3D Model-based Continuous Emotion Recognition

Hui Chen[1], Jiangdong Li[2], Fengjun Zhang[3], Yang Li[1], Hongan Wang[2,3]
[1]Beijing Key Lab of Human-computer Interaction, Institute of Software, Chinese Academy of Sciences. [2]University of Chinese Academy of Sciences.
[3]State Key Lab of Computer Science, Institute of Software, Chinese Academy of Sciences.

Continuous emotion analysis refers to acquire and process long unsegmented naturalistic inputs and to predicate affective values represented in dimensional space [4]. Traditional emotion recognition algorithms try to classify emotions into different categories. But in natural communications, human emotions are revealed naturally. Thus it is better to estimate emotions as real values of different affective dimensions instead of different categories for higher quality of human-computer interactions.

We propose a real-time 3D model-based method that continuously recognizes dimensional emotions from facial expressions in natural communications. In our method, 3D facial models are restored, which provide crucial clues to overcome large changes including out-of-plane head rotations, fast head motions and partial facial occlusions. Via the reconstructed 3D facial model, temporal information and user-independent emotion presentations are also taken into account through our image fusion process. To ensure accuracy and efficiency, a novel random forest-based model is constructed, which integrates two regressions for 3D facial tracking and continuous emotion estimation simultaneously. The framework is shown in Figure 1.

For every input 2D image, the 42 inner facial landmarks we used are detected via the algorithm proposed by Baltrusaitis *et al* [1]. Then we construct all the 3D facial shapes using fundamental blendshapes of the same identity in FaceWarehouse in order to represent the input images from different persons in a uniform way. The optimum identity can be calculated via formula:

$$E_{id} = \sum_{i=1}^{N}\sum_{b=1}^{42} \left\| P(M^i(C_r \times w_{id}^T \times w_{exp,i}^T)^b - u_i^b) \right\|^2 \quad (1)$$

where N is the number of the picked images; P means the projection matrix; M^i means the extrinsic parameter matrix of camera; $w_{exp,i}^T$ stands for the most similar expression for the i^{th} image; u_i^b is the b^{th} landmark on image. The optimum identity w_{id}^T is the one with minimum energy E_{id}. With the acquired fundamental blendshapes, the 3D facial model of every image can be restored via the linear interpolation of fundamental as [2] did.

With the help of the 3D emotion presentations, an image fusion method is implemented to generate temporal information and user-independent information. First of all, we reconstruct the 3D facial model of input images. Then the 3D facial shape is transformed to the orthogonal position of space coordinate system and projected to the 2D facial coordinate system. With the original landmarks and the projected landmarks, the homographic transform matrix from the original screen space to the facial coordinate space is acquired. The facial part of original image is unified into the 2D facial coordinate system. After transforming all the facial parts of original images to the unified facial coordinate system, these images are superposed and result in one fusion presentation.

We generate user-specific continuous emotion presentation(CEP) and user-independent emotion presentations(UIEP) via a novel 3D model based image fusion method. CEP merges several adjacent frames from a video clip, which is used to contain temporal context of emotions. UIEP fuses images from different videos with the same emotion value into one image , which is used to retain the prominent features and eliminate the differences among different persons.

To construct random forest, we firstly expend the training samples. With the augmented shapes and emotion values, together with the CEP images, training patches for constructing the random forest are generated. When training every CART, only part of the patches are used to avoid over-fitting.

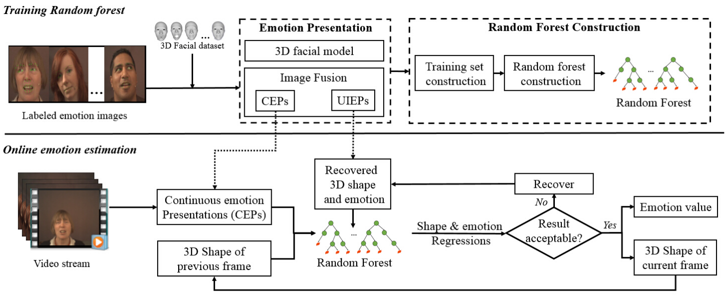
Training Random forest
3D Facial dataset
Labeled emotion images
Emotion Presentation
3D facial model
Image Fusion
CEPs
UIEPs
Random Forest Construction
Training set construction
Random forest construction
Random Forest
Online emotion estimation
Video stream
Continuous emotion Presentations (CEPs)
3D Shape of previous frame
Recovered 3D shape and emotion
Random Forest
Shape & emotion Regressions
Result acceptable?
No
Recover
Yes
Emotion value
3D Shape of current frame

Figure 1: Framework of our 3D model-based continuous emotion recognizing and tracking approach

Similar to [3], in every non-leaf node, a binary test is defined to to split training patches:

$$|F_1|^{-1} \sum_{q1 \in F_1} Int_{q1} - |F_2|^{-1} \sum_{q2 \in F_2} Int_{q2} > \tau \quad (2)$$

where F_1 and F_2 means two fragments from current training patch, Int represents the intensity vector and τ is a random threshold.

For each non-leaf node, we generate a lot of binary tests by randomly choosing the parameters of F_1, F_2 and τ. The quality of every binary test is evaluated by regression uncertainty U_R, which consists two regression uncertainties:

$$U_{R_s}(P \mid t^x) = H(P)_s - w_L H(P_L)_s - w_R H(P_R)_s \quad (3)$$

$$U_{R_a}(P \mid t^x) = H(P)_a - w_L H(P_L)_a - w_R H(P_R)_a \quad (4)$$

where $H(P)$ means the differential entropy of patch set and w_L, w_R are the ratio of patches sent into left and right child node respectively. Then total uncertainty U_R can be presented as:

$$U_R(P \mid t^x) = U_{R_s}(P \mid t^x) + \lambda U_{R_a}(P \mid t^x) \quad (5)$$

where λ is an empirical weight. By maximizing U_R, we can find the best binary test of current node.We save the parameters of the optimal binary test as a part of random regression forest and split the training patches of current node. A node is taken as a leaf if it reaches the deepest level or the patches it contains is less than the minimum threshold. A leaf node stops splitting and saves the information about patches it holds including the mean and covariance of shape displacements together with the average and covariance of affect displacements.

With the constructed random forest, the continuous emotion can be calculated. Taking CEP at the current time step, the 3D emotion shape and the affective value at the previous time step as input, the 3D emotion shape and affective value of current time step can be estimated in a regression way. The experimental results show that our algorithm can achieve good performance and run in real time.

[1] T. Baltrusaitis, P. Robinson, and L. P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *IEEE International Conference on Computer Vision Workshops*, pages 354–361, 2013.

[2] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3d shape regression for real-time facial animation. *ACM Transactions on Graphics*, 32(4):41, 2013.

[3] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool. Random forests for real time 3d face analysis. *International Journal of Computer Vision*, 101(3):437–458, 2013.

[4] H. Gunes and B. Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120–136, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.