Cognitive Computation (2024) 16:3405–3416
https://doi.org/10.1007/s12559-024-10344-7

RESEARCH

Barrier Function to Skin Elasticity in Talking Head

Iti Chaturvedi[1] · Vlad Pandelea[2] · Erik Cambria[2] · Roy Welsch[3] · Bithin Datta[1]

Received: 10 April 2024 / Accepted: 12 August 2024 / Published online: 24 August 2024

Abstract

In this paper, we target the problem of generating facial expressions from a piece of audio. This is challenging since both audio and video have inherent characteristics that are distinct from the other. Some words may have identical lip movements, and speech impediments may prevent lip-reading in some individuals. Previous approaches to generating such a talking head suffered from stiff expressions. This is because they focused only on lip movements and the facial landmarks did not contain the information flow from the audio. Hence, in this work, we employ spatio-temporal independent component analysis to accurately sync the audio with the corresponding face video. Proper word formation also requires control over the face muscles that can be captured using a barrier function. We first validated the approach on the diffusion of salt water in coastal areas using a synthetic finite element simulation. Next, we applied it to 3D facial expressions in toddlers for which training data is difficult to capture. Prior knowledge in the form of rules is specified using Fuzzy logic, and multi-objective optimization is used to collectively learn a set of rules. We observed significantly higher F-measure on three real-world problems.

Keywords Spatio-temporal · Talking head · Sentiment prediction · Finite element

Introduction

Lip-reading is the task of predicting what is being said using only visual cues [1]. This is very challenging due to the presence of homophemes such as 'p' and 'b' that have identical lip sequences [2]. Lip-reading has many applications, such as 'dictation' in a noisy environment or automated speech recognition [3]. Phonemes are the smallest unit size for speech processing instead of characters. This can result in insufficient temporal resolution and retention of spatio-temporal information. Long short-term memory (LSTM) has shown good performance on lip-reading tasks [4].

Most previous approaches rely on textual data for speaker emotion prediction. For example, common-sense databases have been created to disambiguate the meaning of words based on context [5]. However, emotions are better expressed through speech or visual gestures [6]. For example, a scowling expression is commonly labelled as anger. However, the person may just be confused or trying to concentrate. Multimodal algorithms hence aim to fuse features in audio and video with text. Talking heads can generate facial expressions for a piece of text [7]. This is particularly useful when creating dynamic content in virtual reality applications. Generating future responses in a conversation is similar to playing a game [8].

Facial expressions are also an important form of communication for young children who have still not learned a language [3, 9]. It is common for children to show contradictory expressions in a matter of a few seconds. They are unable to distinguish between complex emotions and only spontaneously collected data is available to train the model [10, 11]. Prior rules for different facial actions can help improve the accuracy of predictions. Early identification of emotional trauma in kids can prevent manifestations of autism. Par-

✉ Iti Chaturvedi
iti.chaturvedi@jcu.edu.au

Vlad Pandelea
vlad.pandelea@ntu.edu.sg

Erik Cambria
cambria@ntu.edu.sg

Roy Welsch
rwelsch@mit.edu

Bithin Datta
bithin.datta@jcu.edu.au

1 College of Science and Engineering, James Cook University, Townsville, Australia

2 College of Computing and Data Science, Nanyang Technological University, Singapore, Singapore

3 Sloan School of Management, MIT, Cambridge, MA, USA

ents frequently upload video recordings of their children on YouTube or TikTok. Large-scale analysis of these videos can help parents recognise cognitive disorders early [12]. While it is not ethical to share childhood photographs on the internet, converting them to a numeric representation will preserve their privacy [13].

Figure 1 compares 3D facial landmarks for happy and angry emotions. The top row is an expression of a toddler in the CAFE dataset. The bottom row is for an adult in the IEMOCAP dataset. We can easily distinguish between anger and happiness for adults. However, for children, it is difficult for the classifier to predict emotions from landmark positions. Hence, we can conclude that the ability to express emotions is dependent on skin elasticity. The intensity of spontaneous expression in an individual is dependent on the flexibility of facial muscles [14]. Hence, we define a barrier function to model the rate of increase in expression intensity for an individual. For example, in children, the barrier may be low, while in adults it might be higher.

Multi-objective optimisation is an ideal framework to determine co-existing facial action units responsible for an expression. Here, each facial action is modelled as a separate constraint, and the model is trained to simultaneously achieve all the objectives. For example, during happiness, we observe a smile. To capture the change in intensity of emotions in a face video, we use a spatio-temporal model where the predicted emotion is a function of facial actions spatially and over time [15]. We refer to the proposed model as multi-objective elasticity for spatio-temporal data (MES).

To study the different parameters of MES, we first generate synthetic data for the phenomena of salt water diffusion in coastal areas using the finite element method (FEM). Such a synthetic simulation is ideal to study the phenomena of spatio-temporal diffusion since all the parameters are known. For example, in [16], the authors used FEM to predict cracks due to stress in metals. Here, the total incremental external force is modelled as a product of a stiffness matrix and the vector of nodal displacements. Similarly, for a given configuration of well locations and pumping rates, FEMWATER is used to simulate the diffusion of salt in groundwater by solving the dispersion equation [17]. It lets us capture the movement of salt through water near the seaside in the presence of forces such as wind over time. The accuracy of the model can be measured as the difference between the predicted salt concentration by the neural network and that generated by FEMWATER. Further, in this paper, we show that the proposed MES algorithm has a higher accuracy in predicting salinity levels compared to baselines.

We can summarise the main contributions of this paper as follows:

1. We propose the use of spatio-temporal component analysis to model the conductivity of a medium.
2. We propose a barrier candidate function that can reach the global minima during training.
3. We propose the use of Fuzzy rules during multi-objective optimisation of design parameters.

The organisation of the paper is as follows: 'Related Work' provides a literature review of articles on spatio-temporal analysis; 'Preliminaries' describes the use of Fuzzy rules to model prior information about the system; 'Multi-objective Elasticity Framework' details the proposed approach to solve a spatio-temporal model using multi-objective evolution; finally, in 'Experiments', we evaluate our approach on facial expression generation from speech audio, classification of expressions in toddlers and prediction of salt concentration in coastal areas. Lastly, we provide conclusions in 'Conclusion'.

Related Work

The concept of 'Communication Dynamism' aims to predict future dynamics of a sentence in a conversation [18]. Text-based talking head generation is commonly used in animation movies [19]. This requires generating facial expressions for the sentiments depicted in an utterance. In this paper, we consider a spatio-temporal approach to conserve emotions in speech when generating the corresponding face. In [20], the authors studied the transfer of parameters from several unsupervised spatio-temporal models to a predictive task. Here, we initialise the LSTM using a pre-existing speech to the 68 facial landmark model.

Recently, point clouds are being used to model moving objects and pose estimation [21]. They can vary their density across an object in space and time and hence are robust to irregular sampling errors. However, their accuracy is dependent on the number and angle of 3D images captured. Point-net is a pre-trained model for 3D object classification. Similarly, Matos et al. [22] used a neural network to model the strain response of a conductive polymer that is captured at physical nodes. The input to the model is the homogenised strains, and the output is the change in resistance in three principal directions. A mesh is used to visualise the initial and deformed beam. We see that the change in resistance first increases and then decreases. It is extremely difficult to detect emotions in toddlers due to their hyperactivity. However, the accuracy significantly improved when we used a face mesh of 468 landmarks extracted from a single face image. Since the rate of stress will be different for each direction in this paper, we propose the use of a spatio-temporal component analysis to determine the significant directions of change over time.

The FEM model for a water aquifer defines the number of wells, their pumping rates, and the location on the coast. The total cost of pumping and installing the wells needs to

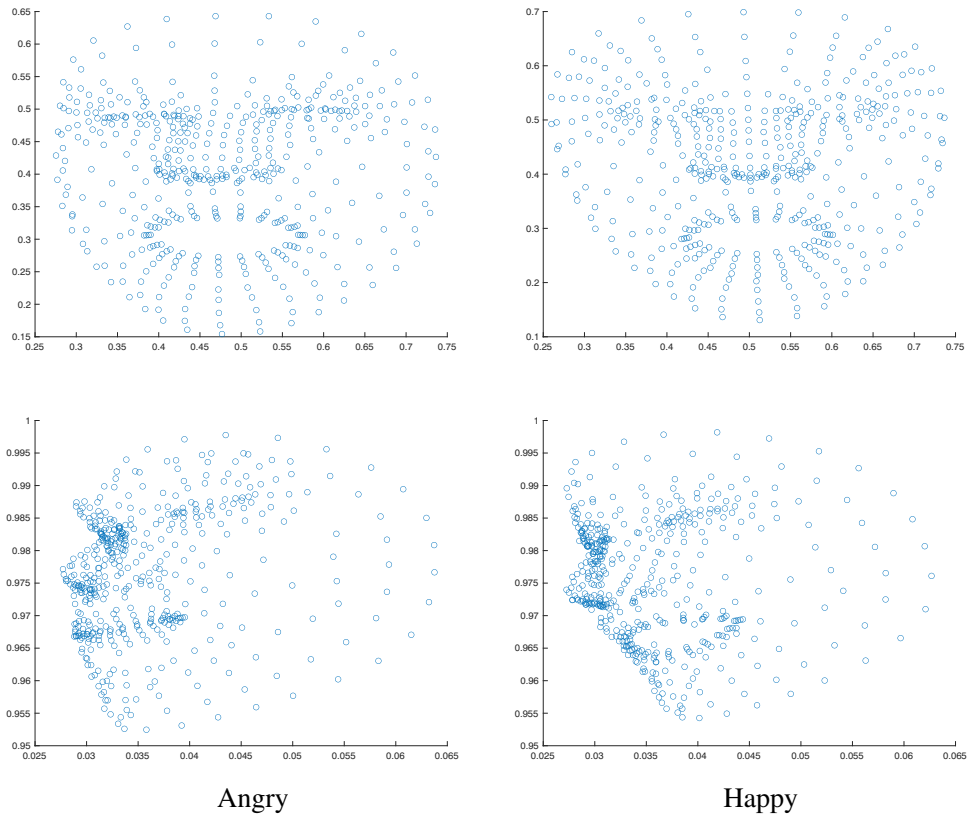

Fig. 1 3D facial landmark for happy and angry emotions. The top row is an expression from a toddler in CAFE dataset. The bottom row is for an adult in IECOMAP dataset. The ability to express emotions is dependent on the skin elasticity

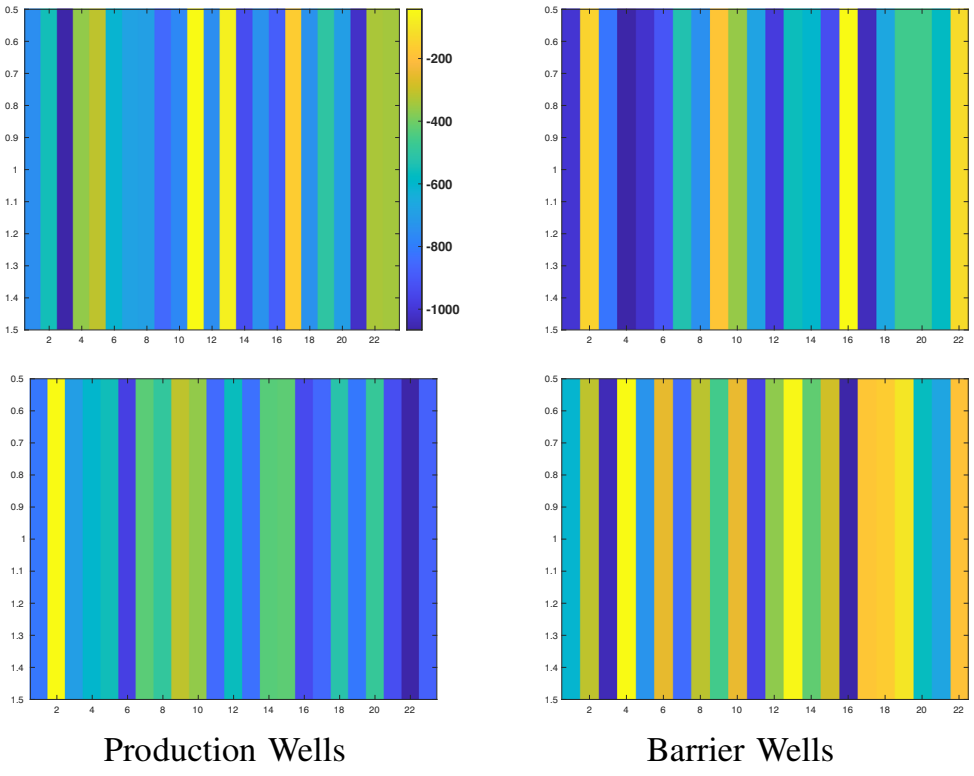

Fig. 2 Well pumping rates for production wells in agricultural lands and barrier wells at the coast. Each vertical bar represents a single well, and the colormap describes the pumping rates ranging from 0 to 1000. The horizontal axis denotes the index number of each well. The top row is a sample from salty water The bottom row is a sample from fresh water. For salty water, the barrier wells are pumping at a higher rate than in fresh water. There appears to be a lower barrier to salt diffusion

be minimised. In the event of salt water intrusion, the aquifer aims to prevent contaminated water from flowing to other parts of the aquifer. Finite element modelling was used to solve the salt water intrusion process using the numerical code FEMWATER [23]. However, each simulation can take several minutes to complete due to the complexity of the underwater terrain. The salinity level is calculated at a monitoring location at the end of the simulation.

Figure 2 illustrates well pumping rates for production wells in agricultural lands and barrier wells on the coast. Each vertical bar represents a single well and the colourmap describes the pumping rates ranging from 0 to 1000. The horizontal axis denotes the index number of each well. The top row is a sample of salty water. The bottom row is a sample of fresh water. Barrier wells maintain the level of fresh water in coastal areas. If the level falls below a threshold, then salt water intrusion will occur into agricultural land. In contrast, the production wells are used to pump fresh water into agricultural land, such as for irrigation. For salty water, the barrier wells are pumped at a higher rate than in fresh water. There appears to be a lower barrier to salt diffusion in salty water.

To allow for real-time salinity prediction, we can train a neural network with pumping rates of geographically dispersed wells as input and the salt concentration at a monitoring well is predicted as output at the end of each simulation. While a neural network is used to predict the salt concentration, we can use a multi-objective genetic algorithm (GA) to simultaneously optimise different constraints for each variable. In [24], the authors showed that search interval adaptation instead of random weight updates shows a lower root mean square error (RMSE) during backpropagation training of FEM. Similarly, multi-objective GA search was used to optimise an antilock brake system in [25]. Here, the pusher position depends on the diameter of the rod that transmits the force. The range of parameters is constrained in the optimisation. The FEM parameters have evolved over several generations, and the optimal solution is tested on the bench.

The neural network trained from FEM data can easily get stuck at a local minimum. To overcome this, we consider a barrier candidate function based on elasticity that can reach the global minima easily. Next, to study the different parameters of the multi-objective elasticity, we define a suitable objective function and specify constraints for the aquifer [26]. Using expert prior knowledge of the aquifer design, we define multiple constraints such as maximise the pumping of production wells and minimise the pumping of barrier wells [27]. We can also include constraints such as maximum and minimum values of salt in fresh water [28]. Lastly, to model the uncertainty in the parameters, we apply Fuzzy membership functions to the output from the neural network [29]. Next, we apply the approach to emotion prediction from face images. For example, during a smile, we observe multiple facial actions such as 'cheek raiser' and 'lip stretcher'.

In summary, we conclude that past approaches to generating a talking head from speech lacked modelling of underlying emotions. Furthermore, classification accuracy on facial expressions of children is very low. 3D elasticity models of the face are computationally very expensive. A neural network trained on 3D data can also get easily stuck in a local minima due to the complexity of the simulation.

Preliminaries

In this section, we provide the preliminary concepts to understand the algorithm. First, we explain spatio-temporal problems and use salt water as an example. Next, we describe the use of Fuzzy rules as prior knowledge to model the uncertainty in spatio-temporal problems.

Spatio-temporal Diffusion

Gradient descent is an optimisation approach to reach the global minima of prediction error. For a non-linear time-series model, gradient descent updates the parameters θ at time instant t of a model using the following equation:

$$\begin{aligned} \theta(t+1) &= \theta(t) + \Delta\theta(t) \\ \Delta\theta(t) &= \lambda \frac{\partial e(t)}{\partial \theta(t)} \end{aligned} \tag{1}$$

where $e(t)$ is the error in prediction and λ is the rate of learning.

Pretrained models for landmark detection have inherent uncertainty depending on the shape and size of facial features. The presence of facial action units is responsible for different facial expressions. For example, 'Lip corner puller' is an action used in 'contempt'. However, when used in combination with 'cheek raiser', it results in a smile. A neural network, in contrast, will extract a dictionary of features from images and then merge them using a layered model. Human beings look for variations of facial features over time in order to distinguish such emotions. Hence, there is a need to use a spatio-temporal model that can detect changes over time. Here, in addition to the error over time, we can also consider the spatial error across samples. Hence, we modify the gradient descent using the following:

$$\Delta\theta(t) = \lambda \frac{\partial e(t)}{\partial \theta(t)} + \gamma \frac{\partial e(z)}{\partial \theta(t)} \tag{2}$$

where *e(z)* is the spatial error in the position of samples. For the case of salt water, $\Delta\theta$ is the change in salt concentration

over time in *moles/litre*^{3}, and the gradient of salt concentration at the pumping wells is given by λ and the hydraulic conductivity or by γ.

Fuzzy Rules for Modelling Prior

Prior knowledge in the form of action units can be formulated as rules on the position of landmarks. Rules will restrict certain combinations of features and hence tremendously reduce the dimensionality of the problem. Here, we use a baseline decision tree to determine significant rules for each emotion. Next, we can specify these rules in a fuzzy neural network for fusion of features. A fuzzy logic classifier has membership functions that can range from partial positive to partial negative. Each membership function has a set of inputs and outputs. In this way, each input feature now has a membership value in a particular function that is not as rigid as in a conventional neural network.

Figure 3 illustrates fuzzy rules for predicting facial expressions from landmarks. The first membership function is used in rules where a landmark is lower or higher than a value. Here, we followed an approach similar to decision trees where the value of a landmark determines if we move to the left or right sub-branch in the tree. Here, the data is normalised in the range of -1 and $+1$. A value less than 0 is denoted by 'Low', and a value greater than 0 is 'High'. We have one input membership function for each landmark. For example, 68 landmarks would result in 136 inputs corresponding to x and y coordinates. To reduce the computational complexity, we extract a latent representation of landmarks with a lower dimension using a neural network. The second membership function is used for the output emotion. Here, we consider the intensity of the emotion so that a very strong expression is denoted by 'Happy' and a week expression is denoted by 'Low'.

In our experiments, we only considered two output emotions namely 'Happy' and 'Angry'. As shown in Fig. 3, each rule takes the form of 'if else then' statements. For example, 'If ln2 is high and ln4 is high and ln3 is low then cl1 is happy'. Here, 'ln' denotes a landmark index, and 'cl' denotes an output emotion. To initialise the model, we extracted a few rules using decision trees. Next, we used evolutionary optimisation to learn additional rules and maximise the accuracy on the training data. When new training data is acquired, the model can learn new rules over existing membership functions without altering existing rules. We can also add new membership functions if the existing ones are completely populated with rules. To optimise for speed, the algorithm lets you specify the maximum number of rules learned. The output layer will combine the outputs from all the membership functions to predict the polarity of the face using (3).

Let us formulate each rule as a membership function $m_k(x)$ where x is an input vector of n features and K is the number of membership functions. Then, the change in parameters for each membership function $\Delta\theta_k$ is a weighted average over all input features [30] as follows:

$$\Delta\theta_k = \prod_{i=1}^{n} m_k(x_i) \Big/ \sum_{k=1}^{K} \prod_{i=1}^{n} m_k(x_i) \tag{3}$$

Multi-objective Elasticity Framework

In this section, we describe our proposed approach to solving spatio-temporal problems. We first explain the spatial component analysis method which includes a barrier function for improved convergence. Next, we show how additional constraints can be added into the prediction using multi-objective evolutionary optimisation.

Barrier Function

The elasticity model for facial landmarks can be illustrated in Fig. 4. Here, we consider two landmarks at the corner of the eye and the mouth with a vertical distance of L. The horizontal displacement at time t is given by the function $e(z)$; hence, we can constrain the movement using $|e(z)| \leq c$ where c is the desired maximum movement intensity to detect an emotion. The controlled force applied by an individual to generate an emotion is given by $\Delta\theta(t)$.

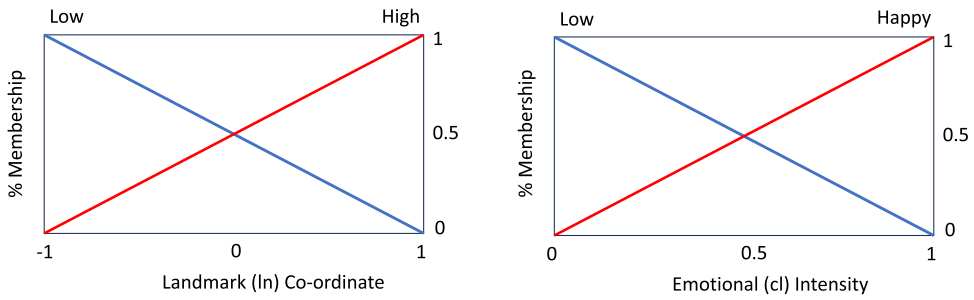

Fig. 3 Fuzzy rules for predicting facial expressions from landmarks. The first membership function is used in rules where a landmark is lower or higher than a value. The second membership function is used for the output emotion

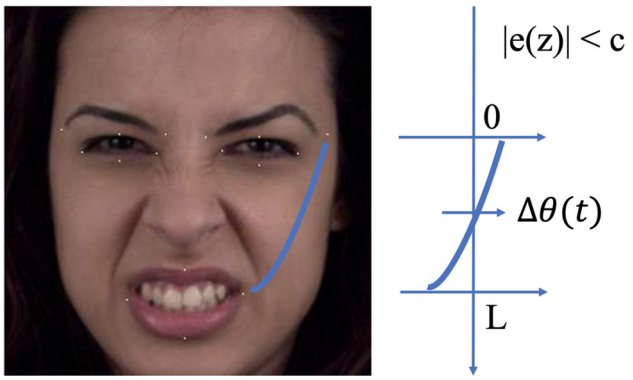

Fig. 4 Elasticity model for a face. For any two selected landmarks, the vertical distance is L and the horizontal displacement over time $\Delta\theta(t)$ will depend on the emotion

Then, the boundary condition for the rate of diffusion of emotion can be given by the following equation [31]:

$$\Delta\theta(t) = \lambda \frac{\partial^2 e(t)}{\partial\theta(t)} - \gamma \frac{\partial e(z)}{\partial\theta(t)} \tag{4}$$

where as previously explained in (2), γ is the elasticity of the medium and λ is the rate of change of salt concentration. This implies that for a displacement of $e(z)$, a person has to apply a force $\Delta\theta(t)$ and overcome the negative acceleration due to gravity.

The following barrier term β when included into applied force $\Delta\theta(t)$ will ensure that the gradient descent will reach a global minima:

$$\Delta\theta(t) = \beta + \mu\left[\frac{\partial e(t)}{\partial\theta(t)} + \frac{\partial e(z)}{\partial\theta(t)}\right] + \lambda\frac{\partial e(t)}{\partial\theta(t)} - \gamma\frac{\partial e(z)}{\partial\theta(t)} \tag{5}$$

$$\beta = -\left\{\frac{\gamma e(z)\frac{\partial e(t)}{\partial\theta(t)}\left[\frac{\partial e(t)}{\partial\theta(t)} + \frac{\partial e(z)}{\partial\theta(t)}\right]}{c^2 - e^2(z)} - \mu\left[\lambda\frac{\partial e(t)}{\partial\theta(t)} + \gamma\frac{\partial e(z)}{\partial\theta(t)}\right]\right\} \Big/ \log\frac{2c^2}{c^2 - e^2(z)}$$

where $\mu > 0$ is the gain of the controller that ensures that enforces the constraint $|e(z)| \leq c$ on movement through the medium. Detailed proof is described in [31].

Each training sample can be decomposed into additive components over time using a method called independent component analysis (ICA). In the real world, a signal may diffuse spatially and affect other signals. Spatio-temporal ICA can model such a phenomenon as it decomposes a signal into sub-components that are independent of each other over time as well as space [32]. We consider the previously defined training data x with n features and a sequence of p time samples. Equation (6) shows that we can now define $\Delta\theta(t)$ is a product of o spatial s_z and temporal s_t components. We also define the symmetric mixing matrices a_z and a_t of dimension $o \times o$.

$$\Delta\theta(t) = \frac{\partial e(t)}{\partial\theta(t)} \times \frac{\partial e(z)}{\partial\theta(t)} \text{ where } \frac{\partial e(t)}{\partial\theta(t)} = s_t \times a_t \text{ and } \frac{\partial e(z)}{\partial\theta(t)} = s_z \times a_z \tag{6}$$

Finally, we can determine the new transformed dataset as follows:

$$\Delta\hat{\theta}(t) = (\Delta\theta(t) + s_t \times \beta \times s_z) \times 0.5 \tag{7}$$

where β is a diagonal barrier matrix as defined previously in (5). Figure 5 illustrates training samples from three datasets with and without the barrier control. We can see that by using the elasticity constraint we can significantly reduce the variance across features resulting in better convergence. We consider (a) pumping rates of different wells in an FEM simulation, (b) position of different 3D landmarks in a face image, and (c) Mel coefficients extracted from a speech sample over four video frames. In practice, we consider a weight sum of the original sample x and transformed sample x_{new} to train the model.

Algorithm 1 Predicting emotions using MES.

```
Input : Input features at time point t : x and its class label
Output : Predicted class label : Δθ×x
% Spatio-temporal ICA with barrier function
   Computer barrier potential β using (5)
   Compute transformed x_new using (7)
% Feature Selection
   Normalise landmarks in the range of [−1, 1]
   Train a neural network using x_new
   Features are activations in last layer
% Fuzzy Neural Network
   Create input membership functions for each feature
   Create output membership functions for each emotion
   Learn Fuzzy Logic rules for each emotion
   Save the trained emotion classifier
% Multi-objective optimisation
   Minimise fuzzy logic output in (3)
   Maximise constraints on input features
   Minimise constraints on input features
```

Multi-objective Optimisation

The entire MES algorithm can be defined as follows using expressions as an example: The training data is a sequence of images for the same facial expression over time. Next, we use spatio-temporal ICA with a barrier function to pre-process the data and extract significant components. The landmarks are extracted using a pre-trained model, and a decision tree classifier is used to determine significant rules for the position of landmarks in a particular emotion. These prior rules are used to initialise the fuzzy logic classifier using 'If then Else' statements. Lastly, we use multiple objectives to define

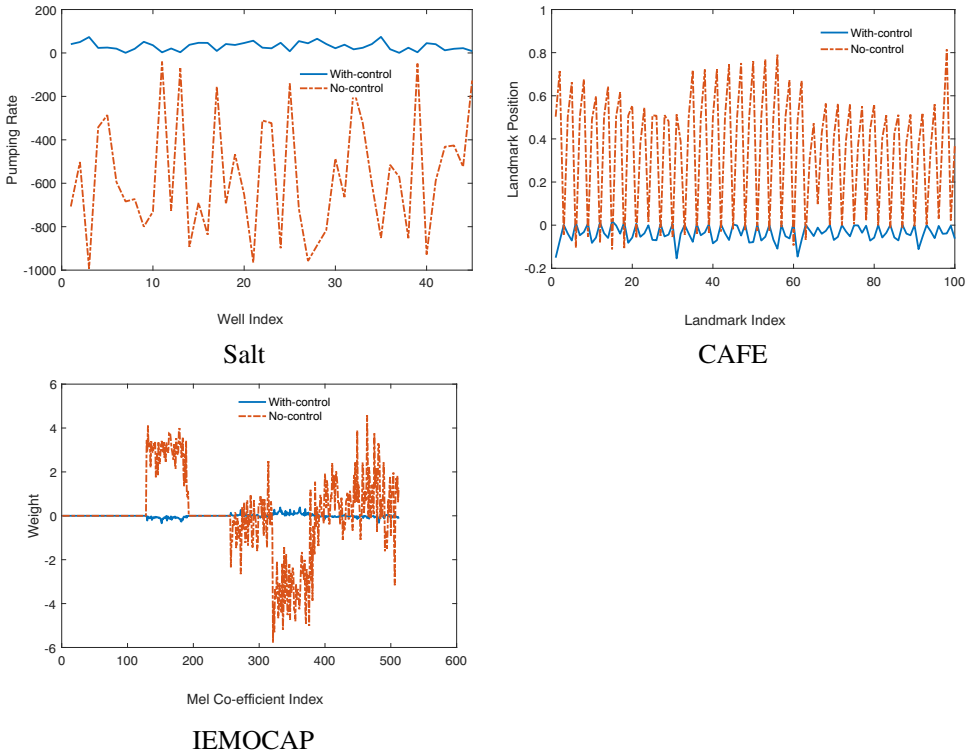

Fig. 5 Training samples in three datasets with and without barrier control. We can see that by using the elasticity constraint we can significantly reduce the variance across features resulting in better convergence. **a** Pumping rates of different wells in an FEM simulation. **b** Position of different 3D landmarks in a face image. **c** Mel coefficients extracted from a speech sample over four video frames

additional constraints such as maximisation or minimisation of a certain facial action. The Pareto front determines possible solutions that simultaneously optimise the different constraints.

Algorithm 1 explains the complete framework for predicting emotions using the proposed model. The input consists of facial landmarks for training images with known emotion labels. We first extract significant components from the data using a spatio-temporal ICA with a barrier function to model the skin elasticity. Next, we perform a feature selection to reduce the number of landmarks using a neural network. The activations in the penultimate layer are used to train a decision tree classifier and extract rules for the presence of different emotions. These rules are used to define the membership functions in a Fuzzy logic-based neural network. We consider a binary classifier for each emotion against the neutral emotion. Lastly, we use our prior knowledge of facial action units to define multiple constraints on the model. The multi-objective model is optimised using evolution, and the fuzzy classifier is used as the fitness function. As an example, for salt water diffusion, the aim of a salt water management system is to (i) maximise the pumping of fresh water at production wells and (ii) minimise the extraction of fresh water at barrier wells. Increased pumping of production wells will hence require a corresponding increase in pumping of barrier wells to maintain the salinity levels of water below a threshold in agricultural land.

The set of feasible solutions can be given by the following:

$$O = \left\{ x | \text{minimize } \Delta\theta_k \times x, \text{minimize } \sum_{i=1}^{n1} x_i, \text{maximize } \sum_{i=1}^{n2} x_i \right\} \quad (8)$$

where $n1$ is the number of barrier wells on the shore and $n2$ is the number of fresh water pumping wells in agricultural lands. Assuming the label for fresh water is 1 and salty water is 2, we aim to mimize the class label $\Delta\theta \times x$ using Fuzzy logic (3). The Pareto front solutions $\Delta\theta$ from the evolutionary model are used to predict the class label for test samples.

Experiments

A neutral expression corresponds to a static face image. On the other hand, a facial expression is generated elastically by the movement of face muscles. In poor illumination or low image resolution, the dynamic information due to movement is more useful in classifying a face image. This is because movement captures a three-dimensional view of the

face. There appear to be two independent cortical areas in the human brain for remembering the static identity features and dynamic social features of a face [33]. The homogenous nature of a face image also requires a model that has high sensitivity in low-contrast vision.

Validation of the proposed MES (available on GitHub[1]) is done on three real-world datasets: (1) salt water diffusion prediction, (2) talking head generation from a piece of audio, (3) classifying facial expressions for children. The first dataset was synthetically generated using an FEM software. The other two datasets have been collected from human subjects.

Parameters

To model the elasticity of the medium, we apply ICA with a barrier function as described in 'Barrier Function'. We consider ten spatial and temporal components for the ICA. Following previous authors [31], we set the gain μ to 0.1 and the maximum movement intensity c to 5, the rate of change of the medium λ is set to 100, and the elasticity of the medium γ is set to 1. It is difficult to define fuzzy rules for a large number of input features; hence, we perform a dimensionality reduction using a NN. The NN is trained to predict the emotions from the landmarks and has a layer of five hidden neurons. The activations at these five neurons are used as training features for the fuzzy logic classifier. We first constructed a decision tree to determine eight starting rules for the fuzzy classifier as explained in 'Fuzzy Rules for Modelling Prior'. We also allow the fuzzy classifier to learn up to 40 rules using genetic algorithm-based optimization with a crossover rate of 0.2. A low crossover rate will ensure that the model does not get stuck in a local minimum. We had to use the parallel computing toolbox in MATLAB to increase the speed of computation.

Salt Water Diffusion

We consider a coastal landscape where 23 barrier wells are installed very close to the sea lines and 23 production wells were installed close to fresh water. Using FEMWATER, we can define the landscape in a coastal region and place the production and pumping wells at desired locations. Next, we specify the pumping rates of the 46 wells using random Latin Hypercube sampling. The experiment was repeated 1000 times for a single configuration. We repeat this process several times for different locations of wells and conductivity ranging from 40 to 240 *moles/day*.

Table 1 compares the F-measure of the proposed algorithm with baselines for the binary task of predicting 'fresh' or 'salty' water at the monitoring well. We first train the model on 70% of 'Batch1' collected from a single FEM simulation. We then test it on the remaining 30% of 'Batch1'. We can see that the simple neural network (NN) has the F-measure of 91%. The proposed model given by FuzzyB also has a very similar F-measure. However, if we do not use the barrier constraint, then the model denoted by fuzzy has a 30% lower F-measure of 59%.

Table 1 Comparison of F-measure of baseline classifiers on salt water intrusion

Test data	Method	Fresh	Salty	Total
Batch1	NN	**0.93**	**0.88**	**0.91**
Batch1	Tree	0.79	0.69	0.74
Batch1	Fuzzy	0.68	0.5	0.59
Batch1	FuzzyB	0.92	0.86	0.89
Batch2	NN	**0.77**	0.6	0.69
Batch2	Tree	0.63	0.48	0.56
Batch2	FuzzyB	0.76	**0.66**	**0.71**

When trained on Batch1 and tested on Batch2, we see an improvement of 5% on salty class
The highest F-measure for each dataset and for each water level is highlighted as bold

Next, we tested the trained model on 'Batch2' data collected from a separate FEM simulation with different starting parameters. Here, the proposed method FuzzyB has a slightly higher F-measure of 71% compared to neural networks (NN). The improvement is 6% on the 'salty' class. The F-measure is over 15% higher than the baseline tree classifier that was trained on 'Batch1'. Hence, we can conclude that the proposed approach has lesser overfitting and can show better accuracy on new datasets.

Lastly, we include the constraints that we wish to maximise the pumping of fresh water wells near agricultural land and minimise the pumping of barrier wells in coastal areas. In order to model the fuzzy classifier as an objective, we mimise the predicted label from FuzzyB. This is because we have set 'fresh' to 1 and 'salty' to 2 in the training data. Table 2 compares the F-measure of multi-objective (MO) and the proposed MES on salt water diffusion. We can see that MES has a higher F-measure than MO when considering the constraints. MO has a higher objective value for maximising the pumping of fresh water wells given by 2.98; however, MES has a lower objective value of 0.19 for minimising the pumping of barrier wells.

Talking Head: Face Audio and Video

Next, we apply the proposed approach to the prediction of facial landmarks from speech. This is a necessary component of models that can generate a talking video from a piece of text. To train the model, we used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) which contains video recordings of conversations between two speakers [6]. There

[1] http://github.com/ichaturvedi/multi-objective-elasticity

Table 2 Comparison of F-measure of multi-objective (MO) and the proposed MES on salt water diffusion and talking head dataset

	Method	Fresh	Salty	Total	Obj1 (max)	Obj2 (min)
Batch1	MO	0.32	0.15	0.24	**2.98**	0.2
Batch1	MES	0.42	0.23	**0.33**	1.96	**0.19**
		Angry	Happy	Total	Obj1 (max)	Obj2 (min)
Happy	MO	0.01	0.62	0.32	2.13	0.28
Happy	MES	0.68	0.1	**0.39**	**2.2**	**0.0034**
Angry	MO	0.001	0.62	0.31	1.4	**0.23**
Angry	MES	0.1	0.62	**0.36**	**1.91**	0.39

We can see that MES has a higher F-measure than MO when considering the constraints. MES also achieves a lower value on the mimisation and a higher value on the maximisation constraint compared to MO
The highest total F-measure for each dataset among the two methods is in bold. Similarly, the bold value for Obj 1 is the higher of two and for Obj 2 is the lower of the two for each dataset

are a total of five female and five male actors and 12 h of audio-visual data. Each video is segmented into utterances that has an emotional label such as happy or angry. The database was designed to capture the relationship between gestures and speech hence most of the faces are captured sideways. Theatre scripts were selected with the requirement that the play conveys target emotions. Subjects were asked to memorise and rehearse the scripts. Here, we only consider a subset of 502 utterances that have been labelled as happy or angry. We extracted 128 Mel coefficients for each frame in a video and used a window size of four frames resulting in an input vector of 512 features. We used pre-trained weights for speech-to-landmark prediction to initialise the LSTM [34].

We test the model on an additional Lip Reading Sentences (LRS2) dataset consisting of thousands of spoken sentences from BBC television recorded between 2010 and 2016 [1]. Video shot boundaries were determined by comparing colour histograms across consecutive frames. Forced alignment was done between the video, audio, and subtitles for each shot. Lastly, sentences were determined using punctuations such as full stops and question marks in the subtitles. Each sentence is restricted to 100 characters in length. Figure 6 shows the Mel

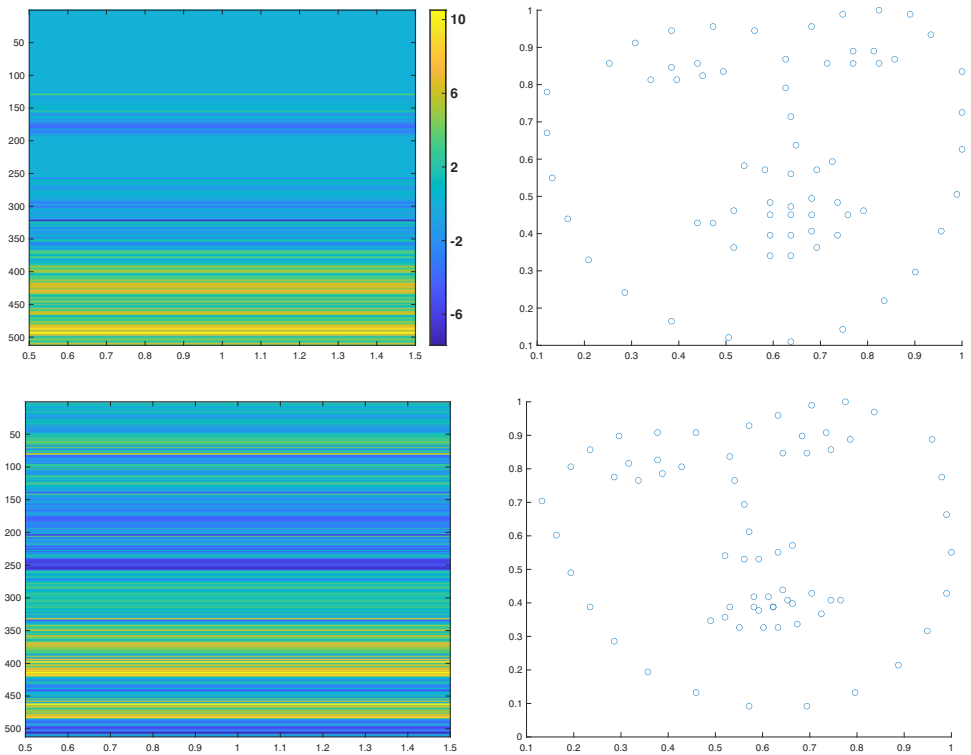

Fig. 6 Mel spectrum for speech and the corresponding facial landmark. The top row is a sample for happy emotion. The bottom row is a sample from angry emotion in LRS2 dataset. The angry emotion has lower values of Mel coefficients. The oral cavity will change shape and hence the barrier to sound depending on the emotion

spectrum for speech and the corresponding facial landmark. The top row is a sample for happy emotion. The bottom row is a sample from angry emotion in LRS2 dataset. The angry emotion has lower values of Mel coefficients. The oral cavity will change shape and hence the barrier to sound depending on the emotion.

Here, we first predict the facial landmarks from speech audio using LSTM, and then we predict the emotion label of the predicted face as 'angry' or 'happy'. Table 3 compares the F-measure of the proposed algorithm with baselines. We transform the speech input using the barrier function and train the model denoted by LstmB. We can see it has a much higher F-measure of 51% compared to the baseline Lstm of 40%. Next, we train the fuzzy classifier with the landmarks predicted by LstmB denoted as LstmB-Fuzzy. We can see that when tested on a new dataset LSR2, the F-measure on 'angry' class is much higher than baselines. This confirms that the fuzzy model is better suited to real-world datasets.

Lastly, we introduce some constraints using facial action units. We find that 'anger' emotion results in 'lip puller' and 'open eyes.' On the other hand, 'happy' emotion has the action units 'lip stretcher' and 'closed eyes'. The third objective is to mimise the label of the FuzzyB model so that it is either 'anger' or 'happy' based on the constraints. Table 2 shows the F-measure of MES is higher than MO on both emotions. It also achieves a lower minimisation and higher maximisation on the constraints specified.

Child Facial Expressions

Lastly, we evaluate the model of facial landmarks for different emotions in children [35, 36]. For each emotion, such as 'happy' or 'surprise', we train a binary classifier with respect to the neutral expression. The Child Affective Facial Expression (CAFE) dataset has 90 female and 64 male children.

Table 3 Comparison of F-measure of baseline speech-to-landmark sentiment classifiers

Test Data	Method	Angry	Happy	Total
Iemocap	Lstm	**0.7**	0.1	0.4
Iemocap	LstmB	0.62	0.4	**0.51**
Iemocap	NN	0.01	**0.8**	0.41
Iemocap	Tree	0.01	0.8	0.41
Iemocap	LstmB-Fuzzy	0.5	0.5	0.5
Lrs2	NN	0	**0.89**	**0.45**
Lrs2	Tree	0.01	0.76	0.39
Lrs2	LstmB-Fuzzy	**0.4**	0.26	0.33

When trained on Iemocap and tested on Lrs2, we see an improvement of 37% on the Angry emotion
The highest F-measure for each dataset and for each emotion is highlighted as bold

Table 4 Comparison of F-measure of baseline sentiment classifiers on facial landmarks

Test data	Method	Angry	Happy	Total
Cafe	NN	**0.85**	**0.79**	**0.82**
Cafe	Tree	0.67	0.66	0.67
Cafe	Fuzzy	0.75	0.74	0.75
Cafe	FuzzyB	0.8	0.74	0.77
Iemocap	NN	**0.69**	0	0.35
Iemocap	Tree	0.69	0	0.35
Iemocap	FuzzyB	0.15	**0.55**	**0.35**

When trained on Cafe and tested on Iemocap, we see an improvement of 55% on the happy emotion
The highest F-measure for each dataset and for each emotion is highlighted as bold

Photographs are captured from children in the age group of 2 to 8 years. Unsuccessful poses were removed from the dataset. The FaceMesh[2] by MediaPipe model detects 468 key face landmarks in real time. For each image, we extract 468 landmark points using FaceMesh. These landmarks define the location of the eyes, nose, mouth, and cheeks. We refer to FACS[3] (Facial Action Coding System) to determine the action units in different emotions. For example, when a person is happy, then the mouth area will be maximised. We use the FACS to determine multiple objectives for each emotion.

Here, we consider the subset of 420 images for 'happy' (215) and 'angry' emotions (205). Each landmark is defined by the X, Y, and Z coordinate, resulting in 1434 input features. Table 4 compares the F-measure of the proposed algorithm with baselines for the binary task of predicting 'angry' or 'happy' expressions from face landmarks. We first train the model on 70% of CAFE data images and test on the remaining 30%. We can see that the NN has a F-measure of 82%. The proposed model given by FuzzyB also has a very similar F-measure. However, compared to a decision tree classifier, the improvement is over 10%.

Next, we tested the trained model on IEMOCAP dataset described in the previous section. This balanced dataset contains 502 images of 'Happy' and 'Angry' face images of speakers. Here the proposed method, FuzzyB has the best result in the Happy class with a 55% F-measure. We can see that the baselines such as NN and tree are unable to classify a new dataset, suggesting that they are overfit to the training data. Hence, we can conclude that fuzzy rules can adapt to new datasets. It is currently difficult to map the 468 3D landmarks to 2D facial actions; hence, we did not report multi-objective results on this dataset.

[2] https://mmla.gse.harvard.edu/tools/face-mesh/

[3] https://imotions.com/blog/learning/research-fundamentals/facial-action-coding-system/

Conclusion

In this paper, we propose a spatio-temporal model that can accurately predict facial expressions from landmark points. We also apply our method to study the problem of salt water diffusion in coastal areas. In order to capture variation spatially and over time, we consider a novel feature selection approach to the dataset that considers the muscle barriers to increase in emotional intensity. Next, we train a multi-objective evolutionary model that is able to simultaneously maximise or minimise multiple constraints in the system. The error function is determined using a fuzzy neural network where prior rules are extracted using a decision tree. We show that the proposed approach has an improvement in the range of 5–30% compared to baselines. We also observed a better minimisation or maximisation of objectives in a constrained multi-objective setting.

Acknowledgements This work is partially supported by the College of Science and Engineering, James Cook University, Australia. Open access publishing facilitated by James Cook University, as part of the Springer - James Cook University agreement via the Council of Australian University Librarians. This work is also partially supported by the School of Computer Science and Engineering, Nanyang Technological University, Singapore.

Author Contribution Dr. Iti Chaturvedi conceptualised the idea and wrote the original draft of the manuscript. Vlad Pandelea helped with the validation of experiments and reviewed the manuscript. Dr. Bithin Datta supervised the synthetic experiments using FEMWATER. Prof. Erik Cambria and Prof. Roy Welsch supervised the work and helped with revising and editing the manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data Availability Data analysis codes are available on GitHub: http://github.com/ichaturvedi/multi-objective-elasticity.

Declarations

Conflict of Interest The authors declare no competing interests.

References

1. Chung JS, Senior A, Vinyals O, Zisserman A. Lip reading sentences in the wild. In: CVPR. 2017. pp. 3444–3453.
2. Stappen L, Baird A, Cambria E, Schuller BW. Sentiment analysis and topic recognition in video transcriptions. IEEE Intell Syst. 2021;36(02):88–95.
3. Chaturvedi I, Noel T, Satapathy R. Speech emotion recognition using audio matching. Electronics. 2022;11(23).
4. Lu Y, Chai J, Cao X. Live speech portraits: real-time photorealistic talking-head animation. ACM Trans Graph. 2021;40(6).
5. Cambria E, Zhang X, Mao R, Chen M, Kwok K. Senticnet 8: Fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing. In: International Conference on Human-Computer Interaction. 2024.
6. Busso C, Bulut M, Lee C-C, Kazemzadeh EA, Provost EM, Kim S, Chang JN, Lee S, Narayanan SS. Iemocap: interactive emotional dyadic motion capture database. Lang Resour Eval. 2008;42:335–59.
7. Cambria E, Schuller B, Liu B, Wang H, Havasi C. Statistical approaches to concept-level sentiment analysis. IEEE Intell Syst. 2013;28(3):6-9
8. Amin MM, Cambria E, Schuller BW, Cambria E. Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of chatGPT. IEEE Intell Syst. 2023;38(2):15–23.
9. Chen Q, Ragusa E, Chaturvedi I, Cambria E, Zunino R. Text-image sentiment analysis. Lect Notes Comput Sci. 2023;13397:169–80.
10. Cambria E, Howard N, Hsu J, Hussain A. Sentic blending: scalable multimodal fusion for the continuous interpretation of semantics and sentics. In: CIHLI. 2013. pp. 108–117.
11. Valdivia A, Luzón MV, Cambria E, Herrera F. Consensus vote models for detecting and filtering neutrality in sentiment analysis. Inf Fusion. 2018;44:126–35.
12. Cambria E, Wang H, White B. Guest editorial: big social data analysis. Knowl-Based Syst. 2014;1-2
13. Cambria E, Mao R, Chen M, Wang Z, Ho S-B, Murugesan S. Seven pillars for the future of artificial intelligence. IEEE Intell Syst. 2023;38(6):62–9.
14. Cambria E, Mazzocco T, Hussain A, Eckl C. Sentic medoids: organizing affective common sense knowledge in a multi-dimensional vector space. In: ISNN. 2011. pp. 601–610.
15. Chaturvedi I, Satapathy R, Lynch C, Cambria E. Predicting word vectors for microtext. Exp Syst. 2024;41(8):e13589
16. Hambli R. Statistical damage analysis of extrusion processes using finite element method and neural networks simulation. Finite Elem Anal Des. 2009;45(10):640–9.
17. Roy D, Datta B. Genetic algorithm tuned fuzzy inference system to evolve optimal groundwater extraction strategies to control saltwater intrusion in multi-layered coastal aquifers under parameter uncertainty. Model Earth Syst Environ. 2017;3:1707–25.
18. Nakano YI, Okamoto M, Kawahara D, Li Q, Nishida T. Converting text into agent animations: assigning gestures to text. In: NAACL. 2004. pp. 153–156.
19. Cheng L, Wang S, Zhang Z, Ding Y, Zheng Y, Yu X, Fan C. Write-a-speaker: text-based emotional and rhythmic talking-head generation. In: AAAI. 2021.
20. Yao Z, Wang Y, Long M, Wang J. Unsupervised transfer learning for spatiotemporal predictive networks. In: ICML vol. 119. 2020. pp. 10778–10788.
21. Charles RQ, Su H, Kaichun M, Guibas LJ. Pointnet: Deep learning on point sets for 3D classification and segmentation. In: CVPR. 2017. pp. 77–85.

22. Matos MAS, Pinho ST, Tagarielli VL. Application of machine learning to predict the multiaxial strain-sensing response of CNT-polymer composites. Carbon. 2019;146:265–75.
23. Lal A, Datta B. Modelling saltwater intrusion processes and development of a multi-objective strategy for management of coastal aquifers utilizing planned artificial freshwater recharge. Model Earth Syst Environ. 2018;4:111–26.
24. Arndt O, Barth T, Freisleben B, Grauer M. Approximating a finite element model by neural network prediction for facility optimization in groundwater engineering. Eur J Oper Res. 2005;166(3):769–81.
25. Lostado R, Villanueva Roldán P, Fernandez Martinez R, Mac Donald BJ. Design and optimization of an electromagnetic servo braking system combining finite element analysis and weight-based multi-objective genetic algorithms. J Mech Sci Technol. 2016;30(8):3591–605.
26. Sawyer CS, Ahlfeld DP, King AJ. Groundwater remediation design using a three-dimensional simulation model and mixed-integer programming. Water Resour Res. 1995;31(5):1373–85.
27. Rajanayaka C, Samarasinghe S, Kulasiri D. Solving the inverse problem in stochastic groundwater modelling with artificial neural networks. iEMSs. 2002;2:154–9.
28. Aly AH, Peralta RC. Optimal design of aquifer cleanup systems under uncertainty using a neural network and a genetic algorithm. Water Resour Res. 1999;35(8):2523–32.
29. Chaturvedi I, Su CL, Welsch RE. Fuzzy aggregated topology evolution for cognitive multi-tasks. Cogn Comput. 2021;13(1):96–107.
30. Rajapakse JC, Chaturvedi I. Stability of inferring gene regulatory structure with dynamic Bayesian networks. Lect Notes Comput Sci. 7036 LNBI. 2011;237–246.
31. He W, Zhang S, Ge SS. Adaptive control of a flexible crane system with the boundary output constraint. IEEE Trans Ind Electron. 2014;61(8):4126–33.
32. Stone JV, Porrill J, Porter NR, Wilkinson ID. Spatiotemporal independent component analysis of event-related FMRI data using skewed probability density functions. NeuroImage. 2002;15(2):407–21.
33. O'Toole AJ, Roark DA, Abdi H. Recognizing moving faces: a psychological and neural synthesis. Trends Cogn Sci. 2002;6(6):261–6.
34. Eskimez SE, Maddox RK, Xu C, Duan Z. Generating talking face landmarks from speech. In: Latent Variable Analysis and Signal Separation. 2018.
35. LoBue V, Thrasher C. The child affective facial expression (CAFE) set: validity and reliability from untrained adults. Front Psychol. 2015;5.
36. LoBue V. The Child Affective Facial Expression (CAFE) set. Databrary. 2014:https://doi.org/10.17910/B7301K