Signal, Image and Video Processing
https://doi.org/10.1007/s11760-021-01943-0

ORIGINAL PAPER

Landmark calibration for facial expressions and fish classification

Iti Chaturvedi[1] · Qian Chen[2] · Erik Cambria[2] · Desmond McConnell[1]

Received: 14 October 2020 / Revised: 18 May 2021 / Accepted: 18 May 2021
© Crown 2021

Abstract
This paper considers the automatic labeling of emotions in face images found on social media. Facial landmarks are commonly used to classify the emotions from a face image. However, it is difficult to accurately segment landmarks for some faces and for subtle emotions. Previous authors used a Gaussian prior for the refinement of landmarks, but their model often gets stuck in a local minima. Instead, the calibration of the landmarks with respect to the known emotion class label using principal component analysis is proposed in this paper. Next, the face image is generated from the landmarks using an image translation model. The proposed model is evaluated on the classification of facial expressions and also for fish identification underwater and outperforms baselines in accuracy by over 20%.

Keywords Facial expressions · Adversarial training · Singular value decomposition · Fish segmentation

1 Introduction

Emotion recognition from social media data such as YouTube or Facebook allows us to understand large-scale opinions about a product or an event. Automatic facial expression prediction can also be used to monitor patients or user experience in a game. The same emotion model may be used on people from different linguistic backgrounds [1,2]. A few lexicons exist for identifying facial actions in face images. However, machine learning methods such as deep learning or belief networks can be trained from annotated face images. Generative adversarial networks (GANs) are ideal for predicting the emotional state of a patient or consumer in real time by transforming any input face using pixel level manipulations [3,4].

Facial expression recognition aims to predict the emotional state of a person from their face image. Participants are asked to watch video clips that elicit spontaneous emotions such as 'happiness' or 'sadness' [5]. The facial expression of the participant at the end of the video clip is captured using a high-definition camera. However, the posture and the lighting conditions are fixed for all participants. In this paper, this challenge is overcome via calibration of the face landmarks with respect to a gold-standard face for a particular emotion. Facial expressions are made up of two or more component 'facial action units.' For example, smile is made of 'cheek raiser' and 'lip stretcher.' This allows understanding the detailed physical changes in a face and study of new relationships between facial movements and the internal state such as 'stress' or 'fatigue' [6].

Most previous models assume that each face image can be classified to a single emotion. This is often not true as a person can experience a complex combination of several emotions simultaneously. Multi-label learning can assign more than one emotion class to the same face image. In [7], the authors tackle this problem by computing several local expression functions to approximate the global function, which requires a lot of training data. Conversely, generating the underlying emotion in an unsupervised manner requires minimal training. Spatiotemporal models such as landmarks, bag-of-words or Gabor transforms are popular for automatic identification of facial action units in different emotions. They are also effective in detecting micro-facial expressions that only last a few seconds [8]. In [9], the authors consider a combined input of image, landmarks and face parts to train a spatiotemporal model. Deep spatiotemporal models have also shown higher accuracy on facial expression benchmarks [10]. Pose-invariant dictionary is another challenge for facial expression prediction. It is observed that the accuracy improves with the number of dictionary features [11].

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

[1] CSE, James Cook University, Townsville, Queensland, Australia

[2] School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

Published online: 16 June 2021

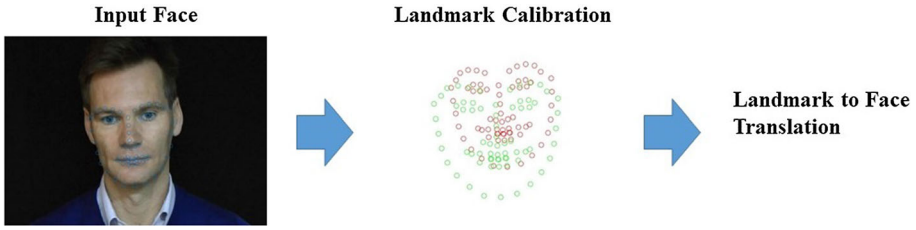

Fig. 1 Illustration of the two tasks in LACE: **a** calibration of landmarks, **b** translation into face images

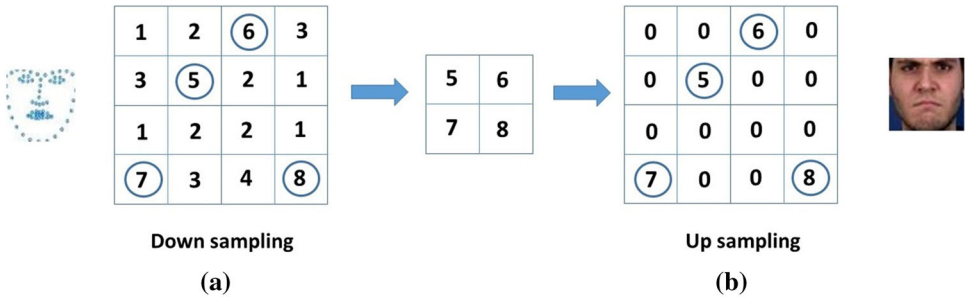

Fig. 2 Illustration of the two steps in LACE translation **a** downsampling of landmarks to convolutional features, **b** upsampling of features to generate face image

We can argue that such a model will over-fit to the training data. An adversarial model has two loss functions competing against each other. A new mathematical inequality is introduced to avoid over-fitting and convergence to a global minima. Figure 1 illustrates the methodology for the proposed framework. The input face is used to extract landmarks which are then calibrated for an emotion. Finally, landmarks are translated to the face using a GAN (see Fig. 2). In this paper, we aim to segment faces in a scene and then predict the emotional state of the person. The idea can be extended to segmentation of objects in any scenery [12]. Next, visual sentiment ontology in the form of adjective-noun pairs can be used to determine the polarity of the scene. For example, if a 'flower' is segmented from an image. The visual sentiment will be 'positive' [13,14].

The paper is organized into five sections. Section 2 reviews related works and dataset on image translation. Section 3 provides the preliminary concepts necessary to understand the present work. Section 4 details the proposed model for classifying facial expressions. Section 5 validates the method on two real-world dataset. Finally, conclusions are provided in Sect. 6.

2 Related works and contributions

Landmark ambiguity is a major challenge in facial expression detection. Previous authors have used supervised convolutional networks to predict landmarks [15]. However, due to large variability in facial features, it is difficult to annotate all of the landmarks accurately. Instead, the underlying expression class such as 'happy' or 'sad' can be used to calibrate landmarks for a face. In [7], the authors modeled facial expressions as a multi-label problem where a face may have two or more emotions. They modeled local label correlations inside clusters using singular value decomposition (SVD). Their method requires additional search for number of clusters and weights of local correlations. Hence, here we choose a high-intensity facial expression for an emotion such as a 'smile' and then calibrate a low-intensity image using SVD.

Similarly, in [16] landmarks in occluded face images were predicted using a distillation module and a low-rank learning module where missing features were interpolated using SVD. Here, again supervised learning was used to learn the shared inter-feature correlations. Lastly, in [6] predicted expressions using a combination of SVM and HMM. The SVM predicted if the action units such as a 'blink' had peak intensity or not. Next, a hidden Markov model used the SVM output to predict the facial expression. In contrast, the model described in this paper uses landmarks to simultaneously generate high-intensity face images and predict the emotions.

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

1. Use of image translation for accurate landmark segmentation of facial features such as the position of 'eye brows' and 'lips.'
2. For the same face image, calibrate the landmarks with respect to two emotions. The quality of image translation is then used as a metric to identify the true underlying emotion.
3. Derive a mathematical inequality between the generator loss and the discriminator loss for global convergence of the model.

Validation of the proposed method is performed on classification of facial expression into six emotions. Next, we apply it to the segmentation and classification of fish species

underwater. The face model can be used to predict the mood of friends in real time over social networks or FaceTime. The fish model will allow us to accurately locate and classify fish underwater or even in the market place. In order to implement such a model on a smart device, we require low latency, memory and power consumption. The traditional deep learning-based segmentation model can only detect around 10,000 objects and requires millions of neurons for high accuracy. Instead, we consider a generative model with only a few layers that can be trained to segment any desired object [17]. Features are extracted from the input image using three convolutional layers. The extracted features are combined with random noise. Next, three layers of upsampling is used to generate the segmented image. A discriminator is trained to determine if the generated image matches the target segmentation. Hence, the generator iteratively generates random images and the discriminator selects the best ones.

3 Preliminaries

In this section, the problem of landmark calibration is formulated with respect to a target emotion class. This is achieved by solving a system of linear equations for the coordinates for each landmark. Next, the entire face image is used instead of just the landmarks for the calibration. Here, an adversarial model where convolution is used to extract features from the images and upsampling is used to generate the complete facial expression for a particular emotion.

3.1 Landmark calibration using SVD

It is difficult to predict landmarks for a face image captured using a smart phone as the angle of the webcam keeps changing. Thus, calibration of the landmarks is necessary using a gold standard for each class. Here, representative landmarks for each emotion are considered such as 'happy' or 'sad' where the emotion is clearly represented using only 64 facial landmarks.

Notation: In this paper, we represent a 2D matrix using upper-case bold $\mathbf{S}$, a vector using lower-case bold $\mathbf{x}$ and a constant using italic lower-case n. Let us consider a vector of n landmark locations for a face in both horizontal $\mathbf{x}$ and vertical $\mathbf{y}$ directions. Hence, we can define a two-dimensional matrix of landmarks $\mathbf{L} = (\mathbf{x}, \mathbf{y})$. For reference, we can manually select a high-intensity emotional face with landmarks $\mathbf{L}^{\dagger}$. Next, each landmark in $\mathbf{L}$ is calibrated such that the distance from the corresponding landmark in $\mathbf{L}^{\dagger}$ is minimal.

Calibration is achieved by solving a system of linear equations where each new facial landmark in $\mathbf{L}$ is represented as a weighted sum over the gold standard $\mathbf{L}^{\dagger}$ as follows:

$$\mathbf{L} = \boldsymbol{\beta}\mathbf{L}^{\dagger} + \mathbf{e} \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of coefficients. Here, we aim to minimize the error $\mathbf{e}$ between the input landmark $\mathbf{L}$ and the gold standard $\mathbf{L}^{\dagger}$. In order to allow for nonlinear calibration in a two-dimensional plane, we consider higher-order transformations for each landmark. Here, we use a six-dimensional landmark matrix $\mathbf{L} = (1, \mathbf{x}^2, \mathbf{y}^2, \mathbf{xy}, \mathbf{x}, \mathbf{y})$. Now, we can rewrite Eq. 2 as follows:

$$\mathbf{x} = \boldsymbol{\beta_x}\mathbf{x}^{\dagger} + \mathbf{e_x}, \mathbf{y} = \boldsymbol{\beta_y}\mathbf{y}^{\dagger} + \mathbf{e_y}$$

where $(\boldsymbol{\beta_x}, \boldsymbol{\beta_y})$ are both six-dimensional coefficient vectors. We can note here that all 64 landmarks are transformed using the same coefficient vector. To solve this linear equation and predict the coefficients, we aim to minimize the error vector $\mathbf{e}$. This can be achieved by maximizing the covariance $\boldsymbol{\Sigma}$ between the six-orthogonal components. The covariance can be computed as $\boldsymbol{\Sigma} = \mathbf{L}^{\mathrm{T}}\mathbf{L}$ where $\mathbf{L}^{\mathrm{T}}$ is the transpose of the landmarks $\mathbf{L}$. To determine the maximum covariance components, we can rewrite the covariance as follows:

$$\boldsymbol{\Sigma} = \mathbf{V}\mathbf{D}\mathbf{V}^{\mathrm{T}} \tag{2}$$

where $\mathbf{D}$ is a diagonal matrix with the magnitudes of each of the six component and $\mathbf{V}$ is the direction vector of each of the six orthogonal components. The eigenvalues are in sorted order such that the top few have the highest covariance and hence minimal error. In order to determine the coefficients $\boldsymbol{\beta}$, we have to determine an orthonormal matrix $\mathbf{U} = \mathbf{I}^{\mathrm{T}}\mathbf{I}$ where $\mathbf{I}$ is the identity matrix. Then, we can derive the equation for the landmarks $\mathbf{L}$ as follows:

$$\boldsymbol{\Sigma} = \mathbf{L}^{\mathrm{T}}\mathbf{L} = \mathbf{V}\mathbf{D}\mathbf{V}^{\mathrm{T}} = (\mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}})^{\mathrm{T}}(\mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}})$$

where $\mathbf{D}\mathbf{D}^{\mathrm{T}}$ is also a diagonal matrix. Hence, we can conclude that $\mathbf{L} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}$. We can determine the three matrices $\mathbf{U}, \mathbf{D}, \mathbf{V}$ using the gold-standard landmark $\mathbf{L}^{\dagger}$. Next we consider the transformed landmarks:

$$\mathbf{m_x} = \mathbf{U}\mathbf{x}^{\dagger}, \boldsymbol{\beta_x} = \mathbf{V}(\mathbf{m_x}./\mathbf{D}) \tag{3}$$

where we only consider the top six values in the vector $\mathbf{m_x}$. Similarly, we can determine the coefficients for $\boldsymbol{\beta_y}$. In order to calibrate any new face, we have to transform the two-dimensional landmarks to six-dimensions and then perform a dot product with the coefficient vector $\boldsymbol{\beta}$. This process is referred to as singular value decomposition (SVD). Figure 3 illustrates an example of calibration of landmarks for 'fear' emotion. Next, we extend the idea from n landmarks to all the pixels in the face image. Given a pair of images (a) the face image and (b) the landmarks image, a model that can translate the landmarks to the complete face for a particular emotion is considered. The landmarks for different emotions of the same person will be different, hence the model will

Fig. 3 Calibration of landmarks for 'fear' emotion

learn to differentiate the different expressions for the same individual.

3.2 Convolutional GAN networks

The n landmarks can be used to identify the emotional state of the face. Here, a model is trained to generate the expression from the landmarks for any person. This allows us to control the expression of the person in a video by simply changing the input landmarks. Convolutional neural networks (CNNs) are commonly used to classify face images to different emotions. The CNN will extract significant features for each emotion and combine them to into the target emotion class. In addition, upsampling can be used to increase the resolution of the generated image to match the size of the target face image. Here, we partition each image into m groups and the coefficients take the form of two-dimensional kernels. Each of the m groups is associated with a $n_x \times n_y$ kernel where n_x is the height of the kernel and n_y is the width of the kernel. The low-resolution landmark image can now be represented as:

$$\hat{y} = \sum_{r,s}^{n_x,n_y} p_{i+r-1,j+s-1} h_{ij} \beta_{rs}. \tag{4}$$

where $\hat{y}$ is the low-resolution representation of the landmark image, h_{ij} is a pixel in the face image, p_{ij} is a pixel in the landmark image, and β_{rs} is the coefficient for a particular group in m. Due to sharing of weights a in each group, several pixels h_{ij} are redundant. Pooling is used to remove duplicate features. Next, upsampling is used to increase the resolution of the image so that it looks similar to the target face image. This results in the generator loss that is the difference between the upsampled image and the target face image. The vector $\hat{y}$ can also be trained to differentiate between the real target image and a fake image with a lot of noise. Here, the CNN is trained to classify an image as the real face or generated face as described above. Hence, we consider two CNN models (a) for generating images and (b) for discriminating images. This discriminator loss does not require upsampling. The combined model is known as a GAN.

In order to preserve the identity of a person, another domain discriminator is used where y is trained to differentiate between the true person in the face and the person in the generated image. In summary, there are three loss functions: (a) generator: e_g, (b) discriminator: e_d and (c) domain discriminator: e_{dd}. During training, we learn kernels using Eq. 4 such that the average over all the three loss functions is minimal. Figure 2 illustrates learning in a GAN made up of two convolutional neural networks. The first network extracts features from a landmark image using sliding window kernels. Downsampling selects the feature with highest activation among closest image features that are redundant. The second network works in the opposite manner; it uses upsampling to increase the resolution of features so that it matches the target face image.

4 Landmark to face mapping

In this section, a mathematical inequality is introduced based on Lipschitz continuity that provides bounds on the error vector e and ensures convergence to the global minima. Next, the complete framework for classifying facial expressions using calibration and adversarial training is described. The resulting model is referred to as landmark adversarial calibration for expressions (LACE).

4.1 Lipschitz inequality

Because the model will randomly select two persons each time to train the domain discriminator, it is likely to get stuck in a local minima. In order to avoid this, a mathematical inequality that ensures global convergence of the model after several iterations is introduced. In particular, the Lipschitz continuity condition to guarantee that the error updates during gradient descent approach zero as the number of iterations approaches infinity. This is achieved by constraining the weights with an upper and lower bound.

Consider the error vector $\mathbf{e} = (e_g, e_d, e_{gg})$ and a candidate function γ that is known to have a global minima solution:

$$\gamma(t) = \mathbf{e}^T \times \mathbf{S} \times \mathbf{e} \tag{5}$$

where t is the iteration index during training and $\{\mathbf{S}\}_{3\times 3}$ is an unknown matrix. Next, compute the eigenvalues of the covariance matrix of the error vector $\boldsymbol{\Sigma} = \mathbf{e}^T\mathbf{e}$. The highest eigenvalue λ of $\boldsymbol{\Sigma}$ represents the component with maximum variance. The last step is to heuristically determine the matrix s so that both values $(\gamma(t), \lambda)$ are almost equal for a given error vector $\mathbf{e}(t)$. Now, simply reject error updates where $\gamma(t) > \lambda$ as these will not follow the Lipschitz continuity requirement.

Figure 4 shows the effect of increasing generator loss e_g on the discriminator loss e_d. There cases are considered (a) $e_g = 0.1$ (b) $e_g = 20$ and (c) $e_g = 50$. This is because the generator loss is over all the pixels and will be higher for

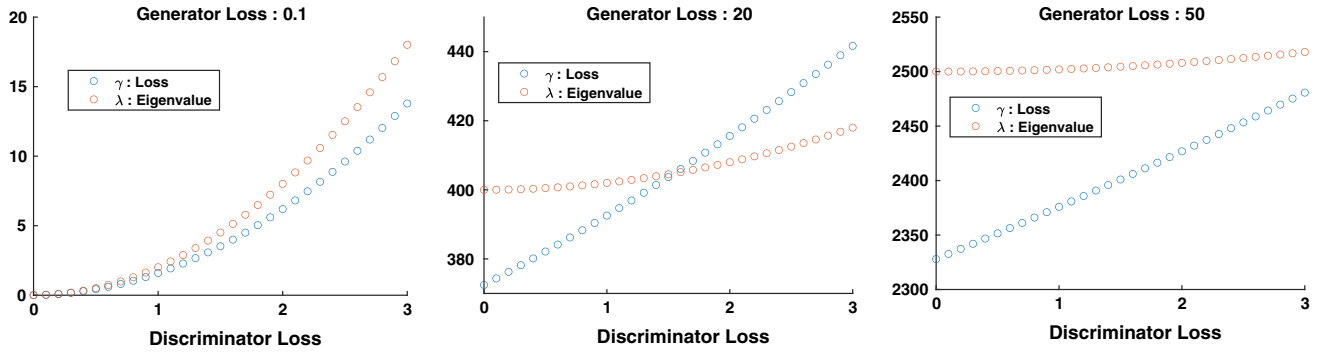

Fig. 4 Plot of change in candidate loss function γ as the discriminator loss e_d increases and the generator loss e_g is fixed

larger images. Next, we slowly increase the discriminator loss from 0 to 3 shown as the x-axis. Here, we assume $e_d = e_{dd}$. The matrix s is chosen heuristically such that the pair of values (γ, λ) are almost equal and $\lambda > \gamma$. For the second case in Fig. 4b, we clearly observe that while $\lambda > \gamma$, the discriminator error e_d is low; however, once e_d increases, we observe that $\lambda < \gamma$. Hence, we can conclude that for a smaller loss we should ensure the above inequality. For example, in the face dataset we arbitrarily select an error vector e $= (0.1, 1.5, 1.5)$ after a few iterations of training and determine the corresponding s matrix.

4.2 Expression classification framework

In Algorithm 1, the calibration method is applied to expression identification. Here, for any given input face image a pre-trained GAN is used to generate the different face emotions. It is worth noting that for the true emotion the GAN is able to easily generate the target image and the corresponding PSNR is highest. The training data contain face images for the same person and for different emotions. The first step at line no:5 is to extract landmarks for the faces using a general tracker trained for 64 facial landmarks including the position of eyes, nose and mouth. Next, we consider each pair of images: (a) the landmark image and (b) the face image. The GAN described in the previous section is used to translate a given landmark image to the corresponding face image. This is a difficult task, especially for subtle emotions such as 'sadness.'

Hence, the next step at line no:6 is to calibrate each landmark with respect to the gold-standard landmark for a particular emotion. Now, the GAN can be trained with all the pairs of images until the loss vector **e** does not reduce significantly. During testing, we only have the face image and the emotion is unknown. For this reason, calibration is attempted with respect to different emotions at line no:15 so as to find the best fit. For each test image, we first extract the landmark and then calibrate it for all the emotions.

The pair of images: (a) the landmark image and (b) the face image are input to the trained GAN. The GAN is again trained for one epoch to generate the face image and also preserve the identity of the person. We can now compute the PSNR between the generated image after one epoch and the input image. At line no:20, we see that the true label will have the lowest PSNR and therefore the best quality of generated image. The same procedure can be applied for fish identification. For the case of FISH dataset, the bounding box and the fish type are known during training. During testing, gold-standard images of the fish available on the Internet are used. The GAN then generates the segmented fish, and the PSNR is used to determine the fish type.

Algorithm 1 LACE Algorithm

```
1: % Training of LACE model
2: for For all Emotions p do
3:     % Train GAN for emotion p
4:     for For all training face images i do
5:         Extract landmarks for image i
6:         Calibrate landmarks L for image i
7:         % Learn GAN translation:(landmarks→image)
8:         while epochs < T do
9:             Update weights using loss functions
10:        end while
11:    end for
12: end for
13: % Testing of LACE model
14: for For all testing face images i do
15:    for For all Emotions p do
16:        Calibrate landmarks for image i for emotion p
17:        Translate using trained GAN for emotion p
18:        Compute PSNR (face image, generated image)
19:    end for
20:    Label image i with lowest PSNR emotion p
21: end for
```

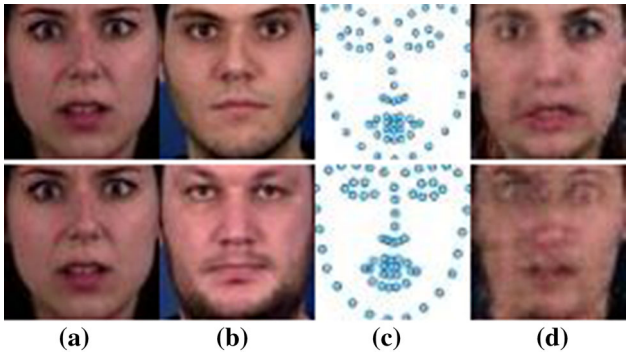

Fig. 5 Comparison of generated face by LACE (top row) and GAN (bottom row). **a** Face image, **b** random person, **c** landmarks, **d** generated face

5 Experiments

Validation of the proposed LACE method (available on GitHub[1]) is performed on two real-world benchmarks. The first is facial expression classification, and the second is fish identification and segmentation. We also detail the parameter setting and visualize the generated face images and the fish segmentation region. The width and height of face images are set to 200 pixels. The GAN is trained for 5000 epochs with a learning rate of 0.002. The running time for training each GAN is about two hours. The unknown matrix s is determined as described in Sect. 4 and fixed for all the simulations.

5.1 MUG

The model is trained on several videos of the individuals where the expression gradually changes from neutral to an emotional face. The beginning of each video is a neutral expression, and the intensity of the emotion (for example a smile) increases slowly until it is maximum. The MUG Facial Expression database contains image sequences from 86 subjects and for six emotions [18], namely 'happy,' 'sad,' 'fear,' 'surprise,' 'anger' and 'disgust.'

The subjects were sitting in a chair, and a camera was used to capture images at the rate of 19 frames per second and a resolution of 900 pixels. There are no occlusions on the face such as hair or spectacles. The participants were informed about the facial action units for the six emotions as defined in the FACS manual.

For training, we only consider the maximum intensity images for a particular emotion. The corresponding landmark points are extracted using a previous benchmark. The landmarks are calibrated using a benchmark landmark for the emotion before training the GAN. During training, the proposed LACE model is provided with the landmark images as the source and the smiling face as the target. For testing, we calibrate the landmarks with respect to different emotions and the generated images with highest PSNR is used to label the test face image. Figure 5 shows the comparison of generated face by LACE (top row) and GAN (bottom row). Table 1 shows that the PSNR is highest for the true expression class in the test data. We compare with a baseline CNN that has extremely low accuracy on anger and fear. The proposed LACE has over 30% higher accuracy on these emotions. Further, improvement is observed when we consider the Lipschitz constraint. The improvement in LACE-l is almost 10% more. For example, accuracy of fear increases to 70%. We also compare the average accuracy over all emotions with baselines in Table 2. We can see that the proposed LACE-l outperforms the baselines by over 20% in accuracy. Methods such as bag-of-words and pre-trained GoogleNet have only 65% accuracy on multi-class emotion classification.

Table 1 Accuracy comparison of LACE with and without Lipschitz continuity

Method	Anger	Fear	Sad	Neutral	Surprise	Happy
CNN	58	39	63	58	73	87
LACE	70	60	85	80	100	55
LACE-l	75	70	100	90	100	90

The baseline CNN shows very poor accuracy on fear and anger

Table 2 Comparison of average classification accuracy for MUG and FISH datasets

Method	MUG [19]	FISH [20]
Bag-of-words	67.4 [21]	–
GoogleNet	65.2 [22]	53.2 [23]
ML	–	49.1 [24]
LACE-l	87.5	80

5.2 Fishnet open image dataset

Commercial fishing boats use electronic monitoring devices and on-board cameras to capture fish images in the sea [25]. The Fishnet Open Image dataset has over 35,000 images from 28 different species. The dataset includes images of fishes on the deck of the vessel or alongside the vessel in the water as long as 80% of the fish body is identifiable. The location and type of each fish are marked using a bounding box by human annotators. As an example, we train a LACE model with images containing 'Albacore' fish. The model is supplied by a pair of images, where the source is the entire scenic image containing the fish and the target image is the the location of the fish with a white background.

[1] http://github.com/ichaturvedi/landmark-adversarial-calibration-for-expressions.

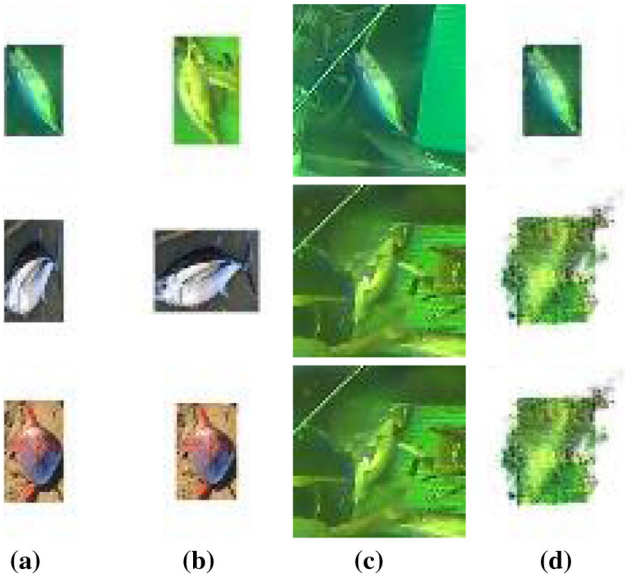

Fig. 6 Comparison of generated fish for training Albacore (top row) and for testing with Opah vs Albacore (bottom two rows) **a** fish species and location, **b** random fish, **c** underwater image, **d** generated fish

For training, 200 scenic images containing the species 'Albacore' are used. During testing, we provide LACE with the entire scene and two benchmark images from 'Albacore' and 'Opah.' We use the PSNR of the generated fish with the target. Table 2 shows that the PSNR for Albacore is higher in 80% of the test samples. This is significantly higher than accuracy of the baseline maximum-likelihood and pre-trained GoogleNet approaches. Thus, it can be concluded that LACE is a good model for fish classification both underwater and on the deck. Figure 6 shows the comparison of generated fish for training Albacore (top row) and for testing with Opah vs Albacore (bottom two rows).

6 Conclusions

Facial expression prediction is challenging for subtle emotions such as 'sadness' and for persons with soft features. In this paper, image translation is used to predict the underlying problem in a multi-label framework. Firstly, we calibrate the landmarks from a simple CNN classifier using a gold-standard image for a single emotion. Next, we train a landmark to face GAN for different emotions of the same person collectively. The resulting GAN has superior accuracy in image translation of subtle emotions that was not possible without calibration. In order to prevent convergence to a local minima, the loss function of the GAN is constrained using Lipschitz inequality. The new loss function is able to achieve higher accuracy in emotion and fish classification.

Acknowledgements This work is partially supported by the Computational Intelligence Lab at the Nanyang Technological University. This work is also partially supported by Information Technology, College of Science and Engineering at James Cook University.

References

1. Poria, S., Chaturvedi, I., Cambria, E., Hussain, A.: Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: ICDM, Barcelona, pp. 439–448 (2016)
2. Chaturvedi, I., Satapathy, R., Cavallari, S., Cambria, E.: Fuzzy commonsense reasoning for multimodal sentiment analysis. Pattern Recogn. Lett. **125**, 264–270 (2019)
3. Chaturvedi, I., Xiang, J.: Constrained manifold learning for videos. In: IJCNN, pp. 1–8 (2020)
4. Li, Y., Pan, Q., Wang, S., Yang, T., Cambria, E.: A generative model for category text generation. Inf. Sci. **450**, 301–315 (2018)
5. Susanto, Y., Livingstone, A., Ng, B.C., Cambria, E.: The hourglass model revisited. IEEE Intell. Syst. **35**(5), 96–102 (2020)
6. Bartlett, M.S., Littlewort, G., Braathen, B., Sejnowski, T.J., Movellan, J.R.: A prototype for automatic recognition of spontaneous facial actions. In: NIPS, pp. 1295–1302 (2002)
7. Jia, X., Zheng, X., Li, W., Zhang, C., Li, Z.: Facial emotion distribution learning by exploiting low-rank label correlations locally. In: CVPR, pp. 9833–9842 (2019)
8. Davison, A.K., Lansley, C., Costen, N., Tan, K., Yap, M.H.: Samm: a spontaneous micro-facial movement dataset. IEEE Trans. Affect. Comput. **9**(1), 116–129 (2018)
9. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: a survey of registration, representation, and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(6), 1113–1133 (2015)
10. Zhang, K., Huang, Y., Du, Y., Wang, L.: Facial expression recognition based on deep evolutional spatial–temporal networks. IEEE Trans. Image Process. **26**, 4193–4203 (2017)
11. Shojaeilangari, S., Yau, W., Nandakumar, K., Li, J., Teoh, E.K.: Robust representation and recognition of facial emotions using extreme sparse learning. IEEE Trans. Image Process. **24**(7), 2140–2152 (2015)
12. Qian, C., Chaturvedi, I., Poria, S., Cambria, E., Malandri, L.: Learning visual concepts in images using temporal convolutional networks. In: SSCI, pp. 1280–1284 (2019)
13. Ragusa, E., Apicella, T., Gianoglio, C., Zunino, R., Gastaldo, P.: Design and deployment of an image polarity detector with visual attention. Cogn. Comput. 1–13 (2021)
14. Ragusa, E., Cambria, E., Zunino, R., Gastaldo, P.: A survey on deep learning in image polarity detection: balancing generalization performances and computational costs. Electronics **8**(7), 783 (2019)
15. Liu, Z., Zhu, X., Hu, G., Guo, H., Tang, M., Lei, Z., Robertson, M.N., Wang, J.: Semantic alignment: finding semantically consistent ground-truth for facial landmark detection. In: CVPR, pp. 3467–3476 (2019)
16. Zhu, M., Shi, D., Zheng, M., Sadiq, M.: Robust facial landmark detection via occlusion-adaptive deep networks. In: CVPR, pp. 3481–3491 (2019)
17. Ragusa, E., Gianoglio, C., Zunino, R., Gastaldo, P.: Image polarity detection on resource-constrained devices. IEEE Intell. Syst. **35**(6), 50–57 (2020)
18. Aifanti, N., Papachristou, C., Delopoulos, A.: The mug facial expression database. In: WIAMIS, pp. 1–4 (2010)
19. Giannopoulos, P., Perikos, I., Hatzilygeroudis, I., Palade, V.: Deep learning approaches for facial emotion recognition: A case study

on fer-2013. In: Advances in Hybridization of Intelligent Methods: Models. Systems and Applications, pp. 1–16 (2018)
20. Siddiqui, S.A., Salman, A., Malik, M.I., Shafait, F., Mian, A., Shortis, M.R., Harvey, E.S.: Automatic fish species classification in underwater videos: exploiting pre-trained deep neural network models to compensate for limited labelled data. ICES J. Mar. Sci. **75**(1), 374–389 (2017)
21. Gidaris, S., Bursuc, A., Komodakis, N., Perez, P., Cord, M.: Learning representations by predicting bags of visual words. In: CVPR (2020)
22. Qian, Y., Deng, W., Hu, J.: Unsupervised face normalization with extreme pose and expression in the wild. In: CVPR (June 2019)
23. Fan, Z., Yu, J.-G., Liang, Z., Ou, J., Gao, C., Xia, G.-S., Li, Y.: gn: fully guided network for few-shot instance segmentation. In: CVPR (2020)
24. Hsiao, Y.-H., Chen, C.-C., Lin, S.-I., Lin, F.-P.: Real-world underwater fish recognition and identification, using sparse representation. Ecol. Inform. **23**, 13–21 (2014) (**special Issue on Multimedia in Ecology and Environment**)
25. Fishnet: The nature conservancy (2020): Fishnet open images dataset v0.1.2 the nature conservancy. dataset. The Nature Conservancy (2020). Data retrieved http://fishnet.ai

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.