USING FACIAL FEATURE EXTRACTION TO ENHANCE THE CREATION OF 3D HUMAN MODELS

Eamonn Boyle[†]*, Bartlomiej Uscilowski*[‡]*, Derek Molloy*[†]*, Noel Murphy*[‡]

[†]Vision System Group
Dublin City University
Dublin 9, Ireland

[‡]Centre for Digital Video Processing
Dublin City University
Dublin 9, Ireland

e-mail: {eamonn.boyle, bartek.uscilowski, derek.molloy, noel.murphy}@eeng.dcu.ie

ABSTRACT

The creation of personalised 3D characters has evolved to provide a high degree of realism in both appearance and animation. Further to the creation of generic characters the capabilities exist to create a personalised character from images of an individual. This provides the possibility of immersing an individual into a virtual world. Feature detection, particularly on the face, can be used to greatly enhance the realism of the model. To address this innovative contour based templates are used to extract an individual from four orthogonal views providing localisation of the face. Then adaptive facial feature extraction from multiple views is used to enhance the realism of the model.

1. INTRODUCTION

The development of personalised interactive computer games and the creation of realistic humans to populate virtual worlds is a challenge that is currently seeing significant advances with different applications for the construction of models being continually developed. These techniques have two main foci; the creation of highly realistic models using animation tools [6] and the development of photo realistic models from images [4]. In addition the process of creating realistic human models has shifted from a research topic involving the use of high specification equipment in controlled environments to creation of highly realistic models using off-the-shelf technologies. This is illustrated using the *EyeToy* camera for the *Sony PlayStation2* which facilitates many interactive applications including the creation of a 3D model of an individual's head facilitating the immersion of the individual in the particular game.

Creation of models for immersion in virtual worlds and games have different requirements. In the gaming environments it is particularly advantageous to modify an underlying model because the moves in the games are often dependent on the size of the character and predefined animation, for example "*take three steps left*". While the complete reconstruction of the individual from a series of images can provide a highly realistic model for use in virtual worlds, its realism may be compromised by constraints within the gaming environment.

To accurately modify an underlying model or create 3D models it is necessary to have more than one view otherwise it is impossible to extract accurate 3D information. Four orthographic images of the individual are captured using the capture process described in [2, 4, 8]. This provides sufficient information to texture and modify the underlying model to take on the appearance of the individual and in particular the four views provide complete and unoccluded views of the head. Our method extends the research by Lee and Hilton by using deformable B-spline templates [1] that are constrained to accurately extract the individual from their environment and to localise the head.

The face of an individual provides significant detail that is important in determining the quality and realism of the model that is produced. Thus in developing any technique it is important that the face is well detailed. This can be achieved by finding correspondences between features on the face of the individual and the face of the model. The predominant features on an individual's face are the nose the eyes and the mouth, which provide geometric dependencies and constraints for precise face localisation. We use these dependencies to accurately position the individual's face on the model in order to enhance the realism of the 3D human model. However the locations of these features are commonly used in other applications, e.g. the normalised facial image for the creation of the MPEG-7 Face Recognition (FR) descriptor is obtained using the predefined eye locations [5]. Although the MPEG-7 FR descriptor creation requires only the frontal view of the face we propose to use both frontal and side views to enhance the automated localisation of the features and for texturing the underlying model.

There are a number of different approaches to automated facial feature extraction which use either gray-scale images or colour images of faces. The deformable templates [1] are among the most popular techniques such as symmetry operator, Hough transform, Gabor Wavelet Transform (GWT) [3] and Artificial Neural Networks (ANN) [9]. Other methods are based on eigenvectors, fisherfaces, fuzzy logic, Hidden Markov Model, stereovision or Independent Component Analysis (ICA). All of these algorithms work on the frontal view images of faces, usually under specified lighting conditions, constant resolution or deal only with good quality images. We propose the colour based segmentation technique, with no predefined conditions for the facial feature localisation in the frontal view images and the shape validation method from the side views.

The structure of the system for creation a personalised 3D model is presented in figure 1. Elements of the process are de-

scribed in details in the following sections, firstly the image capturing and fitting the template for extraction the full body contour followed by description of facial features extraction techniques. Finally the texturing process is outlined and some results of the personalized models are presented.

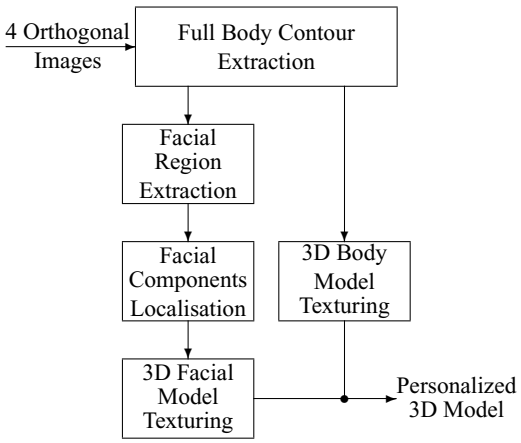


Fig. 1. Facial features localization algorithm

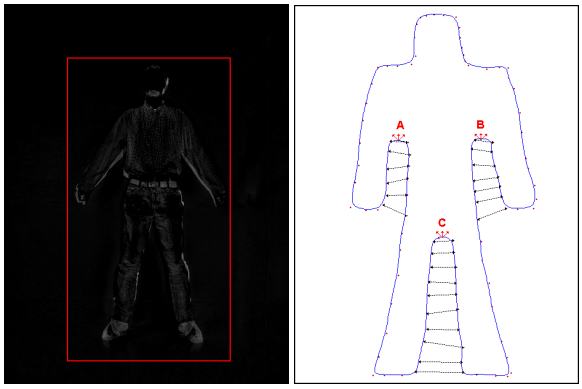
2. IMAGE CAPTURE AND TEMPLATE FITTING

Four images of the individual are captured from a real unconstrained environment. The images are captured with the stationary camera and between each capture the individual rotates 90^o. This enables the capture of the four orthographic projections. In this approach the individual stands approximately 3m from the camera. The user adopts the pose in figure 2. In this figure the individual is required to stand with their legs apart and their arms raised. This pose is necessary to accurately locate of the crotch and the armpits. In addition to this the individual should look directly at the camera or little above the camera. This is to ensure that all the features of an individual's face are visible which is important for the extraction of the facial features.

(a) (b)

Fig. 2. Example of the capture images.

It is necessary to have accurate segmentation of the individual from the background to ensure that the background is not textured to the underlying model. This is achieved using the template in figure 3(b). The template is essentially an active B-spline snake (closed contour) that is continually minimising while being constrained and attracted to strong edges. The actual template is based on the Active Contour Model (ACM) proposed by Kass et al. [7]. One of the initial requirements for using ACMs is that the contour must be initialised in the vicinity of object to be extracted. This is realised by subtracting the front and back views and subtracting the side views. This enables the estimation of the bounding box, see figure 3(a) that is used to scale the template for a particular individual.



(a) (b)

Fig. 3. Results of subtraction and the initial template with the constraints.

The initial template that is used for the front and back views is shown in the figure 3(b). The constraints to force the contour to behave in the desired manner are included. The main constraints were necessary between the two legs and under the arms because sometimes the contours converged to the same edge. In figure 3(b) the constraints are indicated using the dashed lines between the legs and under the arms. The arrows at the end of these lines are used to indicate that the control points cannot move towards each other. In addition to this the control points marked A, B and C are constrained in the direction that they can move, i.e. they can only move in the three directions indicated on the figure. This is to ensure that they will accurately locate the armpits and the crotch.

3. FACE LOCALISATION

Since the facial region is well localised by the final body contour it can be used for the precise localisation of the facial features. In the case of the frontal view of the face the colour segmentation is applied to the facial region and the facial components are found. The segmentation is carried out in three steps: an initial segmentation followed by feature extraction and classification.

The initial segmentation is based on the Recursive Shortest Spanning Tree (RSST) algorithm which is an automated hierarchical segmentation algorithm providing homogeneous connected regions with an easy to define granularity. The feature extraction employs the Expectation-Maximisation (EM) algorithm to obtain the accurate features boundaries. This is based on the observation that the histogram of the facial features such as eyes, eyebrows or lips can be modelled with the Gaussian distribution and these features differ in colour from the rest of the facial region which is skin coloured. Thus they can be extracted using this colour segmenta-

(a) (b)

Fig. 4. Facial features locations in (a) front view and (b) side view head image.

(a) (b)

Fig. 5. The locations of the facial features and the distances between them used for deformating the underlying model.

tion technique. Simple heuristics are added to classify the regions of the particular facial components which are based on the colour features and geometric dependencies between the facial components.The eyes and the centre of the lips localised with the RSST and the EM algorithm are illustrated in figure 4(a).

Other techniques for the frontal view facial feature extraction can be applied instead of the colour segmentation. The deformable templates as proposed by Yuille et al. [1] provide a good alternative for the localisation of the facial features if the initial position of the eye and mouth templates can be placed close to the features. This can be easily achieved since the precise location of the face is obtained from the body contour extracted in section 2.

The feature location in the side views of the head is achieved by scanning the shape of the head from the top to the bottom and searching for the valleys and peaks in the head boundary using gradient analysis. When considering the right hand side view of the head as shown in figure 4(b), the nose tip can be found as the peak on the right hand side boundary, the eyes should be placed in the valley above the nose peak and the lips form small hills below the nose tip. The valley of the chin provides information about the bottom boundary of the face, which in general can not be reliably located in the front view because of shadow. If the ear in the side view of the head is visible and not covered with hair its location can be found using the segmentation technique used for the features extraction from frontal view images.

4. TEXTURING AND PERSONALISING THE MODEL

To texture the underlying model four silhouettes corresponding to the captured images are generated. These are used for the establishment of correspondences between the captured images and the underlying model. The approach is based on feature extraction algorithm in [4]. The establishment of features is essential to enable the accurate texturing of the model. Having the correspondences enables the texturing of the model on a part-by-part basis ensuring that scale of the different body parts is preserved. In the texturing algorithm proposed in [2] the normal vectors for each tri-face of the model are used to texture the model. The major limitation with this approach is that the face is not accurately textured and this reduces the realism of the model.

To overcome the limitation of this approach the facial features are located on the face of the model and on the face in the captured images. These are used to align the individuals face with that of the model. These features are indicated in figure 4 and the geometric relationships are shown in figure 5. These relationships and distances are used for scaling and validation.

The distance between the eyes d_e (see figure 5(a)) is used for scaling the texturing image in the horizontal direction. This distance is used also for the creation of the MPEG-7 FR descriptor and is essential for defining the size of the head [5]. The vertical size of the texturing image is adjusted using the distance between the eyes and the centre of the mouth d_m. The three points representing eyes and mouth locations are used to calculate the centre of the facial region and to position the facial texture on the underlying model in the front view.

The side view images deliver information required for the enhancement of the head model viewed from the sides of the head. As shown in figure 5(b), the triangle drawn between the eye, the mouth and the ear is used for determining the scaling factor for side images. The distance between the ear and the eye d_d determines the depth of the head model whilst the distance between mouth and eye should be equal to the distance d_m in the frontal view image and can be used for the validation of the vertical dimensions of the image. In case of ear covered with hair and not visible the head boundary is used for finding the depth of the head model.

Once the facial information has been aligned the texturing technique in [2] is used to texture the facial model.

5. RESULTS

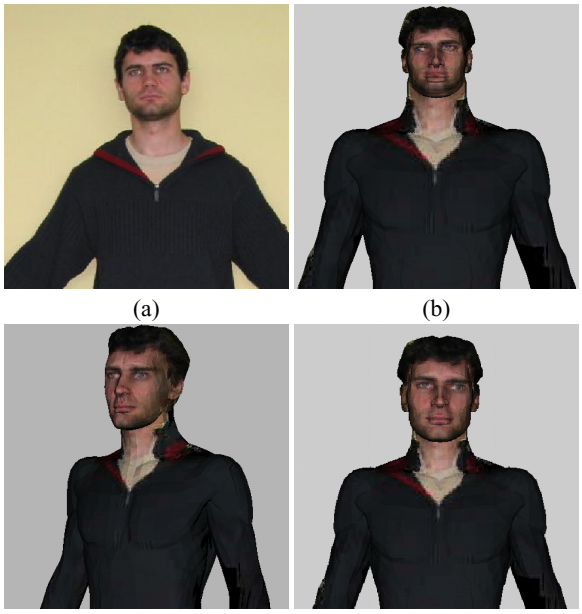
The results obtained using the proposed method to enhance the realism of the human models are presented in figure 6 and figure 7. It can be clearly seen that the quality of the models that use the facial features in the texturing of the model provide superior results. The images in figure 2 are used to texture the underlying model in figure 6. In this figure the complete model is textured using the information from the four orthogonal views. Figure 6(a) shows the model textured without using the facial features to position and scale the facial texture and figure 6(b) shows the model with the aligned facial features.

In figure 7 the results for a second set of images are combined to texture the same underlying model. In this set of results only the upper body an the head are shown and the difference can be easily seen when the features on the model are positioned accurately. In figure 7(a) the frontal image is shown. In figure 7 (b) the simply textured model is shown and figure 7 (c) and (d) shows two views of the textured model when the facial features are used to position the facial texture.

The enhanced model still does not provide high detail of the eye regions, although this technique can be used to texture the model using a separate high resolution facial image or mapping each of the facial feature separately.

(a) (b)

Fig. 6. The created human model with (a) misplaced facial features and (b) aligned facial features.



(a) (b) (c) (d)

Fig. 7. second human model with (a) the original frontal image, (b) misplaced facial features and (c) (d) two views of the model with aligned facial features.

6. CONCLUSIONS AND FUTURE WORK

The approach for enhancing the realism of the 3D human model employing the facial features localisation that is presented in this paper shows that the realism of the model can be substantially increased by considering the facial features when texturing the facial regions. It is believed that the presented method can improve the realism of the 3D human models for low-resolution devices such as PDAs or mobile phones providing a low-cost solution to the creation of personalised 3D models without requiring costly full 3D reconstruction.

The experiments provide promising results as they show improvement in the realism of the models when the texturing facial image is placed and scaled accurately. Further improvements to the realism of the model could be achieved by mapping each of the facial components separately, since the current technique, although improving the realism, still does not provide sufficient detail to warrant facial animation. The extraction of complete body and facial information from a single set of images would be advantageous for many applications. An alternative to achieve the higher level of detail is to capture a separate facial image and to use this to increase the realism of the model.

To gauge the quality of the model it is intended to provide objective measurement of the realism of the 3D human model using the face recognition descriptors in MPEG-7 and to compare the reconstructed face with real photos of the individual.

7. REFERENCES

[1] Blake and M. Isard. *Active Contours*. Springer Verlag, 1998.

[2] Eamonn Boyle. Generation and animation of virtual humans. In *IWSSIP04*, pages 143–146, Sep 2004. Poznan, Poland.

[3] R. Chelappa, C. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83(5), May 1995.

[4] A. Hilton, D. Beresford, T. Gentils, R. Smith, and W. Sun. Virtual people: Capturing human models to populate virtual worlds. In *IEEE International Conference on Computer Animation*, pages 174–185, 1999.

[5] ISO/IEC JTC1/SC29/WG11, N4980. *Overview of the MPEG-7 Standard*, Jul 2002.

[6] N. Kalra, N.M. Thalmann, L. Moccozet, G. Sannier, A. Aubel, and D. Thalmann. Real-time animation of virtual humans. *IEEE Computer Graphics and Applications*, 18(5):42–56, 1998.

[7] M. Kass, A. Watkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):231–331, 1987.

[8] W. Lee, T. Goto, and N.M. Thalmann. Making h-anim bodies. In *Avatars2000*, Nov 2000. Lausanne, Switzerland.

[9] M.J.T. Reinders, R.W.C. Koch, and J.J. Gebrands. Locating facial features in image sequences using neural networks. In *II Int. Conf. on Automatic Face and Gesture Recognition*, pages 230–235, 1997. Killington, USA.