

2D Statistical Models of Facial Expressions for Realistic 3D Avatar Animation

Lukasz Zalewski and Shaogang Gong

Department of Computer Science, Queen Mary, University of London

London, E1 4NS, UK

{lukas,sgg}@dcs.qmul.ac.uk

Abstract

We address the issue of modelling facial expressions for realistic 3D avatar animation. We introduce a hierarchical decomposition of a human face into different components and model them according to their intrinsic functionalities. The parametrisation of the expressions is achieved in a two-level framework. First level accounts for the low level component facial actions and is represented by hierarchical latent variable models. The second level models the final expressions as a combinations of subcomponent information extracted from the lower level using combinatorial logic. Finally we produce continuous animation curves that are used to animate 3D avatar in a morph-based fashion. Our approach is entirely based on 2D information extracted from the input source.

1. Introduction

One of the most powerful and fastest means of relaying emotions between humans are facial expressions. Unfortunately the human face exhibits complex and intricate behaviours which are caused by changes in the facial muscle configurations. These changes depend on many factors such as current emotional state, implied context, surroundings. Facial expressions are individually independent: no two people exhibit the same expression the same way. These factors make modelling and recognising facial expressions a challenging task. Also as our eyes are susceptible to even smallest imperfections - correct parametrisation is a crucial step towards realistic avatar animation. With fast growing quest for realism and advances in the hardware development efficient and fast ways for conveying required information are as real as ever.

One of the most prominent systems and a blueprint for facial expression coding is FACS (Facial Action Coding System [11]). It defines the expressions as a combination of atomic action units corresponding to movements of particular muscle groups. FAP [14] coding system developed for

MPEG-4 standard defines set of facial landmarks placed on predefined location on the face and defines facial motions by creating displacements with respect to neutral state. All of them require very detailed information regarding facial feature landmarks configurations that might not be available at most times. Waters [18] introduced a pseudo-muscle model simulating contraction of real facial muscles which was controlled by FACS parameters. [16] extended Waters model to incorporate skin and fatty tissue of the face. Although very close to the underlying mechanics of human face those model are very complex, requires laborious rigging and therefore are not very suitable for general use. Morph-based techniques [13] has been used widely used and provide easy way to define expressions from set of bases in linear fashion, although deformations are limited to the number of predefined bases.

A lot of work has been done for expression classification [2, 8, 17, 5, 1]. Some of these systems do not deal with classification directly [2, 8] and are focused on the synthesis process, others treat the classification as a on/off state process [17, 5, 1] which is nowhere near acceptable for continuous and realistic animation.

For the animation to be realistic we also need the pose information as the heads position rarely stays at a particular view. Cootes *et al.*[6] used View-Based Appearance Model to find the relationship between the parameters and current pose. It requires building multiple appearance models for such task which presents an overhead that might not be desired in most cases. Dornaika and Ahlberg [10] combined 3D deformable model, statistical texture model and RANSAC paradigm. Both approaches require 3D priori information. Unfortunately none of the above map directly into the shape framework that we wish to use and their application would be an ad-hoc solution.

In this work we wish to model set of six basic expressions such as neutral, smile, grin, surprise/fear, anger and sadness and provide the extent or severity of each of the expressions in a continuous manner such that they can be used in the animation framework. The expression we wish to model are not defined as standard set of emotion targets

used in classification such as neutral, happy, angry, sad, surprised, frightened, disgusted. They are chosen in the way to provide the maximum visual impact (normally smile and grin would be defined as happy). We aim to model the intrinsic functionalities by placing hierarchical constraints to bootstrap the parametrisation process. We provide one unified statistical framework for such task. Our facial appearance under varying expression is based on a statistical appearance model originally introduced by [7]. We focus on person specific AAM, for robustness and given sparse training data. Also such model allows us to better represent intricate facial movements of an individual providing more stable tracking basis. We extend the basic definition of AAM model to implicitly incorporate pose variation into the statistical distribution. To bootstrap the tracking process and to enhance the parametrisation we equip our model with pose estimator which defines 6 DOF.

2. A 3D Animation Model

Realistic animation requires continuous and gradual changes between different facial expressions. We achieve that by producing a set of ROC (Rate Of Change) animation curves. This differs to classical emotion classification approaches which only provide discrete outputs (on/off state). We employ a morph based approach, in which every character is required to have a set of predefined morph bases corresponding to the expressions we wish to model. Then any expression E is given by:

$$E = \sum_i \mathbf{w}(i)\Gamma(i) \quad (1)$$

where \mathbf{w} defines a morph weight vector and $\sum_i \mathbf{w}(i) = 1$. Γ defines a set of morph bases corresponding to predefined expression states. Figure 1 shows an example of such morph bases. Although somehow limited in the sense of available freedom and requiring pre-rigged characters, such an approach offers several advantages. Firstly the representation is compact and independent of the animation engine giving us the ability to model non-human and human characters alike. Secondly complexity of the model and the number of parameters is relatively small compared to [15] which opens possibilities for real-time animation.

Pose is another crucial element to realistic animation. Our heads are in constant motion and those movements also convey emotional messages (contentment, inquisitiveness, nervousness). We estimate 6 DOF from 2D information by using pose estimator created within the probabilistic framework defined in (Section 3) but do not deal with parametrisation of expressions implied by head movements (their nature is implicitly incorporated into pose information).

Our expression parametrisation is based on hierarchical latent variable structure. Low level corresponds to two hier-



Figure 1. An example of morph bases (left to right) for neutral, smile, grin, fear, anger.

archical latent variable models constructed upon the hierarchical structure (Section 4), consisting of three components ($eye_L, eye_R, mouth$) which are modelled according to their intrinsic functionalities. On the conceptual level we represent expressions as a combination of functional states of the subcomponents such that expression E_c can be defined by:

$$E_c = state_{mouth} + state_{eye_R} + state_{eye_L}$$

Following the conceptual breakdown of the expressions the higher level fuses obtained information of subcomponents from low level to produce final classification. The severity of each expressions is produced by examining the cumulative probability density functions of the corresponding subcomponents. Finally smoothing is applied to obtained curves to remove any irregularities present.

3. 2D based 3D Pose Estimation

Pose estimator provides us with continuous 3D pose based on a probabilistic framework. We use a sparse set of training 2D samples, that covers only part of the view-sphere, $(-40^\circ, 40^\circ)$ around yaw and $(-20^\circ, 20^\circ)$ around pitch (using 10° intervals) and are able to estimate the pose for a much larger, continuous view-sphere to novel sequences. Additionally we do not need to utilise any temporal information such that the estimation is done on-the-fly frame-wise in real time and the system is able to cope with very large jumps and discontinuities in pose change. Our pose model is based on Probabilistic PCA [3].

Given any d -dimensional multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix \mathbf{C} its marginal q -dimensional marginal multivariate distribution (where $q \ll d$) is also Gaussian [12]. Let \mathbf{B} be a $q \times d$ dimensional matrix with diagonal elements set to 1 and the remaining ones equal 0. Then the marginal q -component multivariate probability distribution function (p.d.f) f_q is given by:

$$f_q \sim \mathcal{N}(\mathbf{B}\boldsymbol{\mu}, \mathbf{B}\mathbf{C}\mathbf{B}^T) \quad (2)$$

Following the concept of the marginal p.d.f we define the cumulative distribution function (c.d.f) Φ , such that for a q -dimensional random variable \mathbf{x} the c.d.f is given by:

$$\Phi(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} f_q(\mathbf{x}) d\mathbf{x} \quad (3)$$

We are mostly interested in the c.d.fs that are closely related to the components responsible for the yaw and pitch rotation. Let f_{mt} be marginal p.d.fs and Φ_{mt} be marginal c.d.fs corresponding to pose changes. For a given shape \mathbf{t} the estimate of the yaw and pitch rotations r_t where $t \in \{yaw, pitch\}$ is given by Equation (4), where $a_{1t}, a_{2t}, a_{3t}, a_{4t}$ are coefficients of a cubic polynomial estimated during the training stage using least squares, p_t is the marginal cdf and ϵ_t is the error term defined by constant weighted by the marginal probability f_{mt} for rotation of interest:

$$\begin{aligned} r_t &= a_{1t}p_t^3 + a_{2t}p_t^2 + a_{3t}p_t + a_{4t} + \epsilon_t \\ p_t &= \Phi_{mt}(\mathbf{t}) \\ \epsilon_t &= f_{mt}(\mathbf{t})const_t \end{aligned} \quad (4)$$

Roll is estimated using similarity transform [7]. To find the relationship between the angles and the c.d.fs, we use the posterior distribution of the PPCA model. We have found that such a probabilistic framework provides much more accurate estimation than one using conventional PCA (e.g. by finding the relationship between the projected parameters and angles). We performed some evaluation tests by removing set of training samples from the model building stage and using them as a test data. We obtained RMSE of 0.9241 compared to 1.7802 RMSE of PCA. The projection onto rotation shape-space is achieved by down-sampling the larger PDM to the required size.

4. A Hierarchical Expression Model

Instead of using holistic representation, we define expressions as a combination of intrinsic functionalities of the subcomponents (expression implied facial feature independence has been exploited by [9]). The advantages of this are two-fold: First of all, each of the expressions is defined in a more intuitive and quantitative way. Secondly, such a representation allows us to account for similar expressions (smile with eyes open, or smile with eyes closed) without any additional overhead. We define a hierarchical decomposition as follows: The jaw outline, nose and centres of the eyes and mouth form the root of our hierarchy. As leaves, or children, we have eye-eyebrow pairs and mouth. Figure 2 shows an example of such decomposition. The root component is used for estimating pose (Section 3). The leaves are used for expression modelling.

We choose shape component for our hierarchical representation. Our motivation is as follows: As the shape is individually independent (given appropriate normalisation) hence can be efficiently utilised to capture manifolds of the facial expressions. We experimented with combination of shape and texture and AAM parameters but found that shape alone provides the most optimal basis for expression

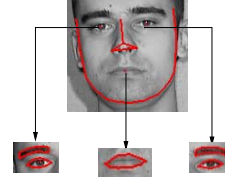


Figure 2. The top row corresponds to the highest point in the hierarchy (root), the middle row corresponds to the leaves.

modelling. This is caused by the lack of texture information which is susceptible to external factors such as illumination or identity changes.

5. Expression Parametrisation

Our main goal is to provide a compact parametrisation of expressions which can be subsequently used to animate the avatar. In doing so we try to avoid complexity and computational demands of FACS [11]. Such complexity is much desired for realistic animation. In the real world scenarios however such information might not be available or might be grossly inaccurate. Additionally we providing a compact and general model that can be applied to various scenarios.

Although in Section 4 we defined two eye components, only one statistical model is required to represent both. As both eye shapes are near symmetric we mirror right one along the vertical axis, apply all the appropriate normalisation and treat it as a left eye. Next we build the model for the left eye only using the training data for left eye and mirrored data for right eye. Although there will be very small discrepancies between the two (our face is not exactly symmetric and the range of motions are not the same either), such an approximation is desired as it will create model that will capture needed variations in a unified manner. Also such representation further simplifies the overall complexity of the model and provides much needed generalisation.

We noticed that the modes of variation for each of the components correspond to their intrinsic functionality. For example for mouth they are mouth open, mouth closed and mouth grin. To represent our data in the most discriminative manner we employ hierarchical latent variable model [3] for such task. Figures 3 and 4 show 2D visualisation of hierarchical structure for the mouth and eye models respectively. Each of the rows reflects the levels in the latent variable hierarchy. The dotted line connecting the plots between the levels means the plot has been copied down and doesn't contain any siblings at that level. Solid line represents the siblings on the current level of the hierarchy pointing to the parent at the previous level.

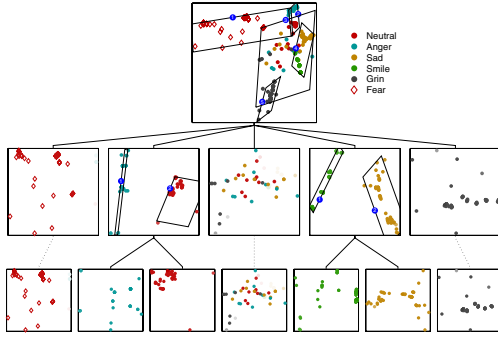


Figure 3. Hierarchical clustering in the mouth space. Colours correspond to the intrinsic functionalities of the components. Each of the rows depicts a level in the hierarchy.

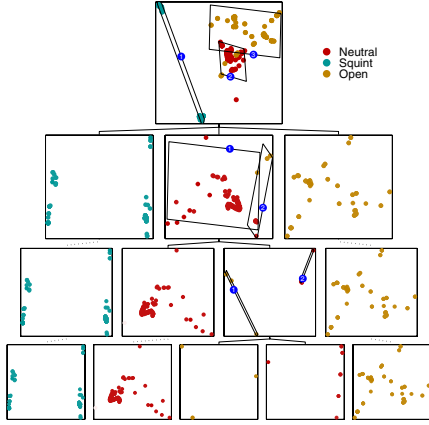


Figure 4. Hierarchical clustering in the eye space. Colours correspond to the intrinsic functionalities of the components. Each of the rows depicts a level in the hierarchy.

The final classification is performed by fusing the information obtained from low level hierarchical components together with their severities with combinatorial logic. Given discrete classification outputs for mouth (M), eyeL (EL) and eyeR (ER) the final expression F is defined as:

$$F = \begin{cases} \textit{smile} & M = 2 \\ \textit{grin} & M = 3 \\ \textit{fear/surp} & (M = 4 \wedge (ER = 2 \vee EL = 2)) \\ & \vee (M = 1 \wedge ER = 2 \wedge EL = 2) \\ \textit{anger} & (M = 5 \wedge (ER = 3 \vee EL = 3)) \\ & \vee (M = 1 \wedge ER = 3 \wedge EL = 3) \\ \textit{sad} & M = 6 \wedge (ER = 3 \vee EL = 3) \\ \textit{neutral} & \textit{otherwise} \end{cases}$$

5.1. Severity Criterion

We measure the degree of severity for each of the component states and expressions. This factor is crucial for realistic animation because a step based on/off parametrisation does not provide necessary continuity as the changes among the expressions are performed in smooth and gradual level. For the subcomponents severities that correspond to their low level behaviour the severity is defined in terms of cumulative distribution of the probability density function of the classified low level behaviour belonging to j -th hierarchical component where $j \in \{\textit{mouth}, \textit{eye}_L, \textit{eye}_R\}$. For the combined classification form if classified expression $\in \{\textit{smile}, \textit{grin}\}$ then the severity S is given by:

$$S = \Phi_j(\mathbf{x}_j) \quad (5)$$

where $\Phi(\mathbf{x}_j)$ is the the cumulative distribution of the probability density function of the classified expression component $j = \textit{mouth}$, otherwise for the expressions other than smile or grin the severity is given by the following linear combination:

$$S = \sum_j w_j \Phi_j(\mathbf{x}_j) \quad (6)$$

where $\Phi(\mathbf{x}_j)$ is the cumulative distribution of the probability density function of the classified expression component $j \in \{\textit{mouth}, \textit{eye}_L, \textit{eye}_R\}$ and w_j are weights such that $\sum w_j = 1$. As the popular belief would suggest the severity should be measured by calculating the Mahalanobis Distance (MD). Unfortunately this is not the case as we found out. Below are two sequences that demonstrate gradual change for two expressions grin (top row) and fear (bottom row). As we can see our Severity Criterion (SC) produces







			
SC	0%	41%	86%
MD	0%	0%	50%
			
SC	0%	24%	60%
MD	0%	0%	98%

Figure 5. Selected frames from two expressions demonstrating gradual change for grin (top row) and fear (bottom row) and the severities associated with them.

correct values compared to MD. This is caused by lack of symmetry in the PCA space and the fact that variation in each of principal components is a combination of many factors not just those caused solely by the expressions.

5.2. Animation Curves

To make the parametrisation complete we generate animation curves for predefined morph targets together with 6DOF pose information. The curves produced by SC are jagged and do not exhibit the smoothness much needed for seamless animation. To address this problem we apply smoothing filter to remove all sudden jumps and provide smooth curvature. Figure 6 shows the original jagged curve (left) and its smoothed out version (right).

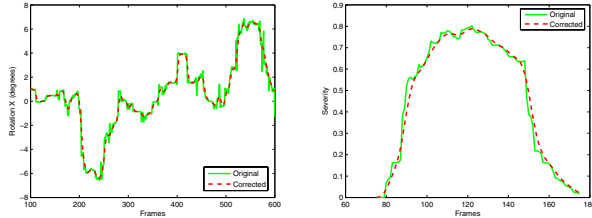


Figure 6. Parts of pitch rotation and grin SC plots: original jagged curves (green solid line) and their smoothed out versions (red dashed line).

6. Experiment

For AAM model training we used a set consisting of 1300 images and shapes (74 landmarks), which included seven basic expressions (neutral, smile, grin, sadness, fear, anger, surprise) and large variations in pose. We focused on person specific AAM (as opposed to generic AAM) for robustness under sparse training samples and also for providing better means for capturing intricate expressions therefore more realistic animation. For the hierarchical decomposition shape model training set of 400 shapes was used. Although we primarily deal with person specific expression parametrisation, our set contained mixture of selected shape samples from AAM training set and selected samples from CMU facial database [5]. This ensured wider range of different expressions being modelled, and accounted for unpredictability and ever changing facial motion for specific individual. For the pose estimator, 640 much sparser shapes (14 landmarks) were used. For the training of hierarchical latent variable models samples from hierarchical shape models were used. All the training samples were hand labelled beforehand. Figure 7 shows selected training samples from hierarchical model and pose estimator.

To test our system we used test sequences containing 30 repetitions of each of the expressions, totalling 10200



Figure 7. Selected training samples from the hierarchical model and and pose estimator.

frames. We compared our hierarchical model with a holistic representation model built in the same fashion and with Naive Bayesian Classifier (NBC) [4]. As our approach did not incorporate any temporal methods we didn't compare it with dynamic models such as MHMM [4]. We also run our system through sample sequences containing frames from the casual conversation (2850 frames). Due to the lack of standardised evaluation tests for continuous and gradual parametrisation the evaluation was based on the discrete basis (expressions on/off) as latter model does not define the severity of the expressions. We obtained the following classification rates:

	Hierarchical	Holistic	NBC
seq smile	79.47%	10.74%	5.34%
seq grin	87.85%	59.09%	19.30%
seq angry	17.70%	74.54%	9.24%
seq sad	61.26%	33.02%	24.20%
seq fear	82.73%	55.36%	16.96%
seq cas1	84.04%	17.15%	34.87%
seq cas2	84.85%	29.78%	30.41%

Figure 9 shows selected frames from an casual conversation sequences. Within each of the boxes the left image corresponds to the currently tracked image frame with the AAM mask superimposed on it. The image on the right corresponds to the synthetic avatar animated according to the classified expression. Figure 8 shows the corresponding classification results.

Finally to test the realism of our system we performed empirical tests by showing human test subjects animated sequences and asking them to rate the realism of the animation in the scale 1 – 10. As realism assessment is very individual specific and relays on the way we perceive the real world, human subject evaluation was the only just evaluation scheme available. From 16 subjects we obtained 75.9% success rate.

7. Discussion

In this paper we demonstrated that hierarchical facial decomposition can be efficiently utilised to capture the manifolds of human facial expressions. Such task can be

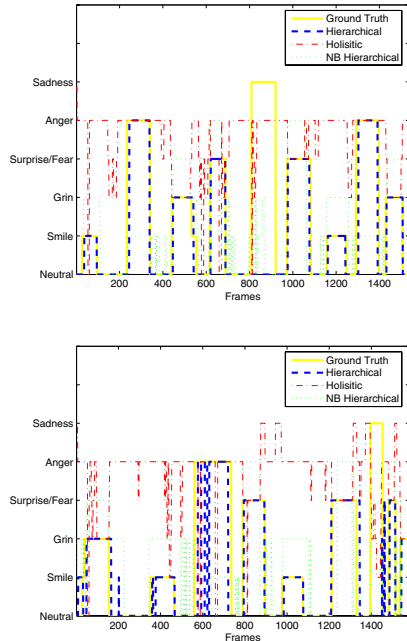


Figure 8. Expression classification for two casual conversation sequences.

achieved by modelling intrinsic functionalities of each of hierarchical components in a two-level framework. We also showed that simple models can be efficiently utilised to produce convincing level of realism. Also we avoided creation of complex animation models by embracing morph-based technique. Our future work involves parametrisation of more intricate expressions and speech modelling.

References

- [1] B. Abboud and F. Davoine. Appearance factorization for facial expression analysis. In *BMVC*, UK, 2004.
- [2] F. Bettinger, T. F. Cootes, and C. J. Taylor. Modelling facial behaviours. In *BMVC*, volume 2, pages 797–806, 2002.
- [3] C. M. Bishop and M. E. Tipping. A hierarchical latent variable model for data visualisation. *PAMI*, 20(3):281–293, 1998.
- [4] E. S. Chuang, H. Deshpande, and C. Bregler. Facial expression space learning. In *10th Pacific Conference on Computer Graphics and Applications*, Beijing, 2002.
- [5] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. S. Huang. Facial expression recognition from video sequences. *International conference on Multimedia and Expo*, 2:121–124, 2002.
- [6] T. Cootes, K. Walker, and C. Taylor. View-based active appearance models. In *IEEE FG*, pages 227–232, France, 2000.
- [7] T. F. Cootes and C. J. Taylor. Statistical models of appearance for computer vision. Technical report, University of Manchester, Manchester, UK, 2001.
- [8] V. E. Devin and D. C. Hogg. Reactive memories: An interactive talking head. In *BMVC*, 2001.
- [9] G. Donato, M. S. Barlet, J. C. Hager, P. Ekman, and T. J. Sejnowski. Classifying facial actions. *PAMI*, 21(10):974–989, October 1999.
- [10] F. Dornaika and J. Ahlberg. Efficient active appearance model for real-time head and facial feature tracking. *2003 IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pages 173–180, October 2003.
- [11] P. Ekman, W. V. Frieser, and P. Ellsworth. *Emotion in the human face*. Pergamon New York, 1972.
- [12] W. J. Krzanowski. *Principles of Multivariate Analysis*. Oxford University Press, 1988.
- [13] J. Noh and U. Neumann. A survey of facial modeling and animation techniques. Technical report, University of Southern California, 1998.
- [14] J. Ostermann. Animation of synthetic faces in mpeg-4. *IEEE Computer Animation*, pages 49–55, 1998.
- [15] F. Parke. *A Parametric Model for Human Faces*. PhD thesis, University of Utah, Salt Lake City, Utah, December 1974. UTEC-CSc75-047.
- [16] D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE PAMI*, 15(6):569–579, June 1993.
- [17] Y. Tian, T. Kanade, and J. F. Cohn. Recognising action units for facial expression analysis. *PAMI*, 23(2):1–19, 2001.
- [18] K. Waters. A muscle model for animating three-dimensional facial expression. *Computer Graphics*, 21(4):17–24, 1987.

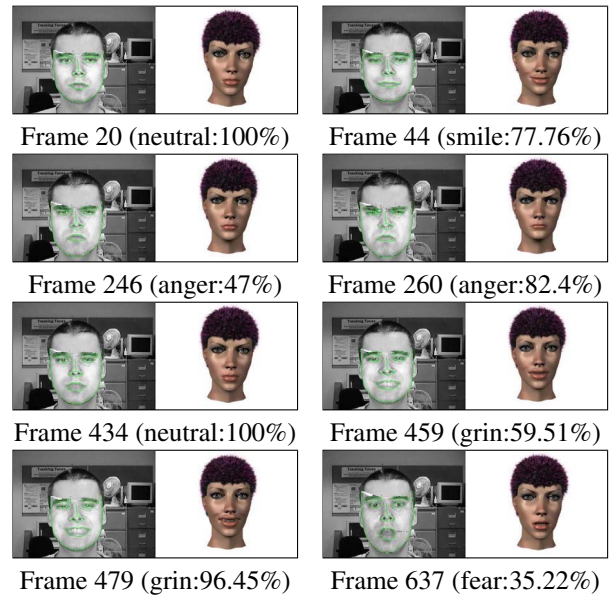


Figure 9. Selected frames from the experiment on expression classification and avatar animation with corresponding labels and severity (in percentage). Each of the images shows tracked frame with AAM mask superimposed on it (left) and corresponding synthesised avatar (right).