Signal, Image and Video Processing (2021) 15:1049–1056
https://doi.org/10.1007/s11760-020-01830-0

ORIGINAL PAPER

Video-based person-dependent and person-independent facial emotion recognition

Noushin Hajarolasvadi[1] · Enver Bashirov[2] · Hasan Demirel[1]

Received: 30 April 2020 / Revised: 14 October 2020 / Accepted: 30 November 2020 / Published online: 19 January 2021

Abstract

Facial emotion recognition is a challenging problem that has attracted the attention of researchers in the last decade. In this paper, we present a system for facial emotion recognition in video sequences. Then, we evaluate the system for a person-dependent and person-independent cases. Depending on the purpose of the designed system, the importance of training a personalized model versus a non-personalized one differs. In this paper, first, we compute 60 geometric features for video frames of two datasets, namely RML and SAVEE databases. In the next step, k-means clustering is applied to the geometric features to select k most discriminant frames for each video clip. Then, we employ various classifiers like linear support vector machine (SVM) and Gaussian SVM to find the best representative k. Finally, five pre-trained convolutional neural networks, namely VGG-16, VGG-19, ResNet-50, AlexNet, and GoogleNet, were used evaluating two scenarios: person-dependent and person-independent emotion recognition. Additionally, the effect of geometric features in keyframe selection for a person-dependent and person-independent scenarios is studied based on different regions of the face. Also, the extracted features by CNNs are visualized using the t-distributed stochastic neighbor embedding algorithm to study the discriminative ability in these scenarios. Experiments show that person-dependent systems result in higher accuracy and suitable to be used in personalized systems.

Keywords Facial emotion recognition · Person specification · Geometric features · Keyframe selection

1 Introduction

With the rapid expansion of multimedia content, facial emotion recognition (FER) from video clips became one of the challenging research areas in recent years. Automatic detection of emotional state through face and voice is proven to be useful across many human–computer interfaces (HCI) [1–3] including pain detection, lie/fraud detection, verification systems and multimedia tagging systems. Meanwhile, one important aspect of studying emotional responses is related to person-dependent and person-independent scenarios.

It is important to highlight that, depending on the purpose of the designed HCI system, the significance of having a personalized model versus a non-personalized one differs. For example, the source of differences rising from gender distribution and ethnicity can result in dramatic errors in a computer-aided diagnosing (CAD) system designed for pain detection [4]. User identification using face verification is another example which is proved to be more accurate when personal biometric features like iris are taken into account [5,6].

On the other hand, applications such as mobile computing and video tagging systems [7,8] benefit from models based on general information of an average user like the distance between the eyes, gaze direction, and head position. Although several research works are carried out to design personal and non-personal FER models, none considered the comparison of these systems based on a priori observation of the person.

In this study, we used two databases that are widely used in the literature for FER: Ryerson Multimedia Laboratory (RML) database [9] and Surrey Audio-Visual Expressed Emotion (SAVEE) database [10]. Firstly, we extract frames and the landmark points using OpenFace framework [11]. Then, we calculate 50 distances and 10 angles from seven

✉ Noushin Hajarolasvadi
noushin.hajarolasvadi@cc.emu.edu.tr

1 Electrical and Electronic Engineering Department, Eastern Mediterranean University, via Mersin 10, Famagusta, Turkey

2 Department of Mathematics, Eastern Mediterranean University, via Mersin 10, Famagusta, Turkey

different face regions to generate a 60-dimensional [8] geometric feature vector for each frame. In the next step, k-means clustering is applied to these vectors to select k most discriminant frames of each video. At the end of this step, each video is represented by k frames called keyframes. Finally, we employ linear support vector machine (SVM) and Gaussian SVM to find the best representative k. Then, pre-trained convolutional neural networks (CNNs) like VGG-16, VGG-19[12], ResNet50 [13], AlexNet [14], and GoogleNet [15] were used to evaluate two scenarios: person-dependent and person-independent emotion recognition. Additionally, we study the effect of k in video processing, the significance of geometric features based on different regions of the face, and the discrimination ability within these scenarios.

It is important to point out that this study is different from verification systems in that we study the effect of prior observation of the subject in FER systems rather than solely identification of the subject. A previous observation of the person expressing any of the six basic emotions helps to improve classification for the other five basic emotions of that person.

We defined the following scenarios to study the effect of a priori observation of the person. In each scenario, it is assumed that s is the number of subjects. Each subject has v videos across all emotional categories, and each video has N frames. In the keyframe selection phase, k keyframes out of N are selected. Having this in mind and based on the given definitions, the following two scenarios are considered in this study:

1. In the first scenario, the classifier is trained based on all videos of some of the subjects. In the test phase, the classifier predicts an emotion label for videos of the subjects who did not participate in the training. In fact, keyframes of each video are classified by SVM and an emotion label is assigned to the video by score averaging. This scenario presents a model for person-independent classification.
2. In the second scenario, the classifier is trained using some of the videos of all subjects. In the test phase, the classifier predicts an emotion label for those videos which are not used during training. Again, keyframes of one video are classified by SVM and an emotion label is assigned to the video by score averaging. This scenario presents a model for person-dependent classification.

It is important to highlight the advantages of such definition. In both scenarios, instead of testing based on a single unseen subject, an ensemble of unseen subjects is used. This makes our system more challenging comparing to the scenarios where leave-one-subject-out (LOSO) is used. Additionally, we try to set a general definition for these scenarios through conducting a series of comparative studies because the definition of person-dependent and person-independent cases is different from one research work to another. It should be noted that the emotion recognition problem is different from user identification. As a result, it requires problem-related strategies for training and test. In our proposed strategy, the person-dependent scenario focuses on recognition of emotional expressions by forcing the classifier to learn based on some emotional status of the subject but performing the test based on unseen emotional expression.

The main contribution of this paper are three folds: 1) it shows that prior observation of the subject's identity is beneficial for emotion detection in applications like CAD or security systems where accuracy is of significant importance, while in non-personalized applications like emotion detection in crowd, one must benefit from general facial features rather than identity of the subjects, 2) it studies various regions of the face based on geometric features for improved performance, and 3) the effect of increasing number of keyframes on emotion recognition rate is investigated.

This paper is organized as follows: Sect. 2 reviews the related works. In Sect. 3 details of the proposed method for comparison of personal and non-personal models are described. Experimental results followed by conclusion are given in Sects. 4 and 5, respectively.

2 Related works

Over the last decades, FER has been an active research topic due to potential applications in areas such as CAD and security systems.

Zhang et al. [16] used the fusion of audio-visual features obtained from convolutional neural networks (CNN) and 3D-CNN as an input to a deep belief networks (DBNs). They achieved a person-independent recognition accuracy of 53.03% and 68.09% for training on the RML database. They adopt the person-independent LOSO strategy with cross-validation for experiments. The proposed person-independent scenario differs from the LOSO strategy in that more than one subject used for testing because the identification is not the target label. In the case of person-dependent scenario, the proposed system does not identify the emotions of a single subject. Such a definition raises classification bias by easily fitting over the information of that specific user.

Noroozi et al. [8] proposed a multimodal ER system where visual geometric features are fused with acoustic features and classified using SVM, random forest, and CNNs. Their work can be categorized under person-independent models. The recognition rate for RML and SAVEE database is reported as 31.67% and 36.10% using the SVM classifier. A mixture of rule-based and machine learning techniques is used by [17]. In essence, the extracted visual and acoustic features are passed into a novel neural classifier called optimized kernel-Laplacian radial basis function. The accuracy of the RML

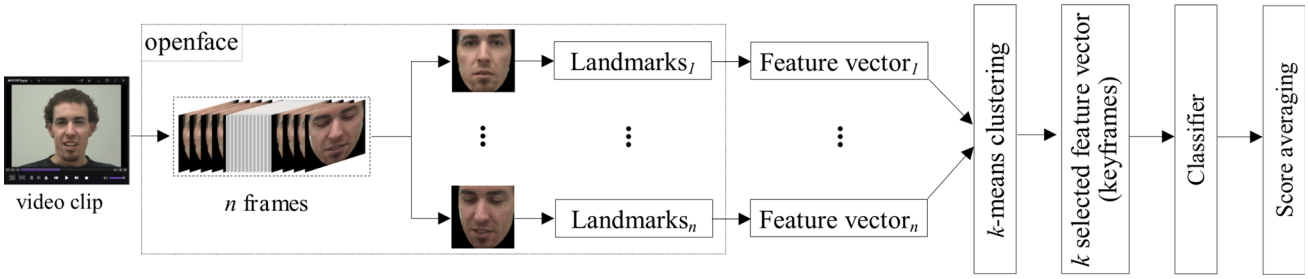


Fig. 1 General framework of the proposed method

dataset using the designed system is about 84%. This work can be categorized as a person-dependent model since all users engaged in training and testing. Different from [8], instead of using a fixed number of keyframes, we studied the effect of k-means algorithm for each database by employing the accuracy of validation set to select an optimum k.

In another study [18], authors present a framework where hidden Markov model (HMM) is applied to active appearance features to perform FER. They achieved 89.04% accuracy on the RML dataset. This study used a person-dependent scenario. In a different study, Wang et al. [19] proposed a FER system for video sequences by using a feature selection method based on Mahalanobis distance. The best recognition rate through the visual channel is reported as 49.29% for studying a person-independent case and 89.20% for studying a person-dependent case on the RML dataset. They used Fisher's linear discriminant analysis (FLDA) classifier.

3 Proposed method

This section explains the main components of the proposed framework as well as the two scenarios mentioned above. In addition to the main objective which is studying the person-dependent and person-independent methods, the effect of the value of k in the k-means clustering algorithm is also analyzed. Figure 1 shows the general framework of the proposed method.

3.1 Face detection and landmark extraction

As previously mentioned, RML and SAVEE databases are used in this study. Initially, the frames from each video clip are extracted using OpenFace software [11]. In a video classification problem, often one deals with the processing of several hundred to thousands of frames. However, most of the frames consist of identical facial expressions. In this context, summarizing each video by a set of discriminant frames, say keyframe, is helpful by providing a reduction in the computational time and complexity of the problem.

Different methods are used for frame selection in the literature. [17] used five images of each subject chosen randomly for training. Random selection does not guarantee the selection of the most informative frames. Other researchers [20] often assume that the peak of emotion is in the middle of the video sequence which is not true all the time.

Generally speaking, there are four approaches for keyframe selection: (a) motion analysis-based approach, where the optical flow for each frame is calculated to detect any changes in facial expression [21], (b) shot boundary-based approach, where the first, the middle and the last frames of each video are considered as the keyframes [20], (c) visual content-based approach, in which the first frame of video is considered as a keyframe and by using a color histogram, the similarities between the current frame and other frames are computed [22], and finally (d) clustering-based approaches, which clusters frames with similar posture and those frames which are closest to the centroid of each cluster are chosen as selected keyframes.

Each of these approaches has its own advantages and disadvantages which are available in more detail in [8]. In our study, the k-means clustering algorithm has been used. The clustering-based approaches are highly sensitive to the motion and noise coming off of the face. In order to reduce this effect, the k-means algorithm is applied to the set of robust features obtained from the well-known tracked facial landmarks. In fact, 68 facial landmark coordinates as shown in Fig. 2 are detected for each frame. The facial landmark vector is shown by $L_n = [l_1, l_2, \ldots, l_{68}]$, where l_i consists of (l_{i_x}, l_{i_y}) as the ith landmark point in the nth frame of a video file.

3.2 Feature extraction

The proposed framework proceeds with the computation of geometric feature vector of each frame using the corresponding facial landmark vector L_j. Two types of geometric descriptors, namely distances and angles, are calculated. In fact, 50 Euclidean distances and 10 angles are computed from the landmarks of each frame. The feature selection theme is adopted from [8].

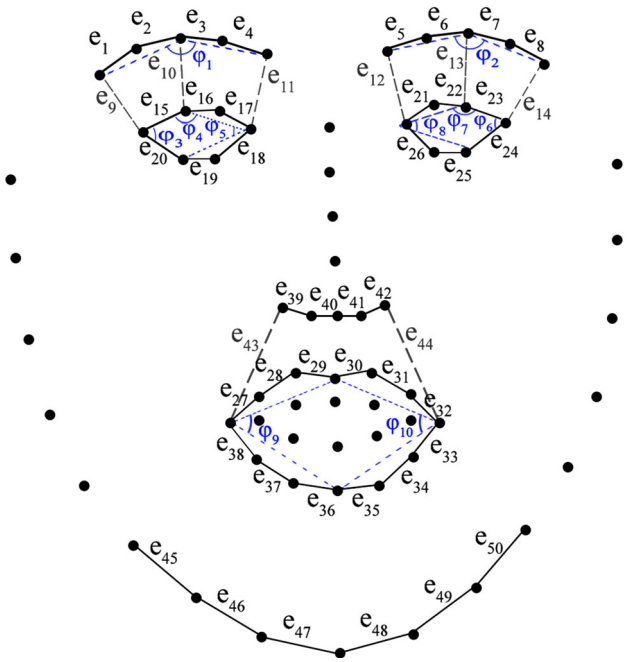


Fig. 2 e_i and ϕ_i are distances and angles, respectively

If the Euclidean distance between two landmarks is $e(l_i, l_j)$, then the calculation is as follows:

$$e(l_i, l_j) = \sqrt{(l_{j_x} - l_{i_x})^2 + (l_{j_y} - l_{i_y})^2} \tag{1}$$

where $j = i + 1$.

In order to have a set of robust features, normalization is applied. Normalizing is carried out by dividing $e(l_i, l_j)$ to the length of corresponding region. In other words, if the distance $e(l_i, l_j)$ belongs to the left or right eyebrow regions, then it is divided by the total length of distances from e_1 to e_4 and from e_5 to e_8, respectively. For the left and right eye regions, length of e_{15} to e_{20} and e_{21} to e_{26} is used as normalizing scale. Similarly, total length of e_{39} to e_{42} is used in calculation of the distances in the nose region, e_{27} to e_{38} for the mouth and e_{45} to e_{50} for the chin. The mathematical calculation is shown in Eq. (2).

$$\hat{e}(l_i, l_{i+1}) = \frac{e(l_i, l_{i+1})}{\sum_j e(l_j, l_{j+1})} \tag{2}$$

This calculation resulted in 50 distance features. In addition, the angles between two distances which share a common landmark are calculated. The calculation of angle descriptors between each triple set of landmark points (l_i, l_r, l_j) is given below:

$$\phi_r = \arccos \frac{e(l_i, l_r)^2 + e(l_i, l_j)^2 - e(l_r, l_j)^2}{2e(l_i, l_r)e(l_i, l_j)} \tag{3}$$

Table 1 List of calculated geometric features

Region	Distances	Angles
Left eyebrow	e_1,e_2,e_3,e_4	ϕ_1
Right eyebrow	e_5,e_6,e_7,e_8	ϕ_2
Left eyebrow to eye	e_9,e_{10},e_{11}	–
Right eyebrow to eye	e_{12},e_{13},e_{14}	–
Left eye	e_{15},e_{16},e_{17},e_{18},e_{19},e_{20}	ϕ_3,ϕ_4,ϕ_5
Right eye	e_{21},e_{22},e_{23},e_{24},e_{25},e_{26}	ϕ_6,ϕ_7,ϕ_8
Upper mouth	e_{27},e_{28},e_{29},e_{30},e_{31},e_{32}	ϕ_9
Lower Mouth	e_{33},e_{34},e_{35},e_{36},e_{37},e_{38}	ϕ_{10}
Nose	e_{39},e_{40},e_{41},e_{42},e_{43},e_{44}	–
Chin	e_{45},e_{46},e_{47},e_{48},e_{49},e_{50}	–

where ϕ_r is the rth angle. The complete list of all descriptors is marked in Fig. 2 and illustrated in Table 1. The calculated distances and angles of nth frame are concatenated to form a feature vector f_n.

3.3 *k*-means clustering

The next step in the framework is applying the k-means clustering algorithm to the extracted feature vectors. The main objective of this algorithm is to group the data into k clusters. The algorithm iteratively assigns each feature vector to one of the k clusters. This results in clustering of the frames based on feature similarities. All the frames (feature vectors) within one cluster are represented by the centroid of that cluster, and the closest frame to the centroid is taken as the representing keyframe of that cluster.

In fact, having k clusters, let μ_{c_j} show the centroid of jth cluster. Then, the closest frame to the jth cluster (c_j) can be found as: $f'_n := \arg\min_n \| f_n - \mu_{c_j} \|$ where f_n is the feature vector of nth frame in a single video. By applying this algorithm, k keyframe is selected.

In order to show the effectiveness of k-means algorithm as a frame selection method, clusters of one video with 336 frames are visualized in Fig. 3. For the visualization purpose, we applied t test score on the 1×60 feature vectors of selected frames and non-selected frames of one video to find the two best representative features. These features were third and fourth angles of left eye. Also, a sequence of selected frames and non-selected ones for one video is illustrated in Fig. 4.

In addition to person-dependent and person-independent experiments, the effect of the value k in k-means clustering is studied. The initial starting point is chosen to be $k = 1$. This generates a selected subset of original data in which each video is represented by only one keyframe. Next, k is increased to 3 which generates another subset of the original data, including three keyframes for each video. The procedure is continued in an odd-heuristic manner by increasing

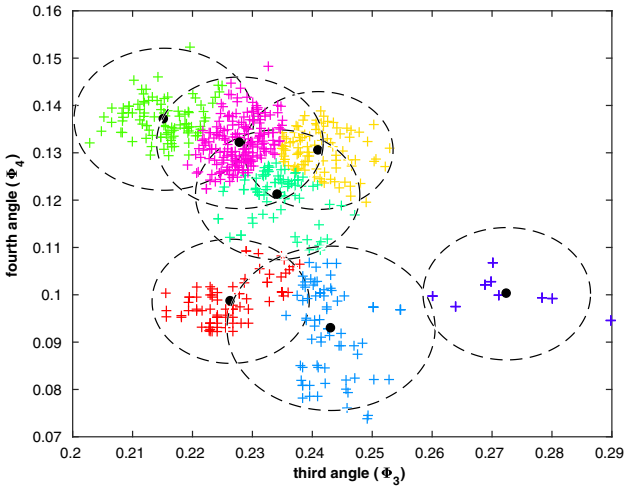


Fig. 3 Generated clusters for k = 7, each cluster is shown by a different color, and centroids are shown as a black dot

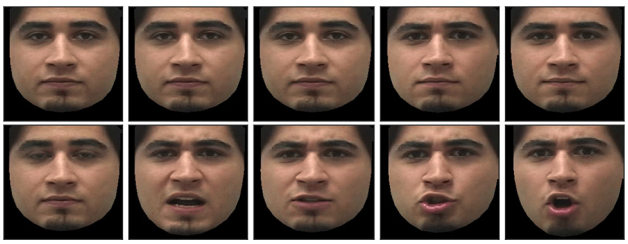


Fig. 4 top row: non-selected frames, bottom row: selected frames, RML database

k to 5, 7, and up to 43. In the end, the keyframe selection procedure resulted in 22 different data subsets for each of the original datasets, RML, and SAVEE. For the sake of simplicity, each of these subsets is shown as $subset_k$. The aim of this observation is to evaluate and optimize the performance of the classifier in terms of accuracy, error rate, and computational time as the number of keyframes increases during training. Choosing k to be odd prevents problems such as having an equal number of votes for different categories while using methods such as majority voting.

Furthermore, we provide the classification results for linear SVM, Gaussian SVM, and the well-known pre-trained CNNs. Considering the high-computational time of CNNs, we do not study the effect of k using these classifiers. In fact, Gaussian SVM had a better performance than linear SVM so it is used as the reference for the best representative k. The corresponding $subset_k$ that maximizes the accuracy on validation data is used as the input for training the pre-trained CNNs, namely VGG-16, VGG-19 [12], ResNet50 [13], AlexNet [14], and GoogleNet [15].

3.4 Experimental framework

The idea behind both models is whether or not the classifier observes any emotional facial information related to a specific subject considering train and test phases. As a result, data division plays an important role in each scenario which is explained in the preceding sections. It is important to note that the size of training data is always larger in person-dependent scenarios. It is known that larger training data help the classifier to achieve a better performance.

3.4.1 Person-independent models

In the first scenario, the classifier is trained based on all videos of some of the subjects. Then, in the test phase, the classifier predicts an emotion label for videos of the subjects who did not participate in training. In fact, each video is represented by k keyframes, and those keyframes are classified by linear and Gaussian SVM. An emotion label is assigned to the video based on a score averaging system. Score averaging calculates a category-based average of frame-level predicted scores and assigns the category with the highest average as the emotion label of video. This removes the uncertainty of dealing with the same number of votes for more than one category. This scenario generates a model for a person-independent classification.

3.4.2 Person-dependent models

In the second scenario, the classifier is trained using some of the videos of all subjects. In the testing step, the classifier predicts an emotion label for those videos that are not used during training. Again, keyframes of each video are classified by SVM and then an emotion label is assigned to the video based on score averaging. This scenario presents a model for person-dependent classification by score averaging.

4 Experimental results

4.1 Datasets

Two datasets have been used to conduct the experiments. The RML database represented by Ryerson Multimedia Laboratory [9] includes 120 videos in each of six basic categories, namely angry, disgust, fear, happiness, sadness, and surprise portrayed by eight subjects. As a result, a dataset of size 720 emotion video samples is obtained. The wide range of ethnicity in this dataset makes it suitable for comparisons between personalized and non-personalized models. The second dataset is SAVEE [10] with 4 male subjects performing the same six basic emotions as the RML dataset. In total, 360 video samples are available in this dataset. The data process-

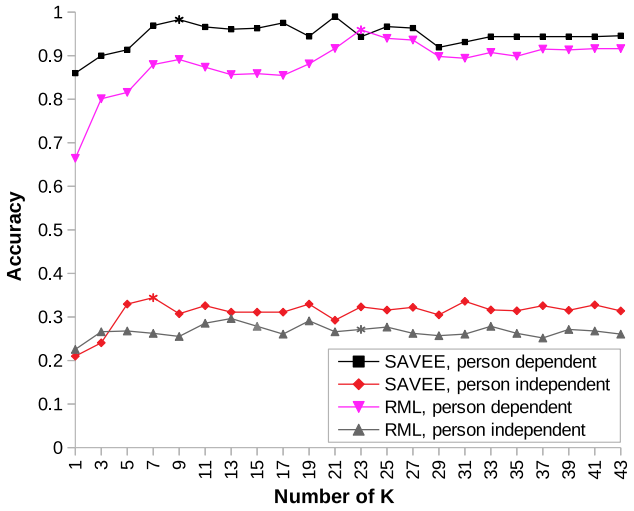


Fig. 5 Accuracy improvement on validation set by increasing # keyframes

ing pipeline explained in Sect. 3.1 is applied to each video sample of both datasets.

4.2 Results and discussion

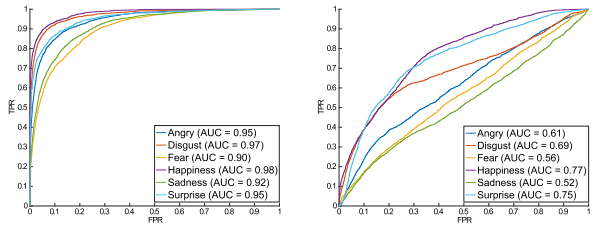
For selecting the best representative k, linear SVM and SVM with Gaussian kernel are used as the classifier. Gaussian kernel has the following format: $K(X_1, X_2) = \exp(\gamma||X1 - X2||^2)$ where $||.||$ is the Euclidean distance function and γ is the spread of the kernel set be $\frac{1}{2\sigma^2} > 0$. The former had a lower performance in both scenarios. Thus, CNN-based experiments are performed using the best representative k of the Gaussian SVM. The kernel scale factor is selected using a heuristic procedure, and standardization is performed. We used cross-validation approach in both scenarios. However, based on the number of subjects within each database and the number of videos recorded for each subject, the percentage of train, test and validation folds changes for person-independent case.

The accuracy of Gaussian SVM for RML and SAVEE dataset on the validation set is shown in Fig. 5. The validation accuracy for each $subset_k$ is assessed, and the subset with maximum value is selected as the representative k for the test set. Also, this subset is used to train the CNNs. The results for the test data are compared with other methods in Tables 2. We report both the test accuracy in percentage and computational time in seconds for classifying test samples using each classifier. This computational time is known as the inference time or the online time which is the time costs during the evaluation of the trained model on an unseen (test) data. ResNet50 had the highest accuracy.

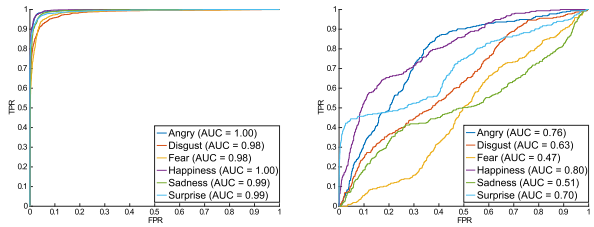
Table 2 Comparing the FER classification performance, **RML** and **SAVEE** databases

Mode	RML					SAVEE				
	# F	Method	Classifier	Acc	Time	# F	Method	Classifier	Acc	Time
Person independent	13	k-means	Linear SVM	30.18	< 1	3	k-means	Linear SVM	29.26	< 1
	19	k-means	Gaussian SVM	36.06	≈ 1	7	k-means	Gaussian SVM	34.44	≈ 1
			VGG-16	49.71	10			VGG-16	42.28	3
			VGG-19	47.32	12			VGG-19	40.32	4
			ResNet50	**51.73**	8			ResNet50	**48.04**	3
			AlexNet	44.13	14			AlexNet	41.78	6
			GoogleNet	48.14	14			GoogleNet	46.66	8
	4	k-means	SVM [8]	31.67	< 1	4	k-means	SVM [8]	36.10	<1
	7	KNN	FLDA [19]	49.29	5	–	–	–	–	–
	19	k-means	Linear SVM	89.97	< 1	5	k-means	Linear SVM	91.36	< 1
Person dependent	23	k-means	Gaussian SVM	95.83	≈ 1	9	k-means	Gaussian SVM	98.77	≈ 1
			VGG-16	94.45	13			VGG-16	96.34	9
			VGG-19	95.60	15			VGG-19	96.56	9
			ResNet50	**98.23**	10			ResNet50	**99.89**	6
			AlexNet	94.86	16			AlexNet	96.21	10
			GoogleNet	97.18	15			GoogleNet	97.02	10
	5	Random	OKL [17]	84.00	NA	All	–	SVM [23]	88.00	< 1
	–	AAM	HMM [18]	89.04	4	All	–	SOM [24]	98.82	NA
	7	KNN	FLDA [19]	89.20	6	All	‡	AlexNet [25]	94.33	9

Bold results are the best ones in our experiments
#F shows number of frames used to represent a video; ‡Pixel-wise difference of consecutive frames; time is reported in seconds



(a) left: Person Dependent, right: Person Independent, RML



(b) left: Person Dependent, right: Person Independent, SAVEE

Fig. 6 ROC for test data using best representative k

Training is performed by transfer learning over the classification layer of the aforementioned networks. As it is shown in Fig. 5, maximum accuracy for RML database occurred on $subset_{19}$ and $subset_{23}$ of the validation set in person-dependent and person-independent cases, respectively. In case of SAVEE database, $subset_7$ and $subset_9$ show the maximum accuracy for each of these scenarios. Increasing the number of frames does not increase the accuracy further after maximum point. These subsets are used as the reference for the test data.

Clearly, CNN-based classifiers have a higher performance than the traditional methods like SVM. This is mainly due to the built-in feature extraction ability of these classifiers.

The area under the ROC curve is an evaluation metric for classification problems measured at different thresholds. ROC is a probability curve, and AUC represents the degree of discrimination. In a multiclass classification problem, there exists one ROC for each class based on one versus all methodology. The ROC curves for Gaussian SVM are given in Fig. 6.

4.3 Feature analysis

Despite the fact that one can improve the performance of the person-independent case using deep learning classifiers instead of SVM, we believe investigating the effect of geometric features in keyframe selection helps interpreting the emotion recognition in both cases. To this end, three regions of the face, namely the whole face, the nose and mouth region, and eye region, are selected to study the contribution of geometric features in person-dependent and person-independent scenarios. In the case of whole face, all geometric features represented in Table 1 are used to select the keyframes. In the case of eye region, only the distances and angles related to eyes and eyebrows are used for keyframe selection. Finally, for the nose and mouth region, SVM is trained using keyframes selected based on geometric features of the nose, upper mouth, and lower mouth. Results of the experiments are represented in Table 3. Explicitly, features of eye region are general and these features suffice to provide a fine performance in the case of person-independent scenario. In the case of person-dependent approach, features from the whole face performed slightly better than features of the eyes region.

Table 3 Feature analysis for both scenarios

Data	Region/Scenario	Independent		Dependent	
		Acc	Loss	Acc	Loss
SAVEE	Whole face	34.44	0.675	**98.77**	0.073
	Eyes & eyebrows	**41.11**	0.642	98.46	0.085
	Nose and mouth	35.19	0.672	95.99	0.094
RML	Whole face	36.06	0.654	**95.83**	0.113
	Eyes & eyebrows	**39.76**	0.720	95.77	0.069
	Nose and mouth	37.51	0.659	93.25	0.081

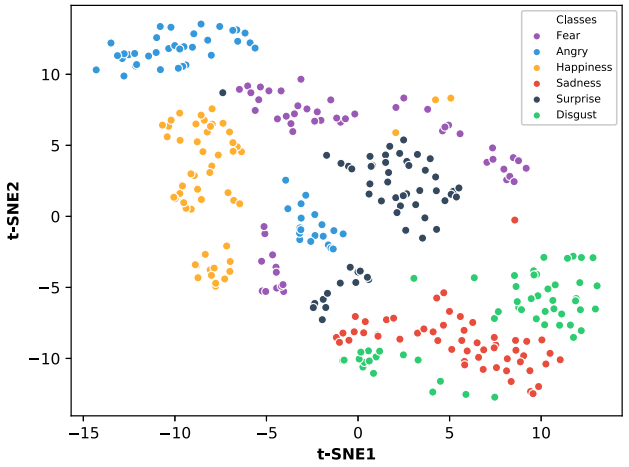
Bold results are the best ones in our experiments

4.4 Qualitative results

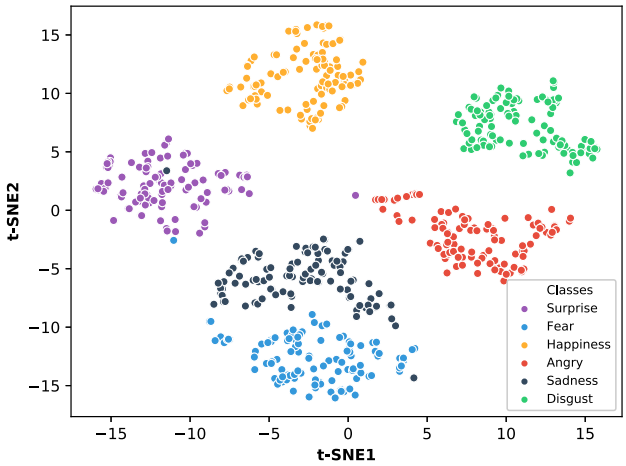
Additionally, to show the significance of features learned by CNN methods for each scenario, we visualize the deeply learned features using t-distributed stochastic neighbor embedding (t-SNE) technique for the highest result among the CNNs (i.e. ResNet50) on SAVEE database. This can show the generality of the learned features. Figure 7 visualizes the features extracted for person-dependent and person-independent scenarios. We qualitatively observe that separation of classes for person-dependent case is higher than the other one. Each video clip is visualized as a point, and samples of same category are illustrated by the same color.

5 Conclusion

This paper defined an explicit definition of person-dependent and person-independent scenarios for evaluating FER systems. We used geometric features for keyframe selection. Then, we compared performance of SVM classifiers in terms of accuracy and computational time with CNN-based classifiers. Person-independent scenarios are more challenging than the person-dependent ones due to no observation of subject's expression. Also, result suggests that the amount of improvement is remarkably high for person-independent scenarios when SVM is replaced by CNN-based classifiers. Our experiments on geometric features show that features of eye region can achieve better results than whole set of them.



(a) person independent



(b) person dependent

Fig. 7 Visualization of features, ResNet50, **SAVEE**

Funding The funding was provided by BAP-C project of Eastern Mediterranean University (Grant No. BAP-C-02-18-0001).

References

1. Hajarolasvadi, N., Demirel, H.: 3d cnn-based speech emotion recognition using k-means clustering and spectrograms. Entropy **21**(5), 479 (2019)
2. Guo, J., Lei, Z., Wan, J., Avots, E., Hajarolasvadi, N., Knyazev, B., Kuharenko, A., Jacques, J.C.S., Baró, X., Demirel, H., et al.: Dominant and complementary emotion recognition from still images of faces. IEEE Access **6**, 26391–26403 (2018)
3. Bolotnikova, A., Demirel, H., Anbarjafari, G.: Real-time ensemble based face recognition system for nao humanoids using local binary pattern. Anal. Integr. Circuits Signal Process. **92**(3), 467–475 (2017)
4. Zen, G., Porzi, L., Sangineto, E., Ricci, E., Sebe, N.: Learning personalized models for facial expression analysis and gesture recognition. IEEE Trans. Multimed. **18**(4), 775–788 (2016)
5. Fierrez-Aguilar, J., Garcia-Romero, D., Ortega-Garcia, J., Gonzalez-Rodriguez, J.: Adapted user-dependent multimodal biometric authentication exploiting general information. Pattern Recognit. Lett. **26**(16), 2628–2639 (2005)
6. Eskandari, M., Toygar, Ö., Demirel, H.: Feature extractor selection for face-iris multimodal recognition. Signal Image Video Process. **8**(6), 1189–1198 (2014)
7. Soleymani, M., Pantic, M., Pun, T.: Multimodal emotion recognition in response to videos. IEEE Trans. Affect. Comput. **3**(2), 211–223 (2012)
8. Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., Anbarjafari, G.: Audio-visual emotion recognition in video clips. IEEE Trans. Affect. Comput. **10**, 60–75 (2017)
9. Xie, Z.: Ryerson Multimedia Research Lab. University of Surrey, Guildford (2014)
10. Jackson, P., Haq, S.: Surrey Audio-Visual Expressed Emotion (Savee) Database. University of Surrey, Guildford (2014)
11. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L-P.: Openface 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 59–66. IEEE (2018)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Preprint arXiv:1409.1556 (2014)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (cvpr). vol. 5, p. 6 (2015)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
15. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., et al.: Going deeper with convolutions. Preprint arXiv:1409.4842, 1409 (2014)
16. Zhang, S., Zhang, S., Huang, T., Gao, W., Tian, Q.: Learning affective features with a hybrid deep model for audio-visual emotion recognition. IEEE Trans. Circuits Syst. Video Technol. **28**(10), 3030–3043 (2018)
17. Seng, K.P., Ang, L.-M., Ooi, C.S.: A combined rule-based & machine learning audio-visual emotion recognition approach. IEEE Trans. Affect. Comput. **9**(1), 3–13 (2018)
18. García, H.F., Álvarez, M.A., Orozco, A.A.: Dynamic facial landmarking selection for emotion recognition using gaussian processes. J. Multimodal User Interfaces **11**(4), 327–340, (2017). ISSN 1783-8738
19. Wang, Y., Guan, L.: Recognizing human emotional state from audiovisual signals. IEEE Trans. Multimed. **10**(5), 936–946 (2008)
20. Doherty, A.R., Byrne, D., Smeaton, A.F., Jones, G.J.E., Hughes, M.K.: Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs. In: Proceedings of the 2008 International Conference on Content-Based Image and Video Retrieval, pp. 259–268. ACM (2008)
21. Guo, S.M., Pan, Y.A., Liao, Y.C., Hsu, C.Y., Tsai, J.S.H., Chang, C.I.: A key frame selection-based facial expression recognition system. In: First International Conference on Innovative Computing, Information and Control, 2006. ICICIC'06. vol. 3, pp. 341–344. IEEE (2006)
22. Zhang, Q., Shao-Pei, Y., Zhou, D.-S., Wei, X.-P.: An efficient method of key-frame extraction based on a cluster algorithm. J. Hum. Kinetics **39**(1), 5–14 (2013)
23. Haq, S., Jackson, P.J.B., Edge, J.: Speaker-dependent audio-visual emotion recognition. In: AVSP, pp. 53–58 (2009)
24. Barros, P., Wermter, S.: Developing crossmodal expression recognition based on a deep neural model. Adapt. Behav. **24**(5), 373–396 (2016)
25. Avots, E., Sapiński, T., Bachmann, M., Kamińska, D.: Audiovisual emotion recognition in wild. Mach. Vis. Appl. 1–11 (2018)