Electronic Research Archive

http://www.aimspress.com/journal/era

ERA, 30(2): 644–660.
DOI: 10.3934/era.2022034
Received: 16 November 2021
Revised: 07 February 2022
Accepted: 14 February 2022
Published: 18 February 2022

Research article

Emotion recognition in talking-face videos using persistent entropy and neural networks

Eduardo Paluzo-Hidalgo*, Rocio Gonzalez-Diaz and Guillermo Aguirre-Carrazana

Department of Applied Mathematics I, University of Seville, Seville, Spain

* **Correspondence:** Email: epaluzo@us.es.

Abstract: The automatic recognition of a person's emotional state has become a very active research field that involves scientists specialized in different areas such as artificial intelligence, computer vision, or psychology, among others. Our main objective in this work is to develop a novel approach, using persistent entropy and neural networks as main tools, to recognise and classify emotions from talking-face videos. Specifically, we combine audio-signal and image-sequence information to compute a *topology signature* (a 9-dimensional vector) for each video. We prove that small changes in the video produce small changes in the signature, ensuring the stability of the method. These topological signatures are used to feed a neural network to distinguish between the following emotions: calm, happy, sad, angry, fearful, disgust, and surprised. The results reached are promising and competitive, beating the performances achieved in other state-of-the-art works found in the literature.

Keywords: topological data analysis; persistent homology; persistent entropy; neural networks; audio-visual emotion recognition; talking-face videos

1. Introduction

(Facial) Emotion recognition consists of a series of processes to detect human emotions from (facial) human expressions. When people communicate with others, they are constantly sending and receiving nonverbal cues, expressed through body gestures, voice, facial expressions, and physiological changes. Nonverbal cues increase trust, clarity and provide more information supporting what spoken words transmit. A particular emotional state produces certain verbal and nonverbal signals that transmit information regarding personal feelings. Nowadays, (facial) emotion recognition has become an important research area in the fields of computer vision and artificial intelligence due to its potential applications. In general, people express their emotional state (such as joy, sadness, or anger) through facial expressions and vocal tones, and these are features that are often analyzed for emotion recognition. Several factors make emotion recognition in talking-face videos difficult to handle (see, for

example [1]). Among others, one of them is that body language cues are not as available to the listener as opposed to having a face-to-face conversation. Besides, rigid movements of the face in a talking-face video can reduce the accuracy of extracting facial features. Another problem is how to measure a ground truth for emotions. Measuring the right emotion at the right time in a talking-face video is, in general, quite subjective and relies entirely on the research setting. Another challenge to take into account is how to deal with facial micro-expressions. In [2], the facial micro-expressions issue is faced using the knowledge distillation approach (see [3]). These spontaneous and low-intensity expressions are the bottleneck of applying deep learning methods due to the huge amount of data needed to train such methods.

Regarding computer-aided emotion recognition research works, roughly speaking, they are focused on the use of various input types such as facial expressions [4–6], speech [7–12] and physical signals [13].

Several emotion classification and recognition techniques have been proposed in the past. Some of them used speech prosody contour information to recognize emotions through different classification methods such as artificial neural networks, multichannel hidden Markov models, mixture of hidden Markov models, and Active Appearance Models.

Regarding audio emotion recognition, in [14], a preprocessing scheme is proposed to remove the noise from speech signals based on fast Fourier transformation and spectral analysis. The authors evaluated their model using benchmark IEMOCAP and EMODB datasets and obtained a high recognition accuracy, which were 73 and 90%, respectively. The authors in [15] faced the task of audio emotion recognition using a convolutional neural network. Their baseline model included one-dimensional convolutional layers combined with dropout, batch normalization, and activation layer. The proposed framework achieved 71.61% of accuracy for the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS dataset [16]). In [17], a novel approach based on persistent entropy (a computational topology tool) is developed to obtain a single value for the audio signal of a given video of a person expressing emotions. These data were later used as the input of a support vector machine to classify audio signals into eight different emotions, namely, neutral, calm, happy, sad, angry, fearful, disgust, and surprised. The results obtained were close to the existing accuracy of methods with a greater scope such as the ones introduced in [18].

Regarding visual emotion recognition, in [19], a topology-based approach was used to understand and discern existing patterns between emotions in talking-face videos. In that paper, the authors used precomputed landmark points to compute the persistence diagrams (a key tool in computational topology [20]) from the Vietoris-Rips filtration in a given frame. The paper focuses on emotions at instant moments, i.e., no interframe relation is considered in the image sequence. Furthermore, the audio signal is neither used in that approach. The feature obtained is the persistence diagram and pairwise comparisons are made using bottleneck or Wassertein distances. The classification of emotions is not considered in that paper.

Regarding multimodal emotion recognition, different approaches have been explored so far. For example, in the H2020 KRISTINA project*, multimodal computer-aided emotion recognition is used to help in the interaction between health professionals and migrated patients, allowing to overcome linguistic barriers that hinder communication. Besides, the MixedEmotions project† developed an ap-

*http://kristina-project.eu/en/

†https://cordis.europa.eu/project/rcn/194226_es.html

plication based on a complete emotional profile of the person's behavior. It used data from different channels: multilingual text, data sources, audio and video signals, social media, and structured data. The project offered commercial solutions providing an integrated big linked data platform for emotional analysis, using heterogeneous sets of data and addressing the multilingual and multimodality aspects in a robust and large-scale setting.

Moreover, hybrid neural networks combining convolutional neural networks and recurrent neural networks have become the state-of-art for multimodal emotion recognition. For example, the authors in [21] proposed an audiovisual-based hybrid network that combines the predictions of five models for emotion recognition in the wild. The overall accuracy of the proposed method achieved 55.61 and 51.15% classification accuracy for the audio-only and video-only dataset, respectively. The authors used the Afew-va Database for Valence and Arousal Estimation In-the-wild introduced in [22]. In general, solutions based on deep learning methods require a large amount of training data and, up to now, the public datasets are limited [3].

In this paper, we consider both the audio signal and the image sequence of a given talking-face video and we use a novel tool introduced in [23], called persistent entropy, to "vectorize" the topological information provided by the persistence diagrams. Such vectors are used to feed a neural network to classify emotions. Specifically, the method incorporates the idea consisting of computing the persistent entropy of the lower-star filtration of the 1-dimensional simplicial complex obtained by a discretization of the audio signal, together with a particular construction of a 3-dimensional cell complex obtained from the image sequence as follows. First, the precomputed landmark points are used to build the 2-dimensional Delaunay triangulation in each frame, and then they are stacked in a particular way to obtain a 3-dimensional cell complex. Later, we compute the persistent entropy of each of the eight filtrations, considering, respectively, the distance to eight fixed planes (two horizontal planes, two vertical planes, and four oblique planes). This way, the method is able to completely capture the movement in the image sequence. The computed topological features together with the features obtained from the audio-signal, make up a 9-dimensional vector, also called the *topological signature*, of the given talking-face video. Besides, thanks to [24], we are able to prove that topological signatures are stable to small changes in the input audio signal and image sequence. Finally, the topological signatures computed are used to train a neural network to classify emotions.

As said before, the use of persistent entropy to compute the topological signature of a talking-face video is supported by its demonstrated stability under small perturbations in the input data [24] and by numerous applications of this technique in other fields such as, for example, pattern recognition [25], complex systems [26], clustering [27], heart-rate-based sleep-wake classification [28], glioblastoma detection [29], and audio emotion recognition [17].

The idea of computing a 3-dimensional cell complex from an image sequence is not new. In [30], this idea was used for gait classification, but there, the 3-dimensional simplicial complex was obtained from the silhouettes of the person walking in the video. That methodology was also used in [31] to monitor human activities at distance and in [32] for gait-based gender classification.

The novelty of the method presented here is the computation of a topological signature associated to a talking-face video, combining topological information from the audio signal and the image sequence. This topological signature is obtained by computing the persistent entropy of certain filtrations constructed on specific cell complexes aimed to capture topological changes along the video that characterize the different emotions considered. The topological signatures computed are then used to

feed a neural network to classify emotions. Let us observe that the neural network considered in this paper is extremely simple because of the low dimension of the input.

The contribution of this work is a workflow of emotion recognition, being its stability guaranteed by topology-based theoretical results. The experimentation section shows that our method outperforms several state-of-the-art emotion recognition methods.

The paper is structured as follows. The needed background is introduced in Section 2. The description of the proposed method is provided in Section 3. The stability of the method is proven in Section 4. The experimentation made is presented in Section 5, together with comparisons with state-of-the-art methods. Finally, Section 6 provides conclusions and future work ideas.

2. Background

In this section, the main concepts of topological data analysis and neural networks, needed to understand our method for facial emotion recognition, are recalled.

2.1. Topological data analysis

Topological data analysis has emerged as an important approach to characterize the behavior of datasets using techniques from topology. Tools from topological data analysis, specifically persistent homology, allow assigning shape descriptors to large and noisy data across a range of spatial scales. We will compute such a descriptor to capture the topological changes along the video and these changes will be used to classify emotions.

To compute the persistent homology, we first have to provide the input data with a combinatorial structure that reflects the topology of the underlying space where the input data lay. The combinatorial structure used in this work is the one of cell complexes, whose elements in each dimension d, called d-cells, are d-dimensional topological spaces homeomorphic[‡] to a d-dimensional ball. This way, a 0-dimensional cell is a point (vertex), a 1-dimensional cell is a curve, a 2-dimensional cell is homeomorphic to a disk, and so on. A *cell complex* K is a collection of cells constructed inductively: 1) The 0-skeleton $K^{(0)}$ (i.e., the set of 0-cells of K) is a set of points in an ambient n-dimensional space $\mathbb{R}^n$. 2) The d-skeleton $K^{(d)}$ is constructed from the $(d-1)$-skeleton $K^{(d-1)}$ by attaching d-cells via homeomorphisms.

From now on, we will assume that the given cell complex K has a finite number of cells. The boundary set of a d-cell $\sigma \in K$ can be informally defined as the set of $(d-1)$-cells in the $(d-1)$-skeleton $K^{(d-1)}$ used to attach the d-cell σ. Successively adding to $F = \{\sigma\}$ the boundary set of each cell in F, we obtain the set of faces of σ. For example, the boundary set of an edge is its two endpoints (vertices). A d-dimensional cell complex K is a cell complex satisfying that the dimension of the cell of the higher dimension in K is d. A subcomplex of a cell complex K is a subset $K' \subset K$ which itself is still a cell complex. An example of a subcomplex is the closed star of a vertex v, denoted by St v and defined as follows: A cell σ is in St v if there exists $\mu \in K$ such that σ and v are faces of μ. A filtration is an increasing sequence of cell complexes $\emptyset \subset K_1 \subset K_2 \subset \cdots \subset K_r = K$. See an example of a filtration of cell complexes in Figure 1.

There are several methods to compute cell complexes and filtration from input data depending on the nature of the data and the purpose of the analysis. In this work, given a talking-face video, we will

[‡]A homeomorphism is a bicontinuous and bijective function between two topological spaces.

collect the set of landmark points S precomputed in the image sequence of the video. Such landmark points will be embedded in $\mathbb{R}^3$, being the last coordinate, the position of the frame where the landmark points are allocated. Let V_s be the set of points of $\mathbb{R}^3$ that are closer to $s \in S$ than to any other point of S. That is, for $s \in S$, $V_s = \{x \in \mathbb{R}^3 \mid d(x, s) \leq d(x, s')\ \forall s' \in S\}$. The collection of the sets V_s is a covering for $\mathbb{R}^3$ and it is called the Voronoi decomposition of $\mathbb{R}^3$ concerning S. The nerve of this covering is a a simplicial complex called the Delaunay triangulation of S. The construction of this complex is costly in high dimensions, although there exist efficient algorithms for computing it when $n = 2$ and $n = 3$. See [33] for more details on Voronoi diagrams and Delaunay triangulation.

The filtration considered in this paper is the lower-star filtration (see [20, page 135]). Let us see how to define it. Consider a real-valued function h, also called *height function*, on a finite set of points V. Suppose K is a cell complex with set of vertices V. The lower star of $v \in V$ is defined as the subset of cells of K for which v is the vertex with maximum function value, that is, $\operatorname{low}\operatorname{St} v = \{\sigma \in \operatorname{St} v : x \in \sigma \Rightarrow h(x) \leq h(v)\}$. Sort the vertices by their function values, in a non-decreasing order, $V = \{v_1, v_2, \ldots, v_r\}$. The lower-star filtration $K_1 \subset K_2 \subset \cdots \subset K_r = K$ satisfies that K_j is the union of the lower stars of the first j vertices of V, that is, $K_j = \bigcup_{i \leq j} \operatorname{low}\operatorname{St} v_i$, for all j.

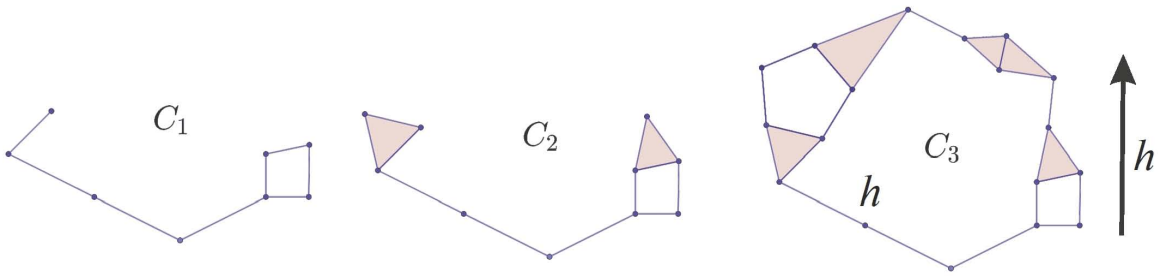

Figure 1. From left to right, the lower-star filtration $K_1 \subset K_2 \subset K_3 = K$ obtained using the height function h.

The next step is to compute the persistent homology of a filtration $K_1 \subset K_2 \subset \cdots \subset K_r = K$ [20, 34] that tracks the moment t where a homology class is born and the moment s where the same class dies leading to a topological descriptor called persistence diagram. Specifically, a d-dimensional homology class of a cell complex K_t is an element of the d-dimensional homology group $H_d(K_t)$ which is defined as follows. First, the d-dimensional chain group $C_d(K_t)$ is obtained by summing up[§] the d-cells of K_t. Then, the boundary operator is extended to a linear map ∂_d from $C_d(K_t)$ to $C_{d-1}(K_t)$ in an obvious way. Since the boundary of the boundary of a cell is always zero, then the image $B_d(K)$ of ∂_{d+1} is a subgroup of the kernel $Z_d(K)$ of ∂_d. The d-dimensional homology group of K_t is defined as the quotient group $H_d(K_t) = B_d(K_t)/Z_d(K_t)$. Then, the given filtration leads a family of homology groups and homomorphisms $\{H_d(K_t) \to H_d(K_s) : t \leq s\}$ from which we can track the births and deaths of homology classes. Now, each homology class α that was born in $H_d(K_t)$ and died in $H_d(K_s)$ can be stored as a point (t, s). The result is a multiset of points in $\mathbb{R}^2$ called the persistence diagram for the given filtration. The persistence of the homology class α is the difference $\operatorname{pers}(\alpha) = s - t$. In this work, homology classes with infinity persistence correspond to points of the form $(t, N + 1)$, where N is a fixed big positive integer. This way, all points in the persistence diagram have finite coordinates. The

[§]The ground ring considered in this paper is $\mathbb{Z}/\mathbb{Z}2$.

features of higher persistence are represented by the points furthest from the diagonal, while points nearby to the diagonal may be interpreted as topological noise.

Finally, we summarize the information described by a persistence diagram in a quantity called persistent entropy (introduced in [23]) which consists in the Shannon entropy of the probability distribution obtained from the given persistence diagram. Specifically, given a filtration and the corresponding persistence diagram $\mathrm{Dgm} = \{(a_j, b_j) : j \in J\}$, the persistent entropy of the filtration is defined as $E = -\sum_{j\in J} p_j \log(p_j)$ where $p_j = \frac{\ell_j}{L}$, $\ell_j = b_j - a_j$, and $L = \sum_{j\in J} \ell_j$. Let us notice that if $p_j \leq 1$, then $\log(p_j) \leq 0$, so the persistent entropy is always positive. Intuitively, the persistent entropy measures how different the persistence of the homology classes that appear along the filtration are.

2.2. Neural networks

In this paper, we deal with a supervised classification problem where a set of labelled examples are provided with the aim of making predictions for unlabelled points. A widely extended machine learning model for classification problems is neural networks. In general, we could say that a neural network is a mapping $\mathcal{N}_{\omega,\Phi} : \mathbb{R}^n \to \mathbb{R}^k$ that depends on a set of weights ω and a set of parameters Φ describing the synapses between neurons, layers, activation functions and any other characteristic of its architecture. A good introduction to artificial neural networks was given in [35].

The specific kind of neural network architecture used in this paper is a feedforward neural network composed of a set of neurons hierarchically organized in layers that are fully connected. In this paper, a $9 \times 512 \times 128 \times 64 \times 7$ fully-connected feedforward neural network will be used (see Figure 2). Neural networks can be seen as directed graphs where the input is transmitted and transformed along the graph using different activation functions such as ReLu, sigmoid, or softmax. In this paper, the ReLU activation functions are used in the hidden layers and the Softmax activation function in the output layer.

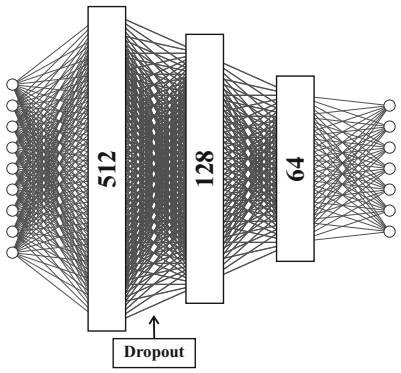

Figure 2. A $9\times512\times128\times64\times8$ fully-connected feedforward neural network composed of an input layer with 9 neurons, three hidden layers with 512, 128 and 64 neurons, respectively, and an output layer with 7 neurons.

To train the neural network $\mathcal{N}_{w,\Phi}$ for a supervised classification task, we will use a labelled dataset $D = \{(x, c_x)\}$ consisting of a finite set of pairs where, for each pair (x, c_x), point x lies in $\mathbb{R}^n$ and label c_x lies in $\{0, 1, \ldots, k\}$, for some $k \in \mathbb{N}$. During the training process, the set of weights of the neural network is updated trying to minimize a loss function which measures the difference between the output of the network (obtained with the current weights) and the desired output (dictated by the labelled dataset).

The loss function used in this paper is the cross-entropy loss function which is related to the Kullback-Leibler divergence: Given two probability distributions $P(x)$ and $Q(x)$ over the same random variable x, the cross-entropy is computed as $H(P, Q) = -\sum_{(x,c_x)\in D} P(x)\log(Q(x))$. To iteratively update the weights, the loss-driven training method used in this paper is the Adam algorithm (introduced in [36]) which is a stochastic gradient-based optimization algorithm.

The goal of training a neural network is generalization. That is, we want our neural network to learn from the given data and to apply the learnt information to new data. One way to measure the performance of the trained neural network is to split the given dataset into two subsets called the training set and the test set. When the trained neural network reaches high accuracy on the training set but performs badly on new data, we say that there is an overfitting. Among the different approaches to prevent overfitting existing in the literature, in this paper, we will use dropout regularization that consists in randomly invalidating a certain percentage of the neurons of the neural network during the training procedure (consult [37] for more information).

3. Description of the method

In this section, we develop an emotion recognition method using persistent entropy and neural networks as the main tools. Overall, the method works as follows. The input data are talking-face videos with precomputed facial landmark points. For each video, we compute a topological feature obtained from the audio signal together with eight topological features obtained from the image sequence, deriving a 9-dimensional vector called the topological signature of the video. The set of topological signatures obtained from the video dataset will then be used to feed a neural network. The summary of the process workflow is outlined in Figure 3.

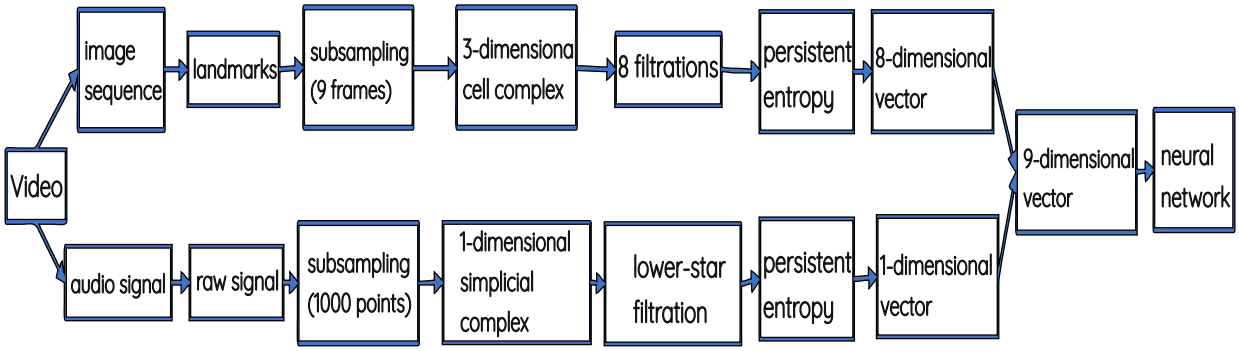

Figure 3. Summary of the process workflow for emotion recognition used in this paper.

Let us explain, step by step, the procedure outlined above. We start by extracting the landmark points on each frame of the input image sequence (see Figure 4). We assume that these landmark points are already precomputed in the given dataset.

For each frame, we use the landmark points to compute the 2-dimensional simplicial complex that consists of the Delaunay triangulation of the set of points corresponding to the spatial position of the landmark points. To connect the topological information along the image sequence, the landmark points corresponding to the same part of the face in consecutive frames are joined by an edge. A 2-dimensional cell is obtained when the two endpoints of an edge are joined to the two endpoints of the corresponding edge in the neighbor frame. A 3-dimensional cell is obtained when the vertices

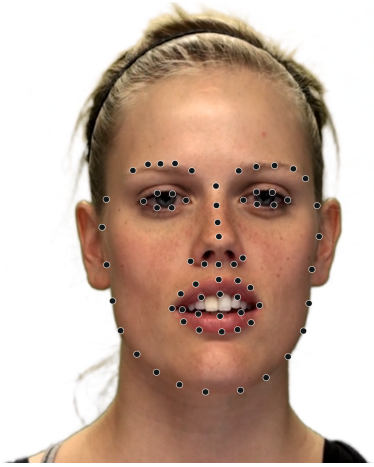

Figure 4. The landmark points considered in this paper drawn on a face in a video frame extracted from the RAVDESS dataset.

of a triangle of the Delauney triangulation associated to one frame are joined with the vertices of the corresponding triangle in the neighbor frame.

The output of the steps described above is a 3-dimensional cell complex K for each input image sequence, which condenses all gestures the person is making while recording on video. In Figure 5, the 1-skeleton of the 3-dimensional cell complex K obtained from an image sequence, is pictured.

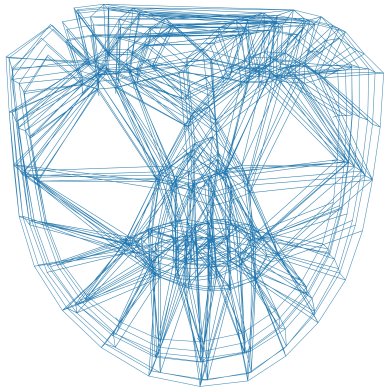

Figure 5. The 1-skeleton of the cell complex K obtained from an image sequence of a video extracted from the RAVDESS dataset.

The next step in this process is to sort the cells of K in different ways to obtain different filtrations with the aim of capturing the small details that characterize an emotion. In this work, eight different filtrations (two horizontal, two vertical, and four obliques) are used to obtain eight different persistence diagrams (see Figure 6 to have intuitions). The way to define a filtration is as follows: Given a plane π, we define the filter function $h_\pi : K \to \mathbb{R}$ that assigns to each vertex of K its distance to the plane π, and to any other cell of K, the maximum distance of its vertices to the plane π. The cells are sorted according to the function values of their vertices, and then, the lower-star filtration K_π associated with the plane π is computed.

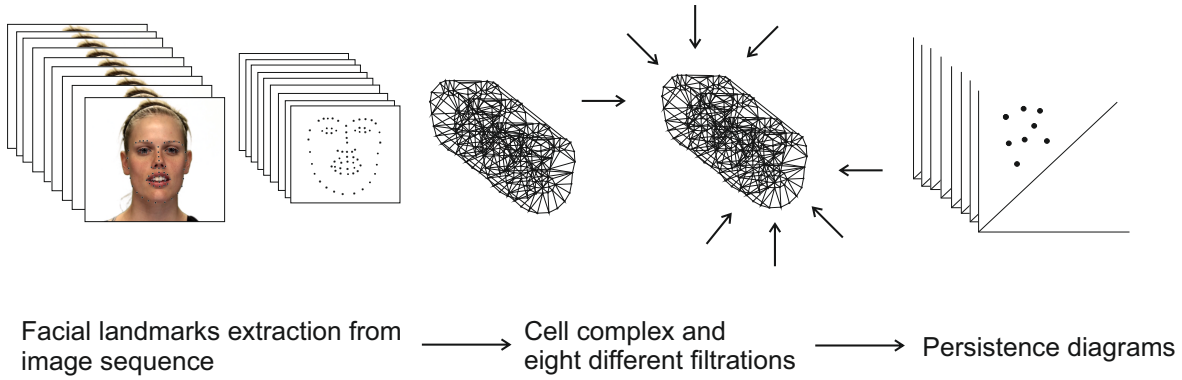

Figure 6. Illustration of the steps followed to compute the eight persistent entropy values from an image sequence with precomputed landmark points. From left to right: firstly, the landmark points are extracted from the image sequence. Then, a cell complex K is computed. Eight different filtrations are used to obtain eight persistence diagrams, one for each filtration. Finally, the eight persistent entropy values are computed, one for each persistence diagram.

Next, the persistence diagram is computed for each of the eight filtrations. The algorithm used for this step is described in Algorithm 1 with complexity $O(n^3)$ in theory but linear in practice [20, page 159].

Algorithm 1: Computing the persistence diagram for a filtration [38].

Input: A filtration $\emptyset = K_0 \subset K_1 \subset K_2 \subset \cdots \subset K_n = K$ and an ordering of the cells $\{\sigma_1, \ldots, \sigma_m\}$ of K such that if $i < j$ then $\text{ind}(\sigma_i) < \text{ind}(\sigma_j)$ where $\text{ind}(\sigma_i) = \min\{r : \sigma_i \in K_r\}$

Output: The persistence diagram Dgm

Initialize $H = \emptyset$, Dgm $= \emptyset$, and $f(\sigma_i) = 0$ for $i \in \{1, \ldots, m\}$

for $i = 1$ **to** m **do**

  if $f\partial(\sigma_i) == 0$ **then**

    $H \cup \{\sigma_i\}$ (a new homology class was born)

    $f(\sigma_i) = \sigma_i$

    Dgm $\cup\{(\text{ind}(\sigma_i), \infty)\}$

  if $f\partial(\sigma_i) \neq 0$ **then**

    Let $\sigma_j \in f\partial(\sigma_i)$ such that $j == \max\{\text{ind}(\mu) : \mu \in f\partial(\sigma_i)\}$

    $H \setminus \{\sigma_j\}$ (an homology class died)

    foreach $x \in K$ *such that* $\sigma_j \in f(x)$ **do**

      $f(x) = f(x) + f\partial(\sigma_i)$.

    Dgm $\setminus\{(\text{ind}(\sigma_j), \infty)\} \cup \{(\text{ind}(\sigma_j), \text{ind}(\sigma_i))\}$

The persistent entropy is then computed for each of the eight persistence diagrams. Due to its formulation, persistent entropy can be computed in linear time. As a result, an 8-dimensional vector is obtained for each image sequence.

Besides, for each talking-face video, we add a new entry to the 8-dimensional vector computed consisting of the persistent entropy of the lower-star filtration obtained from the 1-dimensional simplicial complex computed from the raw audio signal of the video as it is done in [17].

Putting all together, we obtain a 9-dimensional feature vector called the *topological signature* of the video. Finally, the topological signatures computed from the talking-face video dataset are then used to train a feed-forward neural network to classify the videos into the different emotions considered.

4. Stability of the method

Thanks to the work presented in [24] we have the following result.

Lemma 4.1. *The so-called topological signature associated with a given talking-face video is stable in the sense that small changes in the input video produce small changes in the topological signature.*

Proof. In [24], it is proved that persistent entropy is stable. It means that small changes in the input data produce small changes in the persistent entropy value. In this case, the input data are, first, the eight filtrations obtained from the image sequence and, second, the filtration obtained from the audio signal. Small perturbations in the filtrations are equivalent to a small displacement of the landmark points in the image sequence or small changes in the audio signal, that is, they consist of small perturbations in the input data used to compute the persistent entropy values, concluding the proof. □

5. Experimentation

For experimentation, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS [16]) is used, that is a talking-face video dataset where facial landmark points composed of 62 points have been precomputed. This dataset contains the vocalization of two statements in a neutral North American accent by 24 professional actors (12 female, 12 male). Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. The intensity fulfils an important role in emotional theory (see the works in [39, 40]). The strong intensity is useful when we are looking for clear emotional examples. However, as explained in [41], the normal intensity is generally used if we are interested in providing classification for daily life. All actors produced 60 spoken expressions and 44 sung expressions. These vocalizations are available in three formats: audio-only, video-only and audio-video.

In this paper, we focus on the 60 speech videos provided in the RAVDESS video dataset. The total tracked files used is 24 actors $\times$ 60 speeches. Since we do not consider neutral emotions to avoid an unbalanced dataset, we used a total of 1344 videos.

The number of frames used as well as the number of points in the subsampled audio signal were an experimental choice consisting of the minimum number of frames and points needed to obtain good results and to develop the experiment in a feasible time. The neural network considered was the simplest one that provided satisfactory results and the weights of the neural network were tuned using a traditional training procedure.

The steps for experimentation follow the methodology explained in Section 3 and outlined in Figure 3. We first consider the image sequence obtained for each video of the audio-video dataset. For each image sequence, nine equally spaced frames were selected to have an appropriate representation of the full image sequence. The landmark points of those 9 frames were then used to build an 8-dimensional vector following the method described in Section 3. Then, for each video, we added a new entry to the 8-dimensional vector computed consisting of the persistent entropy of the lower-star filtration of the

Table 1. Confusion matrix of the audio-video experiment for one of the repetitions measured on the test dataset.

Emotion	Calm	Happy	Sad	Angry	Fearful	Disgust	Surprised
Calm	61	0	0	0	0	0	3
Happy	0	48	0	4	0	0	0
Sad	0	0	55	0	0	0	0
Angry	0	0	0	60	0	0	2
Fearful	0	0	0	0	60	0	2
Disgust	0	0	0	0	0	45	1
Surprised	0	0	0	0	4	0	55

1-dimensional simplicial complex obtained from a subsampling of the raw audio signal consisting of 10,000 points. The subsampling process was done uniformly on the signal, maintaining its shape and main distribution of the spikes. As a result, we obtained a set of 1344 9-dimensional feature vectors, one for each video of the dataset considered. Finally, this set was split into a training set with 944 vectors and a test set for validation with 400 vectors.

Then, the training set was used to train a neural network with the following standard architecture: It is composed by five layers with a total of $n \times 512 \times 128 \times 64 \times 7$ neurons, using dropout (20%) in the first hidden layer with $n = 9$ being the dimension of the input (i.e., the 9-dimensional topological feature vectors). The ReLU activation function is used in the hidden layers and the Softmax activation function in the output layer. We used the default learning rate in the Tensorflow package, which is 0.001.

The neural network was trained during 500 epochs and the experiment was repeated 10 times using sparse categorical cross-entropy as the loss function and the Adam training algorithm. The accuracy values for those repetitions are shown in Figure 7 for the training set and in Figure 8 for the test set.

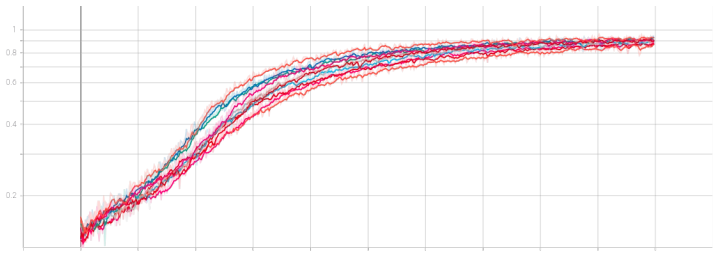

Figure 7. Accuracy values on the training set during 500 epochs. 10 repetitions.

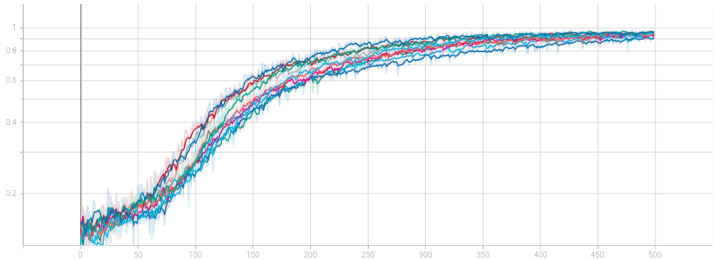

Figure 8. Accuracy values on the test set during 500 epochs. 10 repetitions.

The highest values reached were 99.9% of accuracy on the training set with 98.02% of accuracy on the test set. Average accuracy was 95.97% on the test set. A confusion matrix for the experiment is shown in Table 1.

To test the importance of the eight filtrations in the computation of the topological signatures used to feed the neural network in the methodology presented in this paper, Table 2 shows how the accuracy of the feed-forward neural network applied increases when we use more filtrations, concluding that using the eight filtrations we reach the highest accuracy. Notice that all the parameters considered to train and test the neural network are the same as above, except for the input layer that increases in each experiment from $n = 2$ to $n = 9$ at the same time that we increase the number of filtrations considered.

Table 2. Experimental results highlighting the need of the eight different filtrations to compute the topological signature of a talking-face video to recognise emotions.

Number of filtrations		Filtrations	Average accuracy on the test set
Audio signal	Image sequence		
1	1	$\rightarrow$	21.49%
1	3	$\rightarrow \uparrow \nearrow$	33.47%
1	5	$\rightarrow \uparrow \nearrow \nwarrow$	66.05%
1	6	$\rightarrow \uparrow \nearrow \nwarrow \downarrow$	75.05%
1	7	$\rightarrow \uparrow \nearrow \nwarrow \downarrow \swarrow$	79.98%
1	8	$\rightarrow \uparrow \nearrow \nwarrow \leftarrow \downarrow \swarrow \searrow$	95.97%

The state-of-the-art methods we compare ours to are [42–45]. All of them used the audio-video RAVDESS dataset. As we can see in Table 3, our method outperforms them. In [42], a model is proposed based on three deep networks that are fed by image sequences, facial landmark points, and acoustic features, respectively. Such a method depends on three deep neural networks and they might need more input data to reach higher accuracy. Nonetheless, no drawback is reported in that paper, perhaps because it outperformed the state-of-the-art methods at the time of publication. In [43], the fusion of visible images and infrared images with speech are used to feed an ensemble method based on convolutional neural networks. The method does not use the precomputed landmark points provided in the RAVDESS dataset. Nevertheless, a similar accuracy to that of the method presented in [42] was obtained, perhaps because a bigger training data was used, obtained through augmentation methods, to feed the deep neural network system designed. In [44], the authors proposed a multimodal emotion recognition. For the speech-based modality, they obtained good accuracy results when used transfer learning techniques, confirming that the training was more robust when it did not start from scratch and the tasks were similar. Regarding the facial emotion recognizers, they propose a pre-trained Spatial

Table 3. Comparison of our method with state-of-the-art methods.

Paper	Year	Dataset	Average accuracy on the test set
[42]	2020	RAVDESS	87.11%
[43]	2020	RAVDESS	86.36%
[44]	2021	RAVDESS	80.08%
[45]	2020	RAVDESS	76.30%
Our method	–	RAVDESS	**95.97%**

Transformer Network with saliency maps and facial images followed by a bi-LSTM with an attention mechanism. These two modalities were then combined with a late fusion strategy. As the authors claimed, the method lacked in modeling the dynamic nature of the emotions represented in the image sequence. Finally, in [45], an attention mechanism is used as a powerful approach for sequence modeling, achieving an enhanced multimodal emotion recognition and highlighting the importance of exploiting the temporal strength of audio and video signals for emotion recognition.

The advantages of the tools applied in our approach are several, including the robustness to perturbations in the input data that is theoretically guaranteed, and the low-dimensional representation obtained using persistent entropy, resulting in simple input data for almost any kind of machine learning model. The latter advantage also allows for fast training and easy model tuning. Roughly speaking, the crucial part of the classification is not the machine learning model but the robust, explainable and interpretable preprocessing persistent homology application.

6. Conclusions and future works

In this work, we have developed a novel method using persistent entropy and neural networks for emotion classification of talking-face videos. The results reached are promising and competitive, beating the performance reached in other state-of-the-art works found in the the literature. We combined audio-signal and image-sequence information to develop our topology-based emotion recognition method. The main drawback of our methodology is the need of precomputed landmarks and a video long enough to be able to select a representative subset of frames to compute the cell complex. This fact makes our method not useful in real-time applications.

The following future works are planned to be explored: To expand the topological signature by extracting more information from the audio signals. To divide the landmark points into different subsets to determine regions or pairs of regions that contain discriminative landmark points for each facial expression. To use the 3-dimensional information provided by the landmark points. To take advantage of the depth information of the landmark points could be a challenging problem for the future together with considering higher dimensional topological information once that we increase the dimension of the data we are dealing with.

Code availability

The code developed is available at the link https://github.com/Cimagroup/AudioVisual-EmotionRecognitionUsingTDA. All the parameters are provided in the implementation to be able to perform a complete replication of the experiments using the RAVDESS database and the provided code. If other dataset different to RAVDESS is used, then the facial landmark points should be computed before applying the algorithm proposed in this paper.

Acknowledgments

The work was partly supported by the Agencia Estatal de Investigación/10.13039/501100011033 grant PID2019-107339GB-100 and the Agencia Andaluza del Conocimiento grant P20-01145.

Conflict of interest

The authors declare that there are no conflicts of interest.

References

1. E. Ertay, H. Huang, Z. Sarsenbayeva, T. Dingler, Challenges of emotion detection using facial expressions and emotion visualisation in remote communication, in *Processing of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Academic Press, (2021), 230–236. https://doi.org/10.1145/3460418.3479341
2. B. Sun, S. Cao, D. Li, J. He, Dynamic micro-expression recognition using knowledge distillation, *IEEE Trans. Affect. Comput.*, (2020), In press. https://doi.org/10.1109/TAFFC.2020.2986962
3. J. Gou, B. Yu, S. J. Maybank, D. Tao, Knowledge distillation: A survey, *Int. J. Comput. Vis.*, **129** (2021), 1789–1819. https://doi.org/10.1007/s11263-021-01453-z
4. I. Ofodile, K. Kulkarni, C. A. Corneanu, S. Escalera, X. Baro, S. Hyniewska, et al., Automatic recognition of deceptive facial expressions of emotion, *Comput. Sci.*, 2017. https://arxiv.org/abs/1707.04061.
5. S. Shojaeilangari, W. Y. Yau, E. K. Teoh, Pose-invariant descriptor for facial emotion recognition, *Mach. Vis. Appl.*, **27** (2016), 1063–1070. https://doi.org/10.1007/s00138-016-0794-2
6. J. Wan, S. Escalera, G. Anbarjafari, H. J. Escalante, X. Baró, I. Guyon, et al., Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges, in *IEEE International Conference on Computer Vision Workshop*, (2017), 3189–3197. https://doi.org/10.1109/ICCVW.2017.377
7. E. Avots, T. Sapiński , M. Bachmann , D. Kamińska, Audiovisual emotion recognition in wild, *Mach. Vis. Appl.*, **30** (2019), 975–985. https://doi.org/10.1007/s00138-018-0960-9
8. A. Kleinsmith, N. Bianchi-Berthouze, Affective body expression perception and recognition: A survey, *IEEE Trans. Affect. Comput.*, **4** (2012), 15–33. https://doi.org/10.1109/T-AFFC.2012.16
9. C. T. Lu, C. W. Su, H. L. Jiang, Y. Y. Lu, An interactive greeting system using convolutional neural networks for emotion recognition, *Entertain. Comput.*, **40** (2022), 100452. https://doi.org/10.1016/j.entcom.2021.100452
10. F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, G. Anbarjafari, Survey on emotional body gesture recognition, *IEEE Trans. Affect. Comput.*, **12** (2018), 505–523. https://doi.org/10.1109/TAFFC.2018.2874986
11. P. Pławiak, T. Sośnicki, M. Niedźwiecki, Z. Tabor, K. Rzecki, Hand body language gesture recognition based on signals from specialized glove and machine learning algorithms, *IEEE Trans. Industr. Inform.* **12** (2016), 1104–1113. https://doi.org/10.1109/TII.2016.2550528
12. T. Sapiński, D. Kamińska, A. Pelikant, C. Ozcinar, E. Avots, G. Anbarjafari, Multimodal database of emotional speech, video and gestures, in *Pattern Recognition and Information Forensics, ICPR 2018 Lecture Notes in Computer Science*, **11188** (2019). https://doi.org/10.1007/978-3-030-05792-3_15

13. R. Jenke, A. Peer, M. Buss, Feature extraction and selection for emotion recognition from eeg, *IEEE Trans. Affect. Comput.*, **5** (2014), 327–339. https://doi.org/10.1109/TAFFC.2014.2339834

14. S. Kwon, Mlt-dnet: Speech emotion recognition using 1d dilated cnn based on multi-learning trick approach, *Expert Syst. Appl.*, **167** (2021), 114177. https://doi.org/10.1016/j.eswa.2020.114177

15. D. Issa, M. F. Demirci, A. Yazici, Speech emotion recognition with deep convolutional neural networks, *Biomed. Signal Process. Control*, **59** (2020), 101894. https://doi.org/10.1016/j.bspc.2020.101894

16. S. R. Livingstone, F. A. Russo, The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english, *Plos One*, **13** (2018), 1–35. https://doi.org/10.1371/journal.pone.0196391

17. R. Gonzalez-Diaz, E. Paluzo-Hidalgo, J. F. Quesada, Towards emotion recognition: A persistent entropy application, in *Processing of the International Conference on Computational Topology in Image Context*, Academic Press, (2019), 96–109. https://doi.org/10.1007/978-3-030-10828-1_8

18. B. Zhang, G. Essl, E. M. Provost, Recognizing emotion from singing and speaking using shared models, in *Processing of the IEEE International Conference on affective computing and intelligent interaction*, Academic Press, (2015), 139–145. https://doi.org/10.1109/ACII.2015.7344563

19. H. Elhamdadi, S. Canavan, P. Rosen, Affective TDA: Using topological data analysis to improve analysis and explainability in affective computing, *IEEE Trans. Vis. Comput. Graph.*, **28** (2021), 769–779. https://doi.org/0.1109/TVCG.2021.3114784

20. H. Edelsbrunner, J. Harer, Computational topology: an introduction, *Am. Math. Soc.*, Academic Press, (2010). https://doi.org/10.1090/mbk/069

21. X. Guo, L. F. Polanía, K. E. Barner, Audio-video emotion recognition in the wild using deep hybrid networks, 2020. https://arxiv.org/abs/2002.09023.

22. J. Kossaifi, G. Tzimiropoulos, S. Todorovic, M. Pantic, Afew-va database for valence and arousal estimation in-the-wild, *Image Vis. Comput.*, **65** (2017), 23–36. https://doi.org/10.1016/j.imavis.2017.02.001

23. H. Chintakunta, T. Gentimis, R. Gonzalez-Diaz, M. J. Jimenez, H. Krim, An entropy-based persistence barcode, *Pattern Recognit.*, **48** (2015), 391–401. https://doi.org/10.1016/j.patcog.2014.06.023

24. N. Atienza, R. Gonzalez-Diaz, M. Soriano-Trigueros, On the stability of persistent entropy and new summary functions for topological data analysis, *Pattern Recognit.*, **107** (2020), 107509. https://doi.org/10.1016/j.patcog.2020.107509

25. M. Rucco, R. Gonzalez-Diaz, M. J. Jimenez, N. Atienza, C. Cristalli, E. Concettoni, et al., A new topological entropy-based approach for measuring similarities among piecewise linear functions, *Signal Process.*, **134** (2017), 130–138. https://doi.org/10.1016/j.sigpro.2016.12.006

26. A. Myers, E. Munch, F. A. Khasawneh, Persistent homology of complex networks for dynamic state detection, *Phys. Rev. E*, **100** (2019), 022314. https://doi.org/10.1103/PhysRevE.100.022314

27. X. Wang, F. Sohel, M. Bennamoun, Y. Guo, H. Lei, Scale space clustering evolution for salient region detection on 3d deformable shapes, *Pattern Recognit.*, **71** (2017), 414–427. https://doi.org/10.1016/j.patcog.2017.05.018

28. Y. M. Chung, C. S. Hu, Y. L. Lo, H. T. Wu, A persistent homology approach to heart rate variability analysis with an application to sleep-wake classification, *Front. Phys.*, **12** (2021), 202. https://doi.org/10.3389/fphys.2021.637684

29. M. Rucco, G. Viticchi, L. Falsetti, Towards personalized diagnosis of glioblastoma in fluid-attenuated inversion recovery (flair) by topological interpretable machine learning, *Electr. Eng. Syst. Sci.*, **8** (2020), 770. https://doi.org/10.3390/math8050770

30. J. Lamar-Leon, R. Alonso-Baryolo, E. Garcia-Reyes, R. Gonzalez-Diaz, Persistent homology-based gait recognition robust to upper body variations, in *Processing of the 23rd International Conference on Pattern Recognition*, Academic Press, (2016), 1083–1088. https://doi.org/10.1109/ICPR.2016.7899780

31. J. Lamar-Leon, R. Alonso-Baryolo, E. Garcia-Reyes, R. Gonzalez-Diaz, Topological features for monitoring human activities at distance, in *Processing of the 2nd International Workshop on Activity Monitoring by Multiple Distributed Sensing*, **8703** (2014), 40–51. https://doi.org/10.1007/978-3-319-13323-2

32. J. Lamar-Leon, A. Cerri, E. Garcia-Reyes, R. Gonzalez-Diaz, Gait-based gender classification using persistent homology, in *Processing of the 18th Iberoamerican Congress on Progress in Pattern Recognition, Image Analysis, Computer Vision, and Apps*, **8259** (2013) 366–373. https://doi.org/10.1007/978-3-642-41827-3_46

33. C. D. Toth, J. O'Rourke, J. E. Goodman, Handbook of discrete and computational geometry, CRC press, Academic Press, (2017). https://doi.org/10.1201/9781315119601

34. A. Zomorodian, G. Carlsson, Computing persistent homology, *Discrete Comput. Geom.*, **33** (2005), 249–274. https://doi.org/10.1007/s00454-004-1146-y

35. S. S. Haykin, Neural networks and learning machines, Pearson Education, Upper Saddle River, NJ, Academic Press, 2009.

36. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, (2017). https://arxiv.org/abs/1412.6980

37. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.*, **15** (2014), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

38. R. Gonzalez-Diaz, P. Real, On the cohomology of 3d digital images, *Discret. Appl. Math.*, **147** (2005), 245–263. https://doi.org/10.1016/j.dam.2004.09.014

39. E. Diener, R. J. Larsen, S. Levine, R. A. Emmons, Intensity and frequency: dimensions underlying positive and negative affect, *J. Pers. Soc. Psychol.*, **48** (1985), 1253. https://doi.org/10.1037//0022-3514.48.5.1253

40. H. Schlosberg, Three dimensions of emotion, *Psychol. Rev.*, **61** (1954), 81. https://doi.org/10.1037/h0054570

41. D. Kamińska, T. Sapiński, A. Pelikant, Recognition of emotion intensity basing on neutral speech model, in *Man-Machine Interactions 3*, Springer, **242** (2014), 451–458. https://doi.org/10.1007/978-3-319-02309-0_49

42. S. W. Byun, S. P. Lee, Human emotion recognition based on the weighted integration method using image sequences and acoustic features, *Multimed. Tools. Appl.*, **80** (2020), 35871–35885. https://doi.org/10.1007/s11042-020-09842-1

43. M. F. H. Siddiqui, A. Y. Javaid, A multimodal facial emotion recognition framework through the fusion of speech with visible and infrared images, *Multimodal Technol. Int.*, **4** (2020), 46. https://doi.org/10.3390/mti4030046

44. C. Luna-Jimenez, D. Griol, Z. Callejas, R. Kleinlein, J. Montero, F. Fernandez-Martinez, Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning, *Sensors*, **21** (2021), 7665. https://doi.org/10.3390/s21227665

45. E. Ghaleb, J. Niehues, S. Asteriadis, Multimodal attention-mechanism for temporal emotion recognition, in *Processng of the IEEE International Conference on Image Processing*, Academic Press, (2020), 251–255. https://doi.org/10.1109/ICIP40778.2020.9191019

AIMS Press