Parametric temporal alignment for the detection of facial action temporal segments

Bihan Jiang[1]
bi.jiang09@imperial.ac.uk
Brais Martinez[1]
b.martinez@imperial.ac.uk
Maja Pantic[12]
m.pantic@imperial.ac.uk

[1] Computing Department
Imperial College London, UK

[2] Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente,
Netherlands

We propose a new methodology for producing temporal alignment of facial behaviour, and apply it to the analysis of the facial action units (AU) temporal segments. Therefore, our contributions are twofold. In first place, we propose a new methodology for temporal alignment of two sequences of facial behaviour. Secondly, we propose a new way of segmenting the AU temporal segments that relies on the temporal alignment of an exemplar sequence (a template) with the test sequence.

Alignment methodology The temporal alignemnt strategy builds on the work of [4]. In this work, the authors managed to project a sequence into a parametric curve embedded into a lower-dimensional space by applying Laplacian eigenmaps. Furthermore, they were able to backproject from this curve into frame space by means of a simple linear transformation. Formally, if $X = \{\mathbf{x}_t\}_{t=1:n}$ is the original sequence, then this technique allows the construction of a continuous parametric approximation of the original sequence as:

$$\mathcal{X}(t) = A(X)\mathcal{Y}(t) + \bar{\mathbf{x}} \tag{1}$$

where $\mathcal{Y}(t)$ is the curve embedded in the lower dimensional space, and $A(X)$ is a matrix that depends on the original sequence. Crucially, $\mathcal{Y}(t)$ has an analytical form and can be derived analytically.

We then consider a family of parametric functions that represent the possible temporal transformations. For example, we can use a linear warp to account for constant differences on the speeds of actions, or a piecewise linear function. $W(-;\theta)$ represents such transformation parametrised by θ. If aligning the test function onto a template sequence, we define the loss function of the alignment between the template and the test sequence as:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i}^{n} \|\mathbf{x}_i^{\text{tmpl}} - \mathcal{X}(W(i;\theta))\|_2^2 \tag{2}$$

Applying the chain rule and the fact that $\mathcal{Y}$ can be analytically differentiated, then we can compute:

$$\frac{\partial \mathcal{Y}(\mathcal{W}(i;\theta))}{\partial \theta_j} = \left.\frac{\partial \mathcal{Y}(t)}{\partial t}\right|_{\mathcal{W}(i;\theta)} \frac{\partial \mathcal{W}(t;\theta)}{\partial \theta_j} \tag{3}$$

It is then possible to minimise the loss function using a Gauss-Newton approach as:

$$\theta^{(it+1)} = \theta^{(it)} - \left(\mathbf{J}_{\theta^{(it)}}^t \mathbf{J}_{\theta^{(it)}}\right)^{-1} \mathbf{J}_{\theta^{(it)}}^t \mathbf{r}\left(\theta^{(it)}\right) \tag{4}$$

where $\mathbf{J}_{\theta^{(it)}}$ is the Jacobian of $\mathcal{X}$ respect to the warp parameters θ.

Application to AU temporal segment detection: The AU temporal segments are defined as neutral (no activation), onset (increase of intensity of the AU), apex (maintain) and offset (decay of intensity of the AU). The task is to label each frame of a sequence accordingly. This is typically done by training per-frame classifiers. However, we propose instead to align the test sequence with an exemplar sequence with known labels (a template). The template labels are then mapped through the alignment function to produce the test sequence labelling.

We define two different warp functions. The first one aligns a full activation episode to the test sequence by using a piecewise linear warping. This model adapts to linear differences in speed independently for each AU segment. This model is illustrated in Fig. 1. Specifically, the warp function is defined as:

$$W(i;\theta) = \begin{cases} \frac{\theta_2-\theta_1}{n_{\text{on}}} i + \theta_1 & : \theta_1 \le i < \theta_2 \\ \frac{\theta_3-\theta_2}{n_{\text{ap}}} i + \theta_2 & : \theta_2 \le i < \theta_3 \\ \frac{\theta_4-\theta_3}{n_{\text{off}}} i + \theta_3 & : \theta_3 \le i \le \theta_4 \end{cases} \tag{5}$$

However, this model does not account for different AU intensities. Smiles can be low intensity (closed mouth and low intensity of the mouth corner pulling) or broad smiles (with open stretched mouth). The second model accounts for this differences. In particular, the action exemplar and the test sequence do not need to be aligned in full. Therefore, the template should reach maximum intensity. This model is illustrated on the right hand side part of in Fig. 1.

$$W(i;\theta) = \begin{cases} 0 & : i < \theta_1 \text{ or } \theta_4 \le i \\ \frac{\theta_5}{\theta_2-\theta_1}(i-\theta_1) & : \theta_1 \le i < \theta_2 \\ \theta_5 & : \theta_2 \le i < \theta_3 \\ -\frac{\theta_5}{\theta_4-\theta_3}(i-\theta_3)+\theta_5 & : \theta_3 \le i < \theta_4 \end{cases} \tag{6}$$

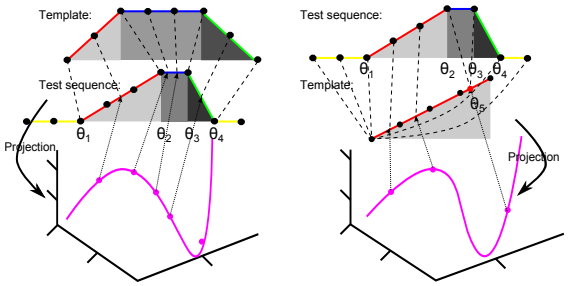

Figure 1: Depiction of the temporal alignment strategy for both of the models presented here (left: model1, right: model2).

The performance achieved by model 2 is the best. However, both models provide superior performance to other state of the art methods, as shown in Table 1.

Table 1: Comparison of AU temporal segment detection methods on the MMI database. $\mathbf{F1}_{\text{act}}$ is the F1-measure after converting into AU activation.

Systems	Neutral	Onset	Apex	Offset	$\mathbf{F1}_{\text{act}}$
Model1	83.42	54.15	78.86	57.87	77.83
Model2	**85.88**	**56.32**	**79.75**	**58.95**	**80.62**
Jiang et al. 2013[1]	78.50	53.38	72.12	48.73	67.53
Valstar et al. 2012[3]	76.60	56.75	69.38	48.87	-
Koelstra et al. 2010[2]	-	-	-	-	62.5

[1] B. Jiang, M. F. Valstar, B. Martinez, and M. Pantic. Dynamic appearance descriptor approach to facial actions temporal modelling. *IEEE Trans. Systems, Man and Cybernetics, Part B*, 44(2):161–174, 2014.

[2] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture based approach to recognition of facial actions and their temporal models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(11): 1940–1954, 2010.

[3] M. F. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Trans. Systems, Man and Cybernetics, Part B*, 42(1):28–43, 2012.

[4] Z. Zhou, G. Zhao, and M. Pietikaïnen. Towards a practical lipreading system. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 137–144, 2011.