CoArt:
Co-articulation Region Analysis for Control of 2D Characters

Douglas Fidaleo and Ulrich Neumann
University of Southern California
{dfidaleo | uneumann}@graphics.usc.edu

Abstract

A facial analysis-synthesis framework based on a concise set of local, independently actuated, Co-articulation Regions (CR) is presented for the control of 2D animated characters. CR's are parameterized by muscle actuations and thereby provide a physically meaningful description of face state that is easily abstracted to higher-level descriptions of facial expression. Independent component analysis on a set of training images acquired from an actor is used to characterize the appearance space of each CR. Within this framework, actor-independent face reconstruction databases can be created by an artist or extracted from video sequences. In addition, the muscle parameter values may be used to drive any similarly parameterized 3D facial model. The flexibility afforded by such a methodology is demonstrated with applications to 2D facial animation control and sample based video synthesis. The analysis runs in real-time on modest consumer hardware.

Keywords: Facial Analysis, independent component analysis, animation control, muscle actuation analysis.

1. Introduction

From the early beginnings of facial animation, biologically inspired parameterizations have been explored for control of animated faces. The set of facial muscles and their actuations provides an intuitive basis for human face state-space, and consequently, the majority of this work has focused on mapping facial motion data to virtual muscles fit to a 3D model. In [12] Essa uses dense optical flow features coupled with a physical model of the face to estimate muscle contraction profiles. Optical flow is also used in [8] for estimation of shape and motion parameters of a deformable face model. Extraction of optical flow features is computationally expensive. Sparse point motion increases the computational tractability of the problem but to ensure stability, physical markers are generally fixed to the face. Subtle details of skin dynamics such as wrinkling are lost in the sparseness.

While muscles are an intuitive parameterization, it is difficult to visualize their effects on the surface of the skin. This is exemplified in recent work by Choe et. al. who used hand generated muscle actuation basis models both to estimate muscle actuation parameters and to produce excellent quality 3D facial animation [5]. An initial set of muscle models was hand sculpted by an artist but an optimization phase was required to accurately reflect the muscle appearances.

In the psychology community, an alternative parameterization of the face based on the static visual manifestations of muscle contractions is widely used for facial analysis; namely, the Facial Action Coding System (FACS) [11]. Devised by Ekman and Friesen, FACS identifies 46 independent areas of facial motion called Action Units and establishes a methodology for the manual estimation of AU activity from static images of facial expression. Automated approaches to AU estimation have been explored in [6, 9, 15].

For AU data to be useful in animation, a fine level of actuation intensity must be determined, however research to date has been focused on binary classification (if an AU is on or off) [6, 9]. Lien et al analyzes intensity for full-face expressions by correlation to existing training images, but AU level intensity is not estimated [15]. The large quantity and relatively small size of AU's makes it a cumbersome interface for creating and editing facial animation. In addition, it is difficult to interpret the actuation of a single action unit as it is a 1D parameterization of a space whose state can be changed by multiple muscles. As noted in [20] this 1D parameterization is insufficient for full facial characterization.

Our approach to facial analysis is a blend between muscle and action unit analysis. We introduce the notion of a Co-articulation Region (CR) that is analogous to a FACS AU augmented with muscle actuation information. A CR is a concise, intuitive, and physically meaningful

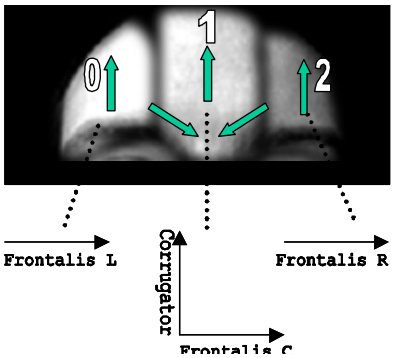


Figure 1. Illustration of selected co-articulation regions and their basis muscles.

facial parameterization. CR analysis with independent component analysis provides a fine level of actuation intensity. The parameters are actor independent and their analysis can be performed in real-time on modest consumer hardware. No encumbering facial markers are required. While the muscle actuations themselves may be independent, the appearance on the surface of the skin may not be (especially in early stages of muscle contraction). We use ICA to further refine the underlying independent image components contributing to the appearance of a given muscle contraction on the surface of the skin.

The collection of CR's intrinsically defines a discrete facial state space that allows an artist to easily create new characters for flip-book style animation. Character creation consists of populating the state space with local face appearance data. Character appearance data can also be extracted from the original training video images for Video Rewrite [2] style sample based video synthesis. As the parameter space is identical for each actor, the created characters are actor independent.

2. Previous Work

2.1. Animation control approaches

Direct approaches to visual animation control not based on physical parameterizations have been explored. Facial motion capture data is used to directly interpolate 3D models using radial basis functions in [13]. Noh transfers motion vectors from existing animation to models with different geometry in [19]. In [1] statistical analysis is performed on synchronized facial motion capture data and digitized speech. Facial animation (2D and 3D) is produced from entirely from speech input. Full-face analysis of expressions is used to interpolate base 3D models in [16].

Physically parameterized animation is performed in [5] where muscle parameters are extracted from sparse point flow and used to compose muscle basis geometry for 3D character animation. Terzopoulos and Waters [21] track contour features on eyebrows and lips to animate their physically based muscle structure of a synthetic character. An approach for model based coding by Eisert et.al. [10] combines MPEG4 facial action parameter (FAP) estimation with traditional block based coding of head and shoulder video sequences. Dynamic facial features such as wrinkles and eye blinking are block coded at each frame. Buck tracks a set of facial features and warps hand drawn images for non-photorealistic 2D facial animation.

CoArt is similar to approaches by [2, 7] that rearrange video samples extracted from footage acquired of speaking people to generate animation. The CoArt system, however, as it derives all information from instantaneous image data is non-causal.

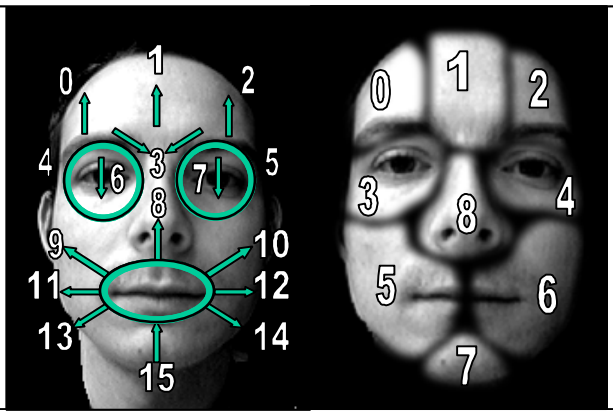
2.2. Facial Analysis

Facial gesture unit analysis considers the parameterization of the face by a discrete set of independently actuated units. These units take the form of Action Units (AU) in the facial action coding system (FACS) by Ekman [11] and FAP's in MPEG 4. Low-level units of facial expression defined by the FACS AU's are analyzed using HMM's trained on dense optical flow data and gradient information in [6] and ICA in [9]. Essa uses optical flow and a finite element model of the face surface to estimate a set of actuation parameters called FACS+ [12] that encapsulates gesture velocity information. Local PCA kernels are used in the eigenfeatures [24] approach that is perhaps most similar to our analysis methods.

Linear models for representing facial gestures, speech, and head pose based on PCA have been explored extensively in the vision literature. A general framework for analysis and synthesis using example based linear models is explored in [23] with applications to facial pose and gesture modeling. Example based learning is also used in [3] for represenation of prior knowledge of faces under pose variations, and subsequently for facial recognition. We use a similar state space discretization for local coarticualtion regions. Okada builds a piecewise linear subspace model of pose variations from images using PCA [18]. PCA is also used in [20] for modeling the space of facial speech parameters and applied to 2D and 3D mouth animation.

In these methods, analysis *and* synthesis are performed using the PCA basis. For our application, as we are associating artist generated image appearances with discrete facial gesture states, classification rather than reconstruction is of primary importance. Donato et al have empirically shown that ICA performs better at classification of facial gestures using the FACS

0. Frontalis L
1. Frontalis C
2. Frontalis R
3. Corrugator
4. Orbicularis Occuli L
5. Orbicularis Occuli R
6. Levator Palpebrae L
7. Levator Palpebrae R
8. Levator Nasii
9. Zygomatic Major L
10. Zygomatic Major R
11. Risorius L
12. Risorius R
13. Triangularis L
14. Triangularis R
15. Mentalis



CR	Muscles
0	0
1	1, 3
2	2
3	4, 6
4	5, 7
5	9, 11, 13
6	10, 12, 14
7	15
8	8

Figure 2. (left) List of muscle groups and their locations and directions of contraction. (right) Co-articulation regions and the list of influencing muscles.

representation [9]. We therefore use ICA to model the space of changes in each CR.

If we consider a single person's face, many of the changes that can occur do so locally, and hence a more concise representation can be obtained by constructing local subspaces that encapsulate these changes. This is the approach used in the eigenfeatures technique that uses PCA for local basis construction [24]. In the CoArt system, we build local subspaces using ICA and parameterize these subspaces with the muscles acting upon the associated regions.

3. Co-articulation Regions

Each facial muscle group listed in Figure 2 is capable of contracting independently and causing secondary motion of the skin in a continuous and local area on the face. This area is defined as the muscle's region of influence (ROI). The state of each muscle is parameterized by a contraction value between 0 (relaxed) and 1 (fully contracted). When a muscle is actuated independently, the changes that are propagated to the skin surface are local and fully determined by the muscle's level of contraction. When two or more muscles have an overlapping region of influence, the resulting skin change is a combination of the affects of the involved muscles. We exploit this fact and model the changes occurring on the surface of the face as a set of contiguous local regions of skin deformation called *co-articulation regions* (CR).

A CR is analogous to a FACS AU but is more specifically defined as a nonempty intersection of n muscle regions of influence. The special case of a CR where n=1 is called an Independent Region (IR). In the CoArt system the activation level of a CR is defined by the n-tuple of activation levels of each muscle contributing to the region's deformation. (In contrast with a FACS AU which defines a 1D activation level for each AU.) Each CR defines the state space for a local region on the face whose dimensionality is given by the number of muscles acting on the region. This CR subspace parameterization allows us to map directly from a point in CR appearance space (the space defined by all possible image samples in a given CR), to a point in muscle space.

The initial set of muscle ROI's is extracted from difference images of the user performing the gestures in figure 3. The intersections are calculated and contiguous regions are collected. The resulting regions are manually merged to produce nine CR's. The state space of a CR is defined by the degrees of freedom of the muscles acting upon the region as shown in figure 1.

4. System Overview

All input images must be normalized with respect to the actor's head pose and environmental lighting conditions prior to analysis. The actor's head frame is defined by a set of four infrared LED's fixed to a pair of empty glasses frames. The LED's are tracked and the image subsequently warped to a canonical pose. The warped image is convolved with a Laplacian kernel provide lighting intensity normalization.

An actor undergoes a training phase in which his/her CR state space is discretized and appearance properties are learned. 2D Character creation is performed offline, and consists of defining a character specific CR space discretization and assigning visual appearance samples for the neutral face support and each discrete CR state.

CR's are analyzed in a new image and a CR state vector is extracted. With the assumption of a neutral state outside of the CR set, an actor's face state is represented by the collection of individual CR states. The state vector indexes a given character's appearance samples that are composited to generate a frame of an animated sequence. Alternatively, as each CR is parameterized by a set of muscles, a muscle contraction vector can be extracted and used to drive muscle parameterized 3D faces.

5. CR Training

A basis for each CR state space is created using independent component analysis of a discrete set of image samples. Training samples are acquired for the basis actuations by instructing the user to perform the specific facial gestures in Figure 3. These gestures are chosen to independently actuate 15 major muscle groups. Each gesture is recorded from the first non-neutral frame indicating gesture motion to the first frame of maximum actuation (this is referred to as a *bounded gesture*). The video frame rate and actuation velocity determines the number of frames in a given bounded gesture, and hence, the quantization resolution of the state space. Figure 3 shows the gestures and the muscles for which they provide appearance information. A set of muscle actuation samples is extracted by sampling the bounded gesture frames at the appropriate CR locations and an image mask is applied at for each region. Each sample associated with a given muscle gesture is transformed by horizontal scan to a column vector and concatenated to form a matrix $X_i = [x_0 x_1 \ldots x_k] | i = 0 \ldots 15$. The subspace S associated with CR j is defined by appending muscle gesture matrices, $S_j = \bigcup_{i \in CR^j} X_i$. From S we compute a *signature basis* B using ICA which is in turn used to compute the muscle signatures.

Gestures	Muscles
0. Forehead Raise	0, 1, 2
1. Eye Squint	4, 5
2. Eye blink	6, 7
3. Smile	9, 10
4. Frown	13, 14
5. Cheek Pinch	11, 12
6. Furrow	3
7. Chin crinkle	15
8. Snarl	8

Figure 3: Muscles involved in training

5.1. Signature basis construction and signature extraction

The ICA model as applied to CR analysis assumes the presence of a set of statistically independent signature space basis vectors B (the "independent components" of S) whose composition by a "mixing matrix" A gives rise to the training samples S.

$$S = AB \tag{1}$$

Though in practice, neither A nor B are known, by assuming the columns of B have non-Gaussian distributions they can be estimated with the fixed-point iteration scheme used in the FastICA algorithm [14]. The set of independent sources B_j derived from the set of muscle appearances S_j is used as the signature space basis. A new sample x is transformed to CR signature space by

$$c = B_j^{-1} x \tag{2}$$

As this is an over determined system of linear equations, c is estimated by least squares approximation where $\hat{c}$ minimizes the residual $r = B_j \hat{c} - x'$. The *muscle signature* σ of x is computed as $\sigma = \frac{\hat{c}}{\| \hat{c} \|}$. Let $\psi(x, B)$ represent the function mapping a sample x to muscle signature σ as described above and compute the muscle signature sets $\Phi_i = \psi(X_i, B_j)$ where muscle i belongs to CR j according to figure 2.

5.1.1. CR state vector computation. Ideally we would like to analyze a new incoming sample from a CR and determine its exact location in CR state space thereby determining the contributions from each muscle basis vector. This turns out to be extremely difficult, as the resulting visual mixture is a nonlinear combination of skin appearances. The problem is simplified by assuming there is a single active muscle in a CR at any point in time.

For a given input sample x, the states ω and scores α for muscles associated with CR j are computed as:

$$\omega_i = \arg\max_n (\psi(x, B_j) \bullet \Phi_{m_i}(n)) \tag{3}$$

$$\alpha_i = \psi(x, B_j) \bullet \Phi_{m_i}(\omega_i) \tag{4}$$

If $c = \arg\max_i \alpha_i$ a single muscle actuation is forced by setting $\omega_i = 0 | i \neq c$. CR states are collected into the *muscle state vector* v and v_i is normalized by the size of the training set acquired for muscle i. As each element in Φ_i is ordered by its temporal progression from neutral to full actuation, v_i is interpreted as the actuation level of muscle i.

By assuming a single active muscle, the analysis does not account for the co-articulation ability of muscles in the CR. For example in a sad expression, the center frontalis and corrugator muscle groups are frequently mutually actuated. A preliminary solution to this problem is to treat the co-articulation gestures as *pseudo-muscles* by acquiring these samples in training and assign the bounded gesture frames to a new basis vector of the CR. Classification proceeds as described. A drawback of this approach is that it puts extra burden on the actor during the training data acquisition phase and limits

classification of co-articulation effects to those seen during training. Future work will involve separation of the underlying muscle contributions from a region sample given only the pure muscle training samples.

5.1.2. Singular value decomposition for real-time signature extraction. Because B is very large, iterative least squares methods are prohibitively expensive for real-time, online signature extraction. Instead, during training the SVD back substitution matrix M is computed which guarantees the least squares solution to (2).

Given D, the diagonal matrix consisting of positive or zero singular values of B, and orthogonal matrices U and V:

$$B = UDV^T \quad (5)$$

is the singular value decomposition of B and $M = VD^{-1}U^T$. For each new sample vector x, $\hat{c}$ is given by $\hat{c} = Mx$. M is constant following the ICA basis computation and may therefore be assembled once, and stored for efficient online signature extraction.

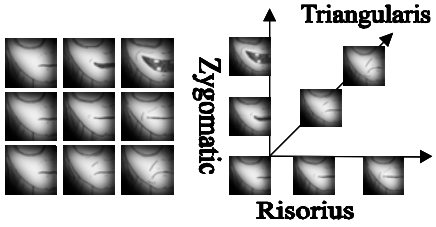


Figure 4: Reconstruction samples and placement for CR 5 of a 2D character.

6. Character creation

As there is a direct mapping from a CR state to its visual appearance, a new character is defined by creating a neutral face frame and assigning an explicit reconstruction sample for each CR state. Samples can take the form of hand drawn frames for the case of cartoon animation, or may be populated from the original training samples. If training samples are used, the region under each CR is used as a mask to extract reconstruction pixel data from each discrete CR state. Figure 4 shows a set of appearance samples assigned to CR 5 in the construction of the character "Maggie" as well as the original samples used for video reconstruction.

Each CR state vector applied to the character definition defines a set of reconstruction elements and alpha masks. Alpha masks are the same size as the reconstruction elements and have a value of 1.0 everywhere with a 10 pixel gradient to 0 approaching the boundary. Reconstruction elements are composited into the final image using simple alpha blending with the neutral image as a base. Animation is thereby performed by reconstructing frames using the stream of CR state vectors.

7. Results

Figure 5 shows results of the classification system applied to a 2D character and to the video reconstruction set acquired from the same actor. Input images consisted of 360x240 pixel grayscale images. Lighting conditions were kept constant throughout the performance. The classification and reconstruction is performed in real-time. Note that multiple CR's are concurrently analyzed and reconstructed. The same training data are for both sequences demonstrating the ability to mix and match reconstruction databases.

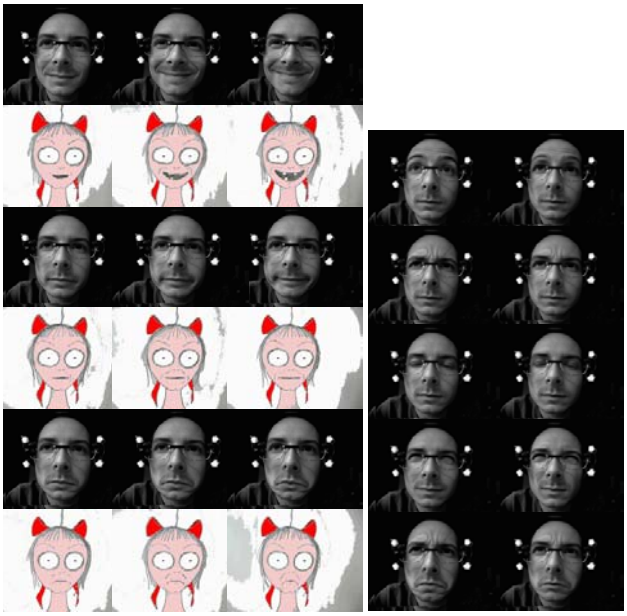


Figure 5. (left) Controlling a 2D character. (right) Reconstruction using original video samples.

8. Discussion and Future Work

Local actuation regions proposed by FACS and others provide an intuitive partitioning of the space of facial motion. Unfortunately such approaches favor many small areas of change that become difficult to manage and interpret independently. Muscle based interfaces where a small number of primary muscle groups are considered are manageable, but the correlation to changes on the face can be difficult to visualize. Co-articulation region parameterization merges the benefits of these two representations. CR analysis by ICA proves to be efficient and robust. Independence between the actor and controlled object is maintained by mapping through a single set of control parameters.

Generating region based flip-book style animation with the CoArt system has been demonstrated. The CoArt system can alleviate the tedium of manual frame arrangement for 2D face animation by allowing an actor to control the gestures using his/her own face. Video footage of the actor is analyzed in real time and used to control expressive gestures. Analysis and synthesis is performed at the region level, and therefore allows for assembly of gestures that were not explicitly constructed by the animator.

For complete characterization of a CR, co-articulation data must be extracted. Future work will involve separation of multiple muscles from a single CR sample. Blending of intermediate states and interpolation between states at CR boundaries is necessary to curtail potential reconstruction artifacts. The correlation of image samples to muscle states is largely dependent on the actor's ability to control their facial gestures. This is a difficult task for untrained individuals. We are exploring other methods for extracting muscle state label information during training. Though the results show promise, we intend to do more thorough testing on a variety of subjects. We also intend to compare the results of ICA with a PCA version of CoArt.

9. Acknowledgements

Thanks to J.P. Lewis, Kazunori Okada, and Antonio Ortega for their time and insight into various aspects of the work. Funding and research facilities were provided by the NSF through its ERC funding of the Integrated Media Systems Center.

10. References

[1] M. Brand, "Voice Puppetry", *SIGGRAPH 99 Proceedings*, 1999, 21-28

[2] C. Bregler, M. Covell, M. Slaney, "Video Rewrite: Driving Visual Speech with Audio", *SIGGRAPH 97 Proceedings*, 1997, 353-360.

[3] D. Beymer, A. Shashua, and T. Poggio. Example Based Image Analysis and Synthesis. *A. I. Memo 1431*, Massachusetts Institute of Technology, November 1993.

[4] Buck, I., Finkelstein, A., Jacobs, C., Klein, A., Salesin, D. H., Seims, J., Szeliski, R., and Toyama, K. Performance-driven hand-drawn animation. *Proceedings of NPAR 2000* (June 2000).

[5] B.W. Choe, H.S. Ko, Analysis and Synthesis of Facial Expressions with Hand-Generated Muscle Actuation Basis, *Proceedings of Computer Animation 2001*, November 2001

[6] J.F. Cohn, A.J. Zlochower, J. Lien, Y. Wu, T. Kanade, "Automated Face Coding: A Computer-Vision Based Method of Facial Expression Analysis", *Psychophysiology*, vol. 35, no. 1, pp. 35-43, 1999.

[7] E. Cosatto and H.P. Graf, "Photo-Realistic Talking-Heads from Image Samples", *IEEE Transactions on Multimedia*, pp. 152-163, 2000.

[8] Douglas DeCarlo and Dimitris Metaxas. Deformable Model-Based Shape and Motion Analysis from Images using Motion Residual Error. *ICCV Proceedings*, 1998, pp. 113-119, 1998.

[9] G. Donato, M. Bartlett, J. Hager, P. Ekman, T. Sejnowski, "Classifying Facial Actions", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, No. 10, October 1999.

[10] P. Eisert, B. Girod, "Analyzing Facial Expressions for Virtual Conferencing", *IEEE CG & A*, 1998, 18(5), pp. 70-78

[11] P. Ekman, W. Friesen, "Facial Action Coding System: A Technique for the Measurement of Facial Movement". Palo Alto, CA, *Consulting Psychologists Press*, 1978.

[12] A. Essa, Analysis, "Interpretation, and Synthesis of Facial Expressions", *PH.D. Thesis*, MIT, 1995.

[13] B. Guenter, C. Grimm, D. Wood, H. Malvar, F. Pighin, "Making Faces", *SIGGRAPH 98 Proceedings*, 1998, 55 – 66

[14] A. Hyvärinen. "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis", *IEEE Transactions on Neural Networks* 10(3):626-634, 1999.

[15] J.J. Lien, T. Kanade, J. Cohn, and C. Li, "Subtly Different Facial Expression Recognition and Expression Intensity Estimation", *CVPR Proceedings*, June, 1998, pp. 853 - 859.

[16] H.C. Lo, R. Chung, "Facial Expression Recognition approach for performance Animation, Proceedings", *Second International Workshop on Digital and Computational Video*, pp. 132-139, 2001.

[17] K. Mase, "Recognition of Facial Expression from Optical Flow", *IEICE Transactions*, vol 74, no. 10, 1991.

[18] K. Okada and C. von der Malsburg, Analysis and Synthesis of Human Faces with Pose Variations by a Parametric Piecewise Linear Subspace Method, *CVPR Proceedings*, pp. I:761-768, Kauai, USA, December 2001.

[19] J.Y. Noh, U. Neumann, "Expression Cloning", *SIGGRAPH 01 Proceedings*, 2001, 277-288

[20] L. Reveret, I. Essa, Visual Coding and Tracking of Speech Related Facial Motion, *GVU Center Tech Report* No. GIT-GVU-TR-01-16, 2001.

[21] D. Terzopoulos and K. Waters, "Analysis of Facial Images Using Physical and Anatomical Models," *ICCV Proceedings*, pp. 727-732, Dec. 1990.

[22] M. Turk, A. Pentland, "Eigenfaces for Recognition", *Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-86, 1991.

[23] T. Vetter and T. Poggio. "Linear object classes and image synthesis from a single example image". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):733--742, 1997.

[24] WJ. Welsh and D. Shah, "Facial feature Image coding using Principle Components", *Electronics Letters*, October 1992, Vol. 28, pp 2066-2067.