecting randomly a feature, noted p, from $subV$ and determine its corresponding index i_p, its minimum value min_p, and its maximum value max_p.
4. Calculating the average value q between min_p and max_p defined as $\frac{(min_p+max_p)}{2}$.
5. $\forall k \in [1..\psi]$, decomposing the current node into sub-nodes: $nodeLeft$ contains samples of $subV$ with $p < q$ (i.e. $subV(k, i_p) < q$); and $nodeRight$ contains samples of $subV$ with $p \geq q$ (i.e. $subV(k, i_p) \geq q$).
   Repeating (1), (2), (3), (4) and (5) for each $nodeLeft$ and $nodeRight$ until each node represent only one instance.

Finally, for a given video sequence, we project it on the obtained model, and we calculate an anomaly score, noted $score$. This score is defined by the Eq. 1:

$$score = 2^{-\frac{avg\_path}{c}} \tag{1}$$

Where $avg\_path = \frac{\sum_{j=1}^{nbT} path^{(j)}}{nbT}$; $path^{(j)}$ is the number of edges of the path between the root node of t_j and the terminating node of the given video sequence; and $c = 2\ \ln(\psi - 1) + 0.58 - (\frac{2(\psi-1)}{\psi})$ is a variable that used to normalize each path length; It is equivalent to the unsuccessful search measure in the Binary Search Tree (BST) [23]. The abnormal video sequence has a shorter average path than a normal video sequence and resides closer to the root of the tree. So, if $score \geq threshold$, the given video sequence is considered as abnormal; otherwise, it is considered as normal [19].

Step 4: Building the facial expression recognition classifier using a supervised learning technique. There is no one optimal algorithm for all situations. It is strongly dependent on the data nature and user expectations. The performance of our approach has been evaluated using three different supervised learning techniques: the Support Vector Machine SMO [16,27], C4.5 decision tree [24], and the Random Forests (RF) [4]. Indeed, SMO represents one of the widely applied techniques for AFER in the literature; C4.5 decision tree offers simple models, in the form of rules, and easily interpretable; and RF uses the bootstrap aggregating strategy in building trees to correct the decision trees' problem of overfitting.

3 Experiments

To validate the proposed approach, we reserve the following subsections for presenting firstly the facial-expression dataset and the different conducted experiments.

3.1 Facial Expression Dataset

Several datasets accessible to the research communities are available in the literature. The majority of these datasets contains images or video sequences labeled by the universal facial expressions captured or collected either under a lab-controlled environment [14,22]. A limited number of datasets present video sequences captured under real-world conditions where the most famous dataset is the Acted Facial Expressions in the Wild (AFEW) [6]. In our work, we have evaluated the proposed approach using the AFEW dataset.

AFEW dataset [6] contains video sequences collected from different movies with various head poses, occlusions, and illuminations [18]. It provides close-to-real-world conditions of varied subjects in terms of sexes and ages. In our experiments, we have used version 7.0 of AFEW. It is divided into three sets. The training set contains 773 samples, the validation set contains 383 samples, and the test set contains 653 samples. The selection of the different subsets is done in an independent manner in terms of subjects and movie/ TV sources. Compared to the version 6 of AFEW [7], only 90 video sequences are added to the AFEW 7.0's test subset, but the training and the validation sets are not modified. The Mixture of Parts (MoP) model [10] is applied to detect and track the faces. However, after face detection and tracking, Only 756 out of 773 training video sequences, and 371 out of 383 validation video sequences are available which are used in our experiments. These video sequences are labeled by seven expressions: 147 (resp. 63) video sequences from the training set (resp. test set) are labeled *"happiness"*, 113 (resp. 59) video sequences *"sadness"*, 80 (resp. 44) video sequences *"fear"*, 74 (resp. 39) video sequences *"disgust"*, 71 (resp. 46) video sequences *"surprise"*, 133 (resp. 59) video sequences *"anger"*, and 138 (resp. 61) video sequences *"neutral"*. As the test video sequences are not labeled, we have not used them to evaluate the performance of our approach.

3.2 Experimental Results

To investigate the performance of the proposed approach, we have carried out all classifiers generated by SMO, C4.5 or RF.

In all these experiments, for the AFEW 7.0 dataset, we have used the video sequences of the AFEW 7.0's training set to build the proposed classifiers and the video sequences of the AFEW 7.0's validation set to estimate the performance. As a validation metric, we have calculated the accuracy rate. In our work, the pyramid representation requires that the length and the width of the frames must be divisible by 2, 4, and 8. Thus, we have defined a size of frames equals 128×128 pixels.

Facial Expression Recognition Without Normal/Abnormal Video Sequence Categorization. We have evaluate the three classifiers generated using the AFEW 7.0 training set. All experiments are conducted on the validation set without applying the normal/abnormal video sequence categorization (Table 1).

Table 1. Experimental results of the three classifiers based on PCA[PTLBP^{u2}] feature space, using the AFEW 7.0 validation set

	Classification algorithm	Number of sub-regions	Number of features	Accuracy (%)
(1)	SMO	33	39	46.09
(2)	C4.5	33	39	44.74
(3)	RF	33	39	47.71

The result of facial expression recognition in the Wild are very low. The best accuracy is obtained by the RF classifier. It does not exceed 47.71%. In fact, the head movements, the pose, the illumination, and the background variation increase the difficulties of the face detection and the facial expression classification as well.

To sum up all the results, we conclude that RF is better than SMO and C4.5 to recognize facial expressions in the Wild.

Impact of Normal/Abnormal Video Sequence Categorization on Facial Expression Recognition in the Wild. In order to verify the impact of using the normal/abnormal video sequence categorization on the performance of facial expression recognition in the Wild, we have carried out several experiments. As described in Subsect. 2, we apply the IF algorithm to generate a model for normal/abnormal video sequence categorization. Three main parameters of IF can influence the effectiveness of the model built: the size of the sub-sample randomly selected for building each tree in the forest ψ, the number of trees nbT, and the anomaly threshold score $threshold$. Based on the empirical studies, through several datasets at different sizes, proposed in [19], we define $\psi = 256$, $nbT = 100$, and $threshold = 0.5$. As a result, 525 (resp. 288) video sequences from the training set (resp. validation set) are classified as normal, and the 231 (resp. 83) remaining video sequences are classified as abnormal. Table 2 shows the number of video sequences, per facial expression, classified as normal and abnormal for both the training and validation sets of the AFEW 7.0 dataset.

When we visualize the result of the normal/abnormal video sequence categorization based on the IF algorithm, we notice that the most abnormal video sequences present poor illumination, different face poses, and occlusions. So, the obtained results show the effectiveness of the proposed IF model.

According to the previous results, RF is more adequate than SMO and C4.5 for facial expression recognition. So, we will use it as the classification technique based on PCA[PTLBP^{u2}] to generate both normal and abnormal classifiers separately. Using the 525 (resp. 231) normal (resp. abnormal) training video sequences, we build the normal (resp. abnormal) classifier. After that, using the validation set, we calculate the accuracy rate of using normal and abnormal classifiers separately and together (normal+abnormal). Obviously, the normal classifier displays better

Table 2. Normal and abnormal video distribution of the training and validation sets of AFEW 7.0 based on the IF model

Facial expressions	Normal video sequences		Abnormal video sequences	
	Training	Validation	Training	Validation
Neutral (NE)	106	55	32	6
Anger (AN)	82	41	51	18
Disgust (DI)	48	31	26	8
Happiness (HA)	103	51	44	12
Fear (FE)	54	30	26	14
Sadness (SA)	84	49	29	10
Surprise (SU)	48	31	23	15
Total	**525**	**288**	**231**	**83**

performance (63.54% of accuracy) than abnormal classifier (49.40%). Therefore, the performance of the "normal+abnormal" classifier reaches 60.38% of accuracy.

Figure 3 shows the performances of predicting each facial expression with and without normal/abnormal video sequence categorization. In fact, we compare the classifier generated by RF based on PCA[PTLBP^{u2}], using all the video sequences of the training set of AFEW 7.0 with the classifier that separates normal from abnormal video sequences ("normal+abnormal" classifier).

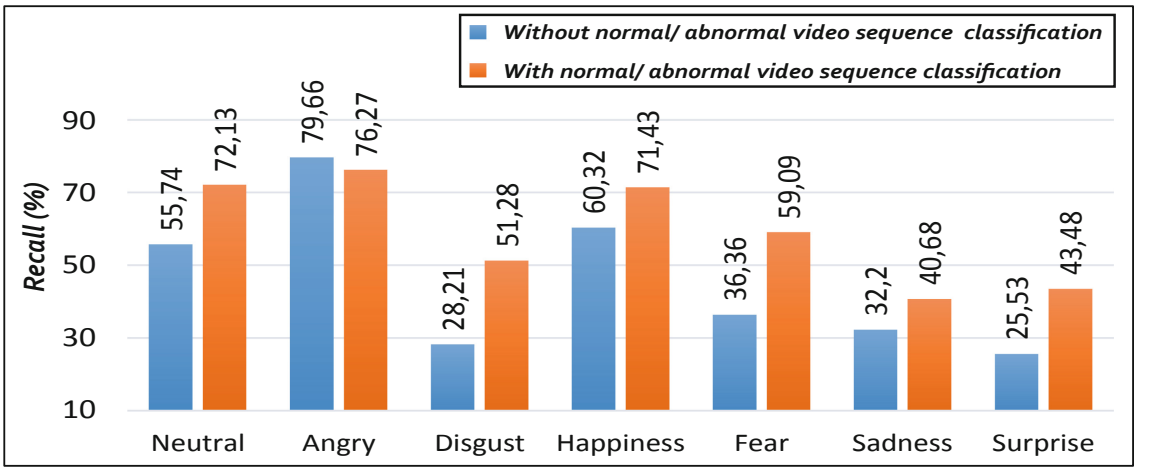

Fig. 3. Impact of the normal/abnormal video sequence categorization on AFER in the Wild using the validation set of the AFEW 7.0 dataset

We observe that using the normal/abnormal video sequence categorization increases the performance of the classifier in the recognition of the majority of facial expressions. The improvement of recall depends on the expression. It varies from 8.48% (the *"sadness"* facial expression) to 23.07% (the *"disgust"* facial expression). The increase in the performance can be assigned to the use

of two different classifiers, each of them adequate to deal with some conditions (normal or abnormal). Moreover, using the normal/abnormal video sequence categorization decreases by 3,39% the recall of only one out of seven expressions, which is the *"angry"* facial expression.

To conclude, the use of normal/abnormal video sequence categorization for facial expression recognition in the Wild can improve the accuracy of the basic classifier generated by RF, using the PCA[PTLBP^{u2}] features to represent each video sequence.

3.3 Analysis and Discussions

Considering that RF is based on a random selection of samples and features, we decide to repeat the experiments based on 10 times in order to evaluate the impact of the random selection on the proposed approach. For each iteration, we calculate the accuracy rate. Then, we measure the standard deviation in order to quantify the magnitude of the dispersion of accuracies from their mean. A low standard deviation indicates that random selection does not affect the effectiveness of the proposed approach. It is worth noting that in all the previous experiments based on RF, we consider the results of the iteration that record an accuracy rate close to the mean. Using the "normal+abnormal" classifier, we find that the mean of accuracies equals 60.24 with 1.35 as a standard deviation. Therefore, we deduce that the effectiveness of the proposed classification algorithm is not affected by the random selection of parameters.

Comparative Study. We have conducted a comparison of our approach with other related approaches in the literature in terms of the number of features, the number of video sequences, and the accuracy rate. Table 3 reports the comparison of our proposal with some recent related works using AFEW (version 6.0 or 7.0).

Table 3. The comparison of the proposed approach vis-a-vis the related approaches on the AFEW dataset

	Feature space	Class. Algo	Nb. Samples	Nb. Features	τ (%)
Dhall *et al.* [7]	LBP-TOP	SVM	383	–	38.80
Yao *et al.* [32]	CNN architecture[a]		383	1024	51.96
Chen *et al.* [5]	Hybrid features	SVM	383	3964	40.20
Fan *et al.* [9]	3DCNN + LSTM architecture[a]		383	–	51.96
Afshar *et al.* [3]	Hybrid features	ELM	378	26588	42.86
Vielzeuf *et al.* [28]	3DCNN + LSTM architecture[a]		380	297	52.20
Kaya *et al.* [15]	Hybrid features	ELM	371	–	52.30
Kaya *et al.* [15]	CNN + Hybrid features	ELM	371	–	57.02
Ours	**PCA[PTLBP^{u2}]**	**RF**	**371**	[24–34]	**60.24**

[a]Unified process (feature detection + classification)

Several of the related works have proposed traditional-based approaches [3,5,7,15]. Recent works have represented the deep learning-based approaches according to different architectures of deep neural networks [9,15,28,32].

According to Table 3, the proposed approach is the best one according to the accuracy that reaches 60.24% for facial macro-expression recognition in the Wild. Comparing the performance of our approach to that of the literature approaches, the enhancement varies from 3.22% [15] to 21.44% [7]. Although the enhancement is not important compared to [15], we provide a lower number of features. Indeed, the proposed approach needs only 39 features to present a video sequence. Also, it is based only on visual features without incorporating audio features unlike some related works such as [15,32]. Probably, the incorporation of additional features can enhance the performance of the proposed approach.

To conclude, our proposal is considered as a strong competitor not only to the traditional-based approaches but also to the deep learning-based approaches in the literature to recognize facial expressions in the Wild.

4 Conclusion and Perspectives

Automatic Facial Expression Recognition in the Wild is a very challenging problem due to different expressions under arbitrary poses, variability of illumination and occlusions. Different from existing approaches, in this paper, we formulate the AFER by separating the analysis of videos categorized as Normal from videos categorized as Abnormal.

To address this challenge, we have proposed a normal/abnormal video sequence categorization approach based on Isolation Forest (IF) algorithm. Our proposal is based on a low-dimensional feature space called PCA[PTLBP^{u2}]. Furthermore, we compare the SMO, C4.5 and RF classification techniques to create robust AFER classifiers.

The findings of our research are quite convincing, and thus the following conclusions can be drawn:

- Separating normal from abnormal video sequences can improve the facial expression recognition in the Wild. With this categorization, we record the best accuracy rate compared to the most famous literature studies. Thus, we proved that building separate classifiers improves significantly the results.
- Using PCA[PTLBP^{u2}] feature space shows its effectiveness to represent the video sequence for both normal and abnormal video sequences. Despite the fact that the proposed feature space is a low-dimensional space, it represents effectively the video sequences.
- Observing that the traditional-based approaches can be under certain conditions more efficient than the deep-learning-based approaches compared to some state-of-the-art works.
- Using an imbalanced distribution of facial has an important effect on the performance of the classifier to detect a particular expression.

In our future research, we intend to concentrate on studying the impact of the different parameters of the IF algorithm on the performance of the proposed approach. We also plan to test other alternatives to classify normal and abnormal videos like the unsupervised Random Forests [2]. We should also estimate the

impact of the normal/abnormal video categorization on other handcrafted features like STLMBP [12] and HOG-TOP [5]. As the distribution of the different expression on AFEW 7.0 is imbalanced, we can also study the impact of applying appropriate learning algorithms to improve the facial expression recognition rate.

References

1. Abdallah, T.B., Guermazi, R., Hammami, M.: Facial-expression recognition based on a low-dimensional temporal feature space. Multimedia Tools Appl. **77**(15), 19455–19479 (2018)
2. Afanador, N.L., Smolinska, A., Tran, T.N., Blanchet, L.: Unsupervised random forests: a tutorial with case studies. Chemometrics **30**(5), 232–241 (2016)
3. Afshar, S., Salah, A.A.: Facial expression recognition in the wild using improved dense trajectories and fisher vector encoding. In: International Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, pp. 1517–1525. IEEE (2016)
4. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
5. Chen, J., Chen, Z., Chi, Z., Fu, H.: Facial expression recognition in video with multiple feature fusion. IEEE Trans. Affect. Comput. **9**(1), 1–12 (2016)
6. Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., Gedeon, T.: From individual to group-level emotion recognition. In: International Conference on Multimodal Interaction, pp. 524–528. ACM, New York (2017)
7. Dhall, A., Goecke, R., Joshi, J., Hoey, J., Gedeon, T.: Video and group-level emotion recognition challenges. In: International Conference on Multimodal Interaction, pp. 427–432. ACM, New York (2016)
8. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Collecting large, richly annotated facial-expression databases from movies. IEEE MultiMedia **19**(3), 34–41 (2012)
9. Fan, Y., Lu, X., Li, D., Liu, Y.: Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In: International Conference on Multimodal Interaction, pp. 445–450. ACM, New York (2016)
10. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. Comput. Vis. **61**(1), 55–79 (2005)
11. Guo, Y., Zhao, G., Pietikäinen, M.: Dynamic facial expression recognition with Atlas construction and sparse representation. IEEE Trans. Image Process. **25**(5), 1977–1992 (2016)
12. Huang, X., He, Q., Hong, X., Zhao, G., Pietikainen, M.: Improved spatiotemporal local monogenic binary patterns for emotion recognition in the wild. In: International Conference on Multimodal Interaction, Istanbul, Turkey, pp. 514–520. ACM (2014)
13. Imotions: Facial Expression Analysis: The Complete Pocket Guide (2016). https://imotions.com/blog/Facial-Expression-Analysis
14. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: International Conference on Automatic Face and Gesture Recognition, Grenoble, France, pp. 46–53. IEEE (2000)
15. Kaya, H., Gürpınar, F., Salah, A.A.: Video-based emotion recognition in the wild using deep transfer learning and score fusion. Image Vis. Comput. **65**, 66–75 (2017)
16. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.: Improvements to Platt's SMO algorithm for SVM classifier design. Neural Comput. **13**(3), 637–649 (2001)

17. Li, J., et al.: Facial expression recognition with faster R-CNN. Procedia Comput. Sci. **107**, 135–140 (2017)
18. Li, S., Deng, W.: Deep facial expression recognition: a survey. Computer Vision and Pattern Recognition (2018, to appear)
19. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: International Conference on Data Mining, pp. 413–422. IEEE, Washington (2008)
20. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E.: A survey of deep neural network architectures and their applications. Neurocomputing **234**, 11–26 (2017)
21. Mehrabian, A.: Communication without words. Psychol. Today **2**(4), 53–56 (1968)
22. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: International Conference on Multimedia, Amsterdam, Netherlands. IEEE (2005)
23. Preiss, B.R.: Data Structures and Algorithms with Object-Oriented Design Patterns in Java. Wiley, Hoboken, first edn (1999)
24. Quinlan, J.R.: C4.5: Programs for Machine Learning, 1st edn. Morgan Kaufmann, San Francisco (1993)
25. Sun, B., Li, L., Zhou, G., He, J.: Facial expression recognition in the wild based on multimodal texture features. Electron. Imaging **25**(6), 1–8 (2016)
26. Sun, N., Li, Q., Huan, R., Liu, J., Han, G.: Deep spatial-temporal feature fusion for facial expression recognition in static images. Pattern Recogn. Lett. **119**, 49–61 (2019)
27. Vapnik, V.: The Nature of Statistical Learning Theory, 1st edn. Springer, New York (1995). https://doi.org/10.1007/978-1-4757-3264-1
28. Vielzeuf, V., Pateux, S., Jurie, F.: Temporal multimodal fusion for video emotion classification in the wild. In: International Conference on Multimodal Interaction, pp. 569–576. ACM, New York (2017)
29. Wang, F., Lv, J., Ying, G., Chen, S., Zhang, C.: Facial expression recognition from image based on hybrid features understanding. Vis. Commun. Image Represent. **59**, 84–88 (2019)
30. Yan, H.: Collaborative discriminative multi-metric learning for facial expression recognition in video. Pattern Recogn. **75**, 33–40 (2018)
31. Yan, J., Zheng, W., Cui, Z., Tang, C., Zhang, T., Zong, Y.: Multi-cue fusion for emotion recognition in the wild. Neurocomputing **309**, 27–35 (2018)
32. Yao, A., Cai, D., Hu, P., Wang, S., Sha, L., Chen, Y.: HoloNet: towards robust emotion recognition in the wild. In: International Conference on Multimodal Interaction, pp. 472–478. ACM, New York (2016)
33. Yu, Z., Liu, G., Liu, Q., Deng, J.: Spatio-temporal convolutional features with nested LSTM for facial expression recognition. Neurocomputing **317**, 50–57 (2018)
34. Zhao, G., Pietikäinen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 915–928 (2007)