A note on modelling a somatic motor space for affective facial expressions

Alessandro D'Amelio[1], Vittorio Cuculo[1 2], Giuliano Grossi[1], Raffaella Lanzarotti[1], and Jianyi Lin[3]

[1] PHuSe Lab - Dipartimento di Informatica, Università degli Studi di Milano
Via Comelico 39/41, Milano, Italy
alessandro.damelio@studenti.unimi.it
{grossi,lanzarotti}@di.unimi.it
[2] Dipartimento di Matematica, Università degli Studi di Milano
Via Cesare Saldini 50, Milano, Italy
vittorio.cuculo@unimi.it
[3] Department of Mathematics, Khalifa University of Science and Technology,
Abu Dhabi, United Arab Emirates
jianyi.lin@kustar.ac.ae

Abstract. We discuss modelling issues related to the design of a somatic facial motor space. The variants proposed are conceived to be part of a larger system for dealing with simulation-based face emotion analysis along dual interactions.

Keywords: Emotion, human-agent interaction, simulation, Kalman filter, probabilistic generative models

1 Introduction

In the course of actual interactions (human-human or human-agent), the unfolding of emotional episodes is likely to follow a different route than pursued by a large body of work in affective facial expression analysis where a computer vision "pipeline-based" approach is followed (feature extraction then recognition/classification [14]). Facial expressions are facial actions and are likely to draw on simulation mechanisms underlying action perception in general [16]. These rely on mirroring processes that ground the capability of own reproduction of the action in question "as if" a similar action were performed or a similar emotion experienced.

At the heart of the simulation-based framework is the modelling of a suitable visuomotor mapping of perceived facial cues to an internal somatic motor space, which, in turn, works side by side with core affect components via forward and backward connections [16]. Importantly, such internal motor space must be endowed with generative capabilities, so to support actual simulation (e.g. facial mimicry). In this note we discuss, from a probabilistic standpoint, some modelling issues that arise in this effort. A relevant one is the hierarchy of levels of predictive control (for an in-depth discussion see [11].

Not much effort has been spent in such direction. We build on [15], addressing a mapping from visual cues to a probabilistic core affect space within a simulation-based paradigm. However, in that case, only static images are considered, and most important, motor representation is not explicitly addressed. An even simpler variant is presented in [7]. Though not addressing the issue of motor simulation, Fan *et. al.* [6] exploit the motor control sequence $\mathbf{m}(t)$ - derived from a 3D shape model as the observation input to a Kalman filter. The authors are mostly concerned with the classification of basic emotions, rather than building a continuous latent space of actions akin to support visuomotor learning and simulation.

2 Modelling issues

We assume that the observer $\mathcal{O}$ perceives the facial action of the expresser $\mathcal{E}$ in terms of the visible cues, say $\mathbf{y}_{\mathcal{E}}$, captured by his visual system and maps such cues onto his own internal motor action representation (*visuomotor mapping* [9]). The observer's internal representation not only "stands for" the visual signalling generated by $\mathcal{E}$, but, in a simulation-based account of facial expression analysis, it must be apt to generate the internal facial dynamics for mirroring that of $\mathcal{E}$.

From a modelling perspective, the egocentric motor representation of the face of agent $\mathcal{I} \in \{\mathcal{E}, \mathcal{O}\}$ is accounted for by the state-space RV $\mathbf{w}(t) = \mathbf{w}(\mathbf{m}(t), \mathbf{s}_{\mathcal{I}})$.

Here, $\mathbf{s}_{\mathcal{I}}$ stands for a set of static parameters that control the biometric characteristics of each individual $\mathcal{I} \in \{\mathcal{E}, \mathcal{O}\}$; we assume that observer's parameters $\mathbf{s}_{\mathcal{O}}$ are given, while expresser's parameter $\mathbf{s}_{\mathcal{E}}$ are inferred by the observer at the onset of the interaction.

The action control is given by the motor parameters $\mathbf{m}(t)$ controlling the facial deformation due to muscle action. Motor control parameters $\mathbf{m}(t)$ tune the actual evolution of the internal facial dynamics $\mathbf{w}(t)$, but are in turn governed by a specific action which we represent as a trajectory in a latent action state-space, formalised via the time-varying hidden RV $\mathbf{h}(t)$. The latent facial action state-space dynamics is affect-driven, since in the context of affective interactions can be assumed to be "biased" by the dynamics of the core affect [13].

The generative stage can be written in the form of an ancestral sampling procedure on the Probabilistic Graphical Model (PGM) shown in Fig. 1a:

1. Sampling a time dependent action state from the latent affect-driven action space:
$$\widetilde{\mathbf{h}}(t+1) \sim P(\mathbf{h}(t+1) \mid \mathbf{h}(t)); \tag{1}$$
2. Sampling facial action control parameters conditioned on the current affect-state and on the inferred control parameters:
$$\widetilde{\mathbf{m}}(t+1) \sim P(\mathbf{m}(t+1) \mid \widetilde{\mathbf{h}}(t+1)), \tag{2}$$
3. Motor-state space dynamics towards visuomotor mapping

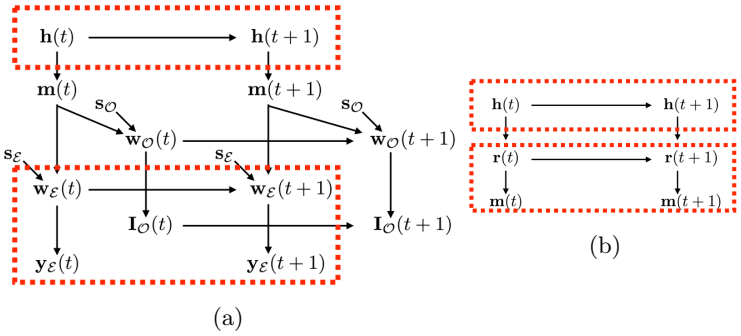


Fig. 1: Modelling issues at a glance. (a): the dynamic PGM representation of the model. The dashed boxes show the two levels of predictive control. (b): the Kalman-based predictive component summarised as a further level of control within the original PGM

(a) Use sampled control parameters, and sample a facial configuration of the expresser $\mathcal{E}$, by setting $\mathbf{w}_{\mathcal{E}}(t+1) = \mathbf{w}(\widetilde{\mathbf{m}}(t+1), \mathbf{s}_{\mathcal{E}})$:

$$\widetilde{\mathbf{w}}_{\mathcal{E}}(t+1) \sim P(\mathbf{w}_{\mathcal{E}}(t+1) \mid \mathbf{w}(t), \widetilde{\mathbf{m}}(t+1)) \tag{3}$$

(b) Sample facial landmarks in expresser visual space

$$\widetilde{\mathbf{y}}_{\mathcal{E}}(t+1) \sim P(\mathbf{y}_{\mathcal{E}}(t+1) \mid \widetilde{\mathbf{w}}_{\mathcal{E}}(t+1)) \tag{4}$$

If external simulation (actual facial mimicry) is enabled, the visible facial expression of the observer can be obtained by setting $\mathbf{w}(t+1) = \mathbf{w}_{\mathcal{O}}(\widetilde{\mathbf{m}}(t+1), \mathbf{s}_{\mathcal{O}})$. Then state is sampled analogously to Eq. 5 and facial mimicry generated via $\widetilde{\mathbf{I}}_{\mathcal{O}}(t+1) \sim P(\mathbf{I}_{\mathcal{O}}(t+1) \mid \mathbf{w}(t+1), \mathbf{I}_{\mathcal{O}}(t))$.

Note that such a generative model, focusing on the expresser's side, can be seen as a hierarchical predictive control model where the lowest level predicts the motor state and then generates an estimate of expresser's visual landmarks. At this level, novel predictions are governed by the error or discrepancy between the estimated landmarks and the observation of expresser's landmarks. Indeed, this level can be seen as an instance of model-based predictive coding that has been widely adopted in the video processing realm.

At the highest level, that is prediction, parameter estimation and error correction are implicitly obtained by relying on the action state-space dynamics, and on the optimization procedures in such latent space. This is the meaning of Eq 2. This has some modelling compactness and efficiency advantages, whilst drawbacks could occur due to the fact that, in principle, the lower dimensionality action space (that is, in turn, related to core affect dynamics) might operate on a coarser time scale than that of motor parameter dynamics. In a more general setting one should consider parameter sampling based on the conditional distribution $P(\mathbf{m}(t+1) \mid \widehat{\mathbf{m}}(t), \widetilde{\mathbf{h}}(t+1))$, where the dynamics is explicitly handled.

To suitably ground the discussion, the observer's internal motor space is formalised as a 3D *deformable shape model* consisting of a collection of N vertices represented by $\mathbf{w} = [\mathbf{w}_1 \cdots \mathbf{w}_N] \in \mathbb{R}^{3\times N}$, where every 3-dimensional vector $\mathbf{w}_i = (X_i, Y_i, Z_i)^T$ corresponds to the i-th vertex in the model. The dynamical evolution of the motor state is captured in the model by the dependence of the vectors upon the time variable t, so that each vertex follows a curve $\mathbf{w}_i(t) = (X_i(t), Y_i(t), Z_i(t))^T$.

It can be shown that under Helmholtz's fundamental theorem for deformable bodies [8] and small rotations, prediction of face motion at vertex i can be written (assuming unitary time step) as:

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \mathbf{R}(t)\mathbf{w}_i(t) + \mathbf{dW}_i^S\mathbf{s} + \mathbf{dW}_i^M\mathbf{m}(t) + \mathbf{t}(t). \tag{5}$$

where the pose parameters $\boldsymbol{\theta}(t) = (\mathbf{R}(t), \mathbf{t}(t))$ represent the rotation matrix $\mathbf{R}(\omega) \in \mathrm{SO}(3)$ with angular velocity vector $\omega = (\omega_x, \omega_y, \omega_z)$ and the translation vector, respectively, that is the global rigid motion constrained by cranial pose dynamics. As to the deformation term, $\mathbf{dW}_i^S \in \mathbb{R}^{3\times N_s}$ and $\mathbf{dW}_i^M \in \mathbb{R}^{3\times N_m}$ are respectively the matrices of Shape Unit (SU) and Action Unit Vector (AUV) deformation. Individual biometric control parameters $\mathbf{s}$ are considered fixed along the interaction, for both expresser and observer. Eq. 5 applied to all vertices represents the motor state of the 3D face model evolving in time, i.e. the forward model.

The generation (estimate) of expresser's visual landmarks is obtained as the projection of the 3D vertices on the 2D image coordinate system, under weak perspective projection (given the small depth of the face [10]), namely $\widehat{\mathbf{y}}_{\mathcal{E},l} = \mathcal{T}\widetilde{\mathbf{w}}_{\mathcal{E},l}$ where l indexes the L vertices that are in correspondence with extracted facial landmarks. Under Gaussian assumption, parameter inference boils down to the negative log-likelihood minimisation problem, which gives the "observed" $\widehat{\mathbf{m}}(t)$ and where the error control is accounted for by term $\|\mathbf{y}_{\mathcal{E},l} - \widehat{\mathbf{y}}_{\mathcal{E},l}\|^2$.

As to the top control level, the latent action space can be specified by resorting to a dynamical variational Gaussian Process Latent Variable Model (DVGP-LVM, [4]). The variational $\mathcal{GP}$ provides an efficient nonlinear mapping. In such setting, Eqs. 1 2 are suitably implemented, and for a single parameter m_k, Eq. 2 becomes

$$m_k(t) = f_k(\mathbf{h}(t)) + \nu_{\mathbf{h}}(t), \;\; \nu_{\mathbf{h}} \sim \mathcal{N}(0, \sigma_{\mathbf{h}}^2), \tag{6}$$

where f_k is a latent mapping from the low dimensional action space to the k-th dimension of the parameter space of $\mathbf{m}$. The individual components of the latent function $\mathbf{h}$ are taken to be independent sample paths drawn from a Gaussian process with covariance function $k_h(t, t')$ and the components of $\mathbf{f}$ are independent draws from a Gaussian process with covariance function $k_f(\mathbf{h}(t), \mathbf{h}(t'))$, which determines the properties of the latent mapping.

To cope with limitations discussed above, we introduce a further control level (see Fig 1b) where $\widetilde{\mathbf{m}}$ and related covariances, say Σ_{td}, serve as top-down bias. To such end we introduce a state variable $\mathbf{r}$ and design a prediction/correct scheme in the form of the Kalman filter shaped as proposed in [12].

In our case the ordinary Kalman filter assumes a predicted observation

$$\overline{\mathbf{m}}(t) = \mathbf{H}(t)\overline{\mathbf{r}}(t) + \boldsymbol{\zeta}(t), \qquad \boldsymbol{\zeta}(t) \sim \mathcal{N}(0, \Sigma_{bu}), \tag{7}$$

with $\Sigma_{bu} = [\boldsymbol{\zeta}(t)\boldsymbol{\zeta}^T(t)]$ is the covariance of the "bottom up" noise $\boldsymbol{\zeta}$ affecting observations $\overline{\mathbf{m}}$. Kalman filter dynamics can be written as a prediction step followed by a measurement or correction step. State prediction can be written as

$$\overline{\mathbf{r}}(t+1) = \mathbf{A}\widehat{\mathbf{r}}(t) + \boldsymbol{\eta}(t) \tag{8}$$

where $\boldsymbol{\eta}(t) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{r}}(t), \Sigma_{\mathbf{r}}(t),\ \Sigma_{\mathbf{r}}(t) = E[(\boldsymbol{\eta}(t) - \boldsymbol{\mu}_{\mathbf{r}}(t))(\boldsymbol{\eta}(t) - \boldsymbol{\mu}_{\mathbf{r}}(t))^T]$. The evolution of $\overline{\mathbf{r}}$ goes together with covariance prediction $\mathbf{M}(t+1) = \mathbf{A}\mathbf{N}(t)\mathbf{A}^T + \Sigma_{\mathbf{r}}(t)$ and $\mathbf{N} = \mathbf{M}^{-1}(t) + \mathbf{H}^T\Sigma_{bu}^{-1}\mathbf{H}$ is a normalization matrix that maintains the covariance of the estimated state.

The update step corrects prediction by taking into account the measurement error

$$\widetilde{\mathbf{r}}(t+1) = \overline{\mathbf{r}}(t+1) + \mathcal{K}(t+1)(\widehat{\mathbf{m}}(t+1) - \overline{\mathbf{m}}(t+1)) \tag{9}$$

where $\mathbf{H}\overline{\mathbf{r}}(t+1)$ is the predicted measurement and $\mathcal{K}$ is the Kalman gain which is updated as $\mathcal{K}(t+1) = \mathbf{N}^{-1}\mathbf{H}^T\Sigma_{bu}^{-1}$.

The Kalman filter equation is obtained by combining Eqs. 8 and 9:

$$\widehat{\mathbf{r}}(t+1) = \mathbf{A}(\overline{\mathbf{r}}(t) + \mathcal{K}(t)(\widehat{\mathbf{m}}(t) - \overline{\mathbf{m}}(t))) + \boldsymbol{\eta}(t). \tag{10}$$

Set $\mathcal{K}_{bu} = \mathcal{K}$ $\mathbf{r}_{td} = \widetilde{\mathbf{m}}$ and define the top-down Kalman gain $\mathcal{K}_{td} = \mathbf{N}\Sigma_{td}$, Σ_{td} being the top-down covariance matrix provided by the upper-most level. Then the update step in Eq. 8 can be rewritten as

$$\begin{aligned}\widetilde{\mathbf{r}}(t+1) = \overline{\mathbf{r}}(t+1) + \mathcal{K}_{bu}(t+1)(\widehat{\mathbf{m}}(t+1) - \overline{\mathbf{m}}(t+1)) + \\ \mathcal{K}_{td}(t+1)(\widehat{\mathbf{r}}_{td}(t+1) - \overline{\mathbf{r}}(t+1)) - \mathbf{N}g(\overline{\mathbf{r}}(t+1))\end{aligned} \tag{11}$$

where the last term is a decay that penalizes overfitting of data and g an exponentially decreasing function. Eventually,

$$\widehat{\mathbf{r}}(t+1) = \mathbf{A}(\overline{\mathbf{r}}(t) + \mathcal{K}_{bu}(t)(\widehat{\mathbf{m}}(t) - \overline{\mathbf{m}}(t)) + \mathcal{K}_{td}(t+1)(\widehat{\mathbf{r}}_{td}(t) - \overline{\mathbf{r}}(t)) - \mathbf{N}g(\overline{\mathbf{r}}(t))) + \boldsymbol{\eta}(t). \tag{12}$$

3 Preliminary results

We focus on the behaviour of the observer's visuomotor simulation component when the motor-state space is controlled either by "raw" or by Kalman filtered parameters. We also compare for completeness with parameters obtained by a Kalman smoother, though this is unsuitable for online processing.

In the simulations, expresser's landmarks $\mathbf{y}_{\mathcal{E}}$ are inferred via the Constrained Local Neural Field (CLNF) [2]; a viable alternative is in [17] (or its sparse variants, e.g. [3]).

For the motor space representation $\mathbf{w}$ and its deformations we exploit the 3D face model Candide-3 [1], which is a 3D wireframe model of approximately 113

vertices $\mathbf{w}_i$ and 184 triangles, that easily fits our needs. Indeed, Candide directly accounts for encoding the matrices of Shape Unit (SU) and Action Unit Vector (AUV) deformations parameters at vertices ($\mathbf{dW}_i^S$ and $\mathbf{dW}_i^M$) together with related control parameters $\mathbf{s}$ and $\mathbf{m}$, respectively. AUVs determines a change in face geometry and implement a subset of the Ekman's Action Units of FACS [5]. The considered AUVs ($N_{AUV} = 11$) are $AUV_k, k = 0, 2, 3, 5, 6, 7, 8, 9, 10, 11, 14$. Observer's parameters $\mathbf{s}_{\mathcal{O}}$ are derived offline, and expresser's parameters $\mathbf{s}_{\mathcal{E}}$ inferred through the perceptual process at the very onset of the interaction.

As to Kalman based control, we consider the state variable as formed by position and velocity for all AUVs. Only the position vectors are eventually used to represent the motor action parameters. Parameter learning is performed via the EM algorithm. In the same framework, we also apply Kalman smoothing for comparison.

Due to limitations of space, we provide an excerpt of typical results so far achieved. Also, to provide clear clues to the reader these are related to motor trajectories of prototypical expressions (basic emotions), though the facial action space is a continuous manifold.

Fig. 2 shows the result of the Kalman filter and smoother, as well as the original motor parameters from the prototypical "disgust" emotion of a subject from the Cohn-Kanade dataset.

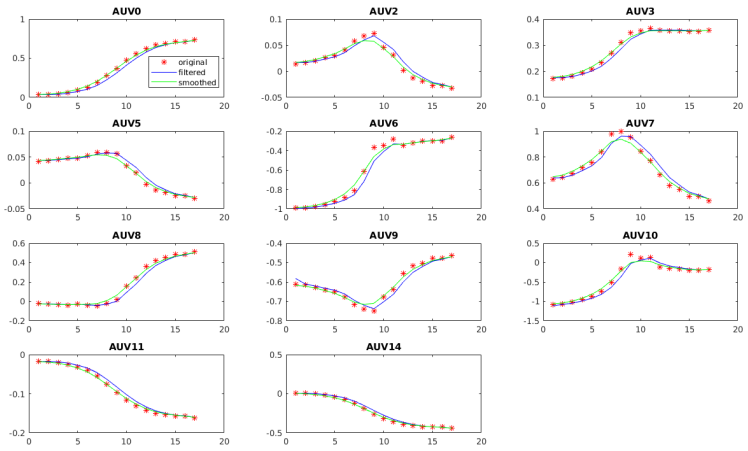
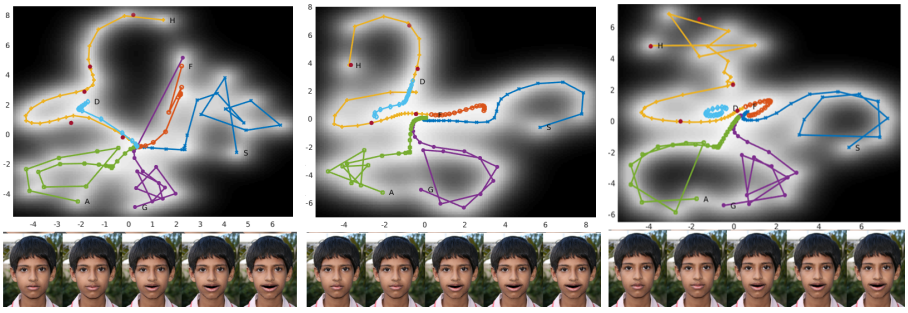


Fig. 2: Result of the Kalman filter (blue) and Kalman smoother (green) observations for each of the considered AUVs, related to the 'disgust' emotion.

Most important, is the latent action manifold as learned by adopting the different control schemes. One example is provided in Fig. 3, where basic emotion trajectories are shown within the GP-LVM latent space.



(a) (b) (c)

Fig. 3: Walking on the 'Happiness' trajectory. Top panels show the learned latent action spaces. To each red dot in top latent space corresponds facial synthesis (bottom panels). Latent space is learned by using raw motor parameters in 3a, Kalman filter state in 3b and Kalman smoother in 3c.

4 Conclusive remarks

We have discussed modelling issues that arise in the design of a somatic facial motor space for affective interactions. We have considered different levels of hierarchical control for the generation and learning of motor control parameters tuning the unfolding of the facial expression. Preliminary results show that it is important to evaluate parameter dynamics not *per se* but related to the construction and the dynamics of the latent action space. On the example provided, and similar to other results, the Kalman level seems, in general, to better separate and constrain trajectories as produced along discrete expressions. This is consistent with the idea that basic expressions originate as prototypes that cluster and partition continuous manifolds [13]. As expected, the Kalman smoother achieves smoother results, however it is unsuitable to provide online control. On the other hand, the direct implicit control via the action space could gain some currency as to the parsimony of such representation.

We surmise that conclusive arguments on the choice between one or the other scheme need to take into account, beyond the latent action space, the continuous manifold of the core affect.

Acknowledgments

This research was carried out as part of the project "Interpreting emotions: a computational tool integrating facial expressions and biosignals based shape analysis and bayesian networks", supported by the Italian Government, managed by MIUR, financed by the *Future in Research* Fund.

References

1. Ahlberg, J.: CANDIDE-3 An updated parameterized face. Tech. Rep. LiTH-ISY-R-2326, Linköping University, Department of Electrical Engineering, Linköping, Sweden (2010)
2. Baltrušaitis, T., Robinson, P., Morency, L.P.: Constrained local neural fields for robust facial landmark detection in the wild. In: Proc. of the IEEE Int. Conf. on Computer Vision Workshops. pp. 354–361 (2013)
3. Cuculo, V., Lanzarotti, R., Boccignone, G.: Using sparse coding for landmark localization in facial expressions. In: 5th European Workshop on Visual Information Processing (EUVIP). pp. 1–6 (Dec 2014)
4. Damianou, A.C., Titsias, M.K., Lawrence, N.D.: Variational inference for latent variables and uncertain inputs in gaussian processes. Journal of Machine Learning Research (JMLR) 17(1), 1425–1486 (2016)
5. Ekman, P., Rosenberg, E.L.: What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press (1997)
6. Fan, P., Gonzalez, I., Enescu, V., Sahli, H., Jiang, D.: Kalman filter-based facial emotional expression recognition. In: International Conference on Affective Computing and Intelligent Interaction. pp. 497–506. Springer (2011)
7. García, H.F., Álvarez, M.A., Orozco, Á.: Gaussian process dynamical models for emotion recognition. In: Bebis, G. (ed.) Advances in Visual Computing: 10th International Symposium, ISVC 2014, Las Vegas, NV, USA, December 8-10, 2014, Proceedings, Part II, pp. 799–808. Springer International Publishing, Cham (2014)
8. von Helmholtz, H.: Über integrale der hydrodynamischen gleichungen welche den wirbelbewegungen entsprechen. Crelles, J. 55, 25–55 (1858)
9. Lopes, M., Santos-Victor, J.: Visual learning by imitation with motor representations. IEEE Trans. on Sys., Man, and Cybernetics, Part B: Cybernetics 35(3), 438–449 (2005)
10. Orozco, J., Rudovic, O., Gonzlez, J., Pantic, M.: Hierarchical on-line appearance-based tracking for 3d head pose, eyebrows, lips, eyelids and irises. Image and Vision Computing 31(4), 322 – 340 (2013)
11. Pickering, M.J., Clark, A.: Getting ahead: forward models and their place in cognitive architecture. Trends in cognitive sciences 18(9), 451–456 (2014)
12. Rao, R.P., Ballard, D.H.: Dynamic model of visual recognition predicts neural response properties in the visual cortex. Neural computation 9(4), 721–763 (1997)
13. Russell, J.A.: Core affect and the psychological construction of emotion. Psychological review 110(1), 145 (2003)
14. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: A survey of registration, representation, and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 37(6), 1113–1133 (2015)
15. Vitale, J., Williams, M.A., Johnston, B., Boccignone, G.: Affective facial expression processing via simulation: A probabilistic model. Biologically Inspired Cognitive Architectures Journal 10, 30–41 (2014)
16. Wood, A., Rychlowska, M., Korb, S., Niedenthal, P.: Fashioning the face: sensorimotor simulation contributes to facial expression recognition. Trends in cognitive sciences 20(3), 227–240 (2016)
17. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Proc. of IEEE CVPR. pp. 2879–2886 (2012)