MVA2005 IAPR Conference on Machine VIsion Applications, May 16-18, 2005 Tsukuba Science City, Japan
3-26

Combining Computer Graphics and Image Processing for Low Cost Realistic 3D Face Generation and Animation

Alexander Mark Woodward, Patrice Delmas
CITR, University of Auckland, Dept. Computer Science
patrice@cs.auckland.ac.nz

Abstract

Realistic 3D human faces have found application in a myriad of different computer situations. The easily recognisable human face provides an emotive bridge for the personalization of many human-computer interactions in todays society. Face models based on real people are thus widely sought after. This paper presents computer vision techniques for low cost acquisition of raw data, a computer graphics model for facial animation, and an interface between the two which binds the process into a viable system for facial synthesis and animation for a wide range of environments and audiences. The presented solution is currently running on standard hardware.

1 Introduction

Facial animation has a wide range of computer applications such as the gaming and cinema industries, medicine, social agents and avatars. Thus a great deal of work has been employed to advance these various techniques and applications [1]. As commodity hardware increases in performance, a renaissance in more advanced and physically accurate techniques for facial animation will come to the fore.

Although facial animation is the domain of computer graphics, there is a need for the acquisition of accurate 3D data from real human subjects. In the absence of sophisticated, often expensive, scanning equipment, low cost techniques must be explored. Binocular stereo is a strong choice since it has readily available texture data, is passive to the scene, and 3D positions can be reconstructed from the acquired depth map with minimal manual interaction.

In this paper, the process of acquiring 3D data is first presented. Stereo vision techniques are utilized to acquire dense depth map information of the face frontal view. Quasi-automatic face features extraction follows, combining boosted classifier filters for rough face detection and active appearance models for accurate facial feature delineation. Using correspondence between the location of 2D face features and their position on a generic 3D face model allows surface registration. Then, the mapping between the generic model and the acquired depth data is achieved to obtain an animatable 3D reconstruction of the test subject with minimal user intervention. Finally, the facial animation system, based on physics based model of human tissue, is briefly described and some results are presented.

2 Acquisition of 3D Data using Binocular Stereo

The human face provides a unique challenge for stereo vision. Within a single image, many areas of differing texture, surface reflectance properties, and colouration may appear. For human faces, a dense disparity map for 3D reconstruction purposes is required. For a comparison specifically regarding algorithm operation on human faces, see [3]. The KZ1 stereo-matching algorithm was chosen for its relatively robust inference scheme on a wide variety of signals that may compose a facial image (for more details on the KZ1 algorithm, see [1]).

2.1 Lab Setup

A stereo rig was setup consisting of two Canon A80 cameras with standard epipolar geometry. The cameras were chosen for their embedded PC-controlled acquisition software and were placed in a vertical configuration as depicted in Fig.1 to take into account the horizontal symmetry of human faces, which can significantly degrade the result of stereo correspondence between pixels [3].

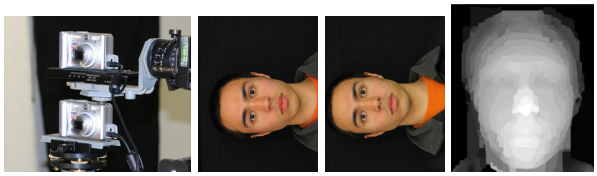

Figure 1. The From left to right: The camera setup, the resulting image pair in vertical epipolar position and the depth map obtained using KZ1.

3 Interface Between the Raw Data and the Facial Animation System

The interface between the raw data and its integration into a facial animation framework involves the selection of a few important landmarks on the raw data, and matching them to their equivalent positions on a generic model possessing our preconfigured animation system. Although manual selection of about 20 landmarks have been implemented and requires only a few minutes to properly register the raw data to the generic 3D face model, automatic registration could limit human intervention. To emphasize both face features contour detection robustness and accuracy, it has been found that boosted classifiers filters followed by active appearance models [8] could capture the face location as well as the delineation of the jaw, eyes, eyebrows, inner and outer lip contours.

4 Face Feature Detection using a Cascade of Boosted Classifiers

This method uses a cascade of classifiers (single layer pereptrons), that are trained using a feature representation set and a set of positive and negative sample images depicting a desired feature [4]. The cascade of classifiers is termed a *detector*. It should be noted that a *feature set* refers to a specific feature representation, or invariant that can be used to describe a certain object such as the head, eye, mouth or nose. The feature representation uses Haar-like basis functions to describe *edge*, *line*, and *centre-surround* features.

A feature value is calculated by the weighted sum of the pixels covered by the white rectangles subtracted by the weighted sum of the pixels covered by the dark rectangles. These features allow for rotation and scaling. Such a feature representation system can operate much faster than one that evaluates pixels alone.

The cascade allows for early rejection of regions of the image that do not possess the feature being searched for. Simpler classifiers are used to reject the majority of subwindows before more complex classifiers are called upon to achieve low false positive rates. A negative outcome at any point leads to the immediate rejection of the sub-window, which allows for a strong speed improvement when searching.

The boosted nature of the classifiers relates to the composition of a strong classifier (one which gives a very strong prediction to a given input) out of a linear combination of weak classifiers (classifiers whose prediction rule is slightly better than random guessing) [7].

The implementation used in this thesis is part of the OpenCV library [5], whose code module for detection is based on the work presented in [4] with such performance to allow real-time tracking of objects in video streams.

4.1 Active Appearance Models

Active Appearance models are generated by combining a model of shape variation with a model of the appearance variations in a shape normalised frame.

First, a training set of labelled images containing the key landmark points (as defined in the generic shape model) are marked, e.g. for a face: features such as the eyes, mouth, nose, jaw, eyebrows have to be delineated. Applying principal components analysis on the above created database, the generated statistical models of shape and of grey level appearance as well as their respective modes of variation are obtained.

When matching a model to an unseen image, the aim is to minimise the residual error between the grey values in the image and the AAM by progressively adjusting the model parameters until a minimum is found. By first using boosted classifiers filters for rough face location, we can approximate the location of an unseen face and its features. The obtained accuracy is then proportional to the variation of faces included in the database.

Figure 2. From left to right: face rough detection using boosted classifier filters, Active Appearance models applied on AAM training set face image and on Alex.

A combination of boosted filters classifiers and ac-

tive appearance models, can extract contours for most face features while preserving their expected shape, thus reducing the landmarks manual clicking process.

4.2 Surface registration

Radial Basis Functions (RBF) [2] are then used to interpolate the generic model to the raw data, and finally, vertices of the generic model are moved closer towards the raw data surface by considering their closest points.

The rationale behind using a generic model is that it firstly allows a great deal of prior knowledge and system to be infused along with its geometric data. The raw data obtained may not be suited for direct integration into an animation framework due to noise and vertex distribution. Facial regions can be defined once on the model, and can then become applicable to a range of different facial data such as jaw, eye regions, and scalp definition. The muscle system used in this application was specifically matched to the generic model and warped accordingly via RBF interpolation.

5 Facial Animation Model

For the purpose of realistic facial animation, a physically based, layered tissue model was created. A mass-spring system has been used to create this layered tissue model which represents the epidermal, fascial and skull layers. With this model the interaction between the nodes and springs provides a more naturalistic behaviour for the skin tissue, stretching and pulling as forces are applied. The basis of this implementation is formed primarily from [6].

5.1 Spring and Node Creation

Given a mesh of facial data, it is first necessary to create the layered tissue model. Fascia and bone nodes need to be generated. This proceeds as follows:

1. Calculate the surface normals at each defined vertex in the facial mesh. This vertex position becomes the position for an epidermal node.
2. Go an increment downward below the face surface, in a direction collinear to the normal, this defines the position of a fascia node.
3. Go a further increment downward in the same direction, this defines the bone node which is assumed attached to the skull surface.
4. Interconnect all nodes in the configuration as presented in Fig. 3.

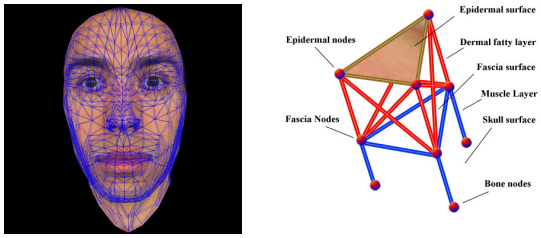

Figure 3. An element of the tissue model and the spring configuration over the face

Each facet of the input mesh represents a triangular prism element with a certain rest volume and location. Together these form a skin patch over the face. Muscle forces are applied to fascia nodes to generate muscle contractions.

5.2 Forces

The various forces when taken holistically, form a model for facial tissue. Once the layered tissue model is in place, abstract muscles apply a contraction force to the Fascia nodes within the system. By doing so the tissue reacts to this application until it reaches a new equilibrium, resulting in a reconfigured facial state.

Spring forces are induced from the effect of the various springs interacting between the nodes, to either push or pull. This element of the model produces a skin tissue continuum.

The **volume preservation force** allows for tissue form restitution by preserving the facial tissue structure, and is proportional to the observed change in volume and displacement of the volume element's nodes.

As well, two muscle models are used to apply forces to the fascia nodes; namely the **linear muscle** and the **ellipsoid muscle**. These abstract muscles are defined and placed in physiologically correct regions of the face. The abstract muscle allows for an important decoupling between the facial mesh and muscle position.

The **linear muscle** is also known as a vector muscle; having an origin where the muscle is attached onto the bone, and another point which is inserted into the skin tissue. Nodes that lie within a certain angular zone to the muscle vector, and within a certain radial distance from its origin, are directly affected by this muscle.

The **ellipsoid muscle** acts like a string bag. Points are squeezed towards the muscle origin. It possesses only a radial weighting to nodal force application, and has no angular weighting.

5.3 Running the simulation

The physical simulation is evaluated over time using explicit Euler integration. Through the use of an appropriately damped system, the simulation can be run at interactive rates on any current Pentium grade machine.
Expressions are the combination of various muscle contractions to generate a reconfiguration of the facial state. To animate the face from a neutral state, one needs only to progressively scale these contraction values from zero to unity over time.

5.4 Results

Results on depth map generation and face landmarks have been already presented on Fig.1 and Fig.2. Once the registration between the raw data and the generic model is achieved a range of facial expressions (namely sadness, anger, joy, fear, disgust, and surprise [6]) can be generated (see Fig.4). Falling within these categories are a range of possible intensities and variations in expression details. It is however not possible to convey through pictures the smoothness of animation produced by the tissue model. An evaluation of the system for one timestep takes 0.016 seconds with 8160 springs comprising the model.

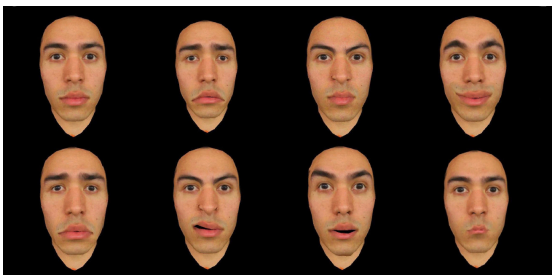

Figure 4. From top to bottom and left to right: Subject 1 neutral face, sadness, anger, joy, fear, disgust,and surprise, and pursed lips.

6 Conclusion

This paper has presented a low cost solution for a complete 3D face animation system using off-the-shelf digital camera for face depth map acquisition, enhanced 3D generic face animation model as well as a tentative automatic method for the mapping between the generic model and the real human data. The framework employs current research into stereo algorithms and a physically based model for animation, providing superior results over simpler geometric deformation alternatives. Although the processing cost is higher than in some other models, the complex interplay between the nodes, springs, and the forces that affect them validate the choice of model needed to attain realism. Current work includes the generation of a full 3D face surface including hair, more realistic lips and eyes as well as ear geometry. Future improvement including a more advanced reflectance model should be incorporated into the final result, and adding fine wrinkles as a further extension to accommodate for low facial tessellation could be pursued. This solution could be used as a foundation in various areas such as interactive avatars, game and movie characters environments. As computers become increasingly more powerful, we will no doubt see the realization of more advanced techniques in these areas.

References

[1] Y. Boykov, O. Veksler, R. Zabih, *Fast approximate energy minimization via graph cuts.* Intl. Conf. on Computer Vision, pp. 377-384, 1999

[2] J.C. Carr, R.k. Beatson, J.B. Cherrie, T.J. Mitchell, W.R. Fright, B.C. McCallum, T.R. Evans, *Reconstruction and Representation of 3D Objects with Radial Basis Functions.* In ACM Siggraph 2001, pp. 67-76, 2001

[3] P. Leclercq, J. Liu, M. Chan, A. Woodward, G. Gimelfarb, P. Delmas, *Comparative study of Stereo algorithms for 3D face reconstruction.* In Advanced Concepts for Intelligent Vision Systems, pp. 201-208, 2004.

[4] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *Proceedings of the International Conference on Image Processing*, volume 1, pages 900–903, 2002.

[5] Intel Corporation. Intel Open CV library. www.intel.com/research/mrl/research/opencv/overview.htm.

[6] F.I. Parke, K. Waters, *Computer Facial Animation.* ISBN 1-56881-014-8, 1996.

[7] R. E. Schapire. The strength of weak learnability. In *Machine Vision*, volume 5(2), pages 197-227, 1990.

[8] M. B. Stegmann, *The AAM-API: An Open Source Active Appearance Model Implementation.* In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Montral, Canada, pp. 951-952, 2003.