Automatic Selection Of Feature Points For Video-Based Facial Animation

Damian Pęszor and Marzena Wojciechowska

Polish-Japanese Academy of Information Technology, Koszykowa 86, 02-008 Warsaw, Poland

Abstract. The paper describes an approach for selection of feature points on three-dimensional, triangle mesh obtained using various techniques from several video footages. This approach has a dual purpose. First, it allows to minimize the data stored for the purpose of facial animation, so that instead of storing position of each vertex in each frame, one could store only a small subset of vertices for each frame and calculate positions of others based on the subset. Second purpose is to select feature points that could be used for anthropometry-based retargeting of recorded mimicry to another model, with sampling density beyond that which can be achieved using marker-based performance capture techniques. Developed approach was successfully tested on artificial models, models constructed using structured light scanner, and models constructed from video footages using stereophotogrammetry.

Keywords: facial animation, feature points, retargeting, mesh sampling
PACS: 07.05.Tp

INTRODUCTION

Facial animation is considered one of the most advanced issues in the field of computer animation. Unlike whole body animation, it cannot be simply defined using weights associated with bones of skeleton-like hierarchical model, due to the soft tissue that forms the facial structure. However, bone-based animation is sometimes used for the purposes of facial animation. It is so because of commonness of tools and frameworks based on bone systems, which can be used to create an animation based on abstract skeleton designed for specific motions. This, however, is not related to the actual structure of human face. Physiological models, on the other hand, are complex, difficult to create, and specific for a given character, which means that their usability is very limited. This, combined with the fact, that most of computer animation's applications are in the area of video games and CGI movies, rather than medical-grade simulations, makes those models very uncommon. Another approach is based on shape interpolation, wherein the vertices' positions are defined for each keyframe for given animation and then interpolated between keyframes to acquire desired information about model in each frame. In case of artificial model, which is manually created by the artist, the amount of work required for each keyframe to contain specific shape as part of animation, is enormous. However, when video-based performance capture is used to obtain three-dimensional model of actor's face in each frame, the amount of work is greatly reduced.

Another problem still occurs though, when one wants to obtain high quality animation. Depending on number of cameras, their parameters and technique used to reconstruct high quality model from video, one can expect facial meshes constructed from ca. 85000 to even as high as 3 millions of polygons. While these numbers reflect the complexity of the model, the actual size of such a models in OBJ format is 8MB and 160MB respectively. If stored in more appropriate way for animation, one could define texture coordinates and topology only once and assume that normals will be recalculated instead of stored. Vertices position in three-dimensional space could then be stored in binary format using 8 bytes for each coordinate. Let us consider barely 'high quality' model and a model that could be seen as extreme 'high quality' (note that in case of facial structure this is much higher than in less complex surfaces or surfaces that do not draw so much focus) in Table 1. The size of motion data depends on number of vertices, but one has to take into account the density of keyframes, lenghts of animations, their number and number of characters to be animated to finally calculate how much data is needed to obtain desired quality. Obviously, the high-end animation is unfeasible due to the size of data required. Proposed approach is based on the concept, that with dense meshes, one does not need information about each vertex movement, but instead, this information can be calculated on the basis of how specific vertices move. These vertices are selected on the basis of mesh's topological and geometrical information, so they contain information that is relative for their surroundings.

TABLE 1. Extremes of data sizes for high-quality animation

	from	**up to**
number of polygons	85000	3000000
OBJ model size [MB]	8	160
number of vertices	43000	1500000
size of vertices [MB]	1	34
keyframes per second	15	60
animation length [s]	1	5
number of different animations	5	20
number of different characters	5	20
approximate data size	360 MB	4 TB

The main purpose of this approach is, however, not the reduction of data size. In our experiments related to marker-based facial animation, we noticed that the amount of markers needed for high-quality animation of dense three-dimensional mesh results in unnatural mimicry. Most often, it also introduces issues in tracking, which can be clearly seen in high amount of errors. This leads to the time-consuming need of manual repair of marker trajectories. Marker-based performance capture is therefore restricted in quality by the number of markers that can be applied to human face. In other words, the restriction of sampling density results in not enough information to produce a realistic facial animation of high-quality model in a fully automated way. Higher rate of sampling can be achieved using video-based performance capture. However, systems based on video are mostly used with a restriction, that there has to be high similarity between mimicry of the actor and the animated model. This is because the mesh is constructed on the basis of actor's appearance, which can then be modified to differentiate between actor and animated character. Using approach proposed in this paper, we are able to use high-density model created during video-based performance capture as a source that can be sampled. This sampling, with much higher density than in case of physical markers, provides us with data that can be retargeted to animated mesh using correspondence of anthropometric points. In result, creating an animation that captures details of actor's mimicry while at the same time applying it to high-definition model of a different person.

METHOD

Input data

Presented approach requires a three-dimensional triangle mesh, which will be analysed in order to select appropriate vertices for the purpose of facial animation. It is assumed, that the mesh is constructed on the basis of video recordings. Depending on the quality of mesh construction, the data might need to be preprocessed to remove erroneous vertices and triangles, ensure that the surface is topologically correct and, depending on exact needs of the user, trimmed to facial expressive area. This preprocessing for the purpose for facial animation was previously described in [1]. In most of stereophotogrammetric techniques, one has to either manually select corresponding points on multiple video frames or find them using facial recognition algorithms. In order to easily find correspondences, selected points are most often anthropometrical and therefore their three-dimensional reconstruction can be used as a basis for selection. If this is not the case, then the mesh can be analysed to estimate placement of basic anthropometrical points, such as in previous work for marker-based facial animation [2].

Curvature-based segmentation

Human face is not a featureless, curved surface, but instead it contains features that are important part of the mechanism of facial expressions. One can not therefore ignore them when trying to select appropriate vertices that will represent deformation of entire facial surface. To find such features in discrete mesh, one must calculate the curvatures in each vertex. There are many different methods of calculating curvatures, however, one has to take into account specificity of the context. In case of high quality meshes, the density of vertices can be high enough so that the curvature will not be sufficiently emphasized to justify recognizing vertex as part of facial feature. Because of that, one has to employ a method that will not be restricted to one-ring neighbourhood, but work well with broader patches of data. Moreover, due to the fact that density of the meshes obtained using different techniques and with different quality parameters may vary greatly, a method is needed to select appropriate neighbourhood size for curvature estimation. Since in most cases the corresponding physical length of unit vector in mesh's coordinate system will not be known, anthropometric distance can be used to obtain the scale ratio. [3] contains comprehensive data about anthropometric facial distances for various populations which can be used as a reference for facial mesh analysis. While some other distances could be used, in our approach we use intercenthal width as a basis for further calculations. This width is stable; it does not differ much between populations, genders, or persons, so it can be used reliably. It is also one of most basic distances, based on most commonly found feature points; inner eye corners. While we used data for representatives of Polish population, one can assume international mean instead of population-specific to calculate physical diameter of neighbourhood that will be used for curvature estimation. Our tests proved that diameter of 3mm is enough to provide significant curvature differences on vertices that are part of anthropometric features. This corresponds to slightly below 10% of mean intercenthal width for men and slightly above 10% of same value for women.

[4] analyses methods for curvature estimation and proposes N-ring neighbourhood methods that are adequate to the issue at hand. Based on this analysis, we implemented polynomial fit with Desbrun parametrization which was based on [5]. For each vertex of the facial mesh, we took at least one-ring neighbourhood to calculate curvatures. If any of two-ring neighbourhood's vertices was connected to analysed vertex by edges of distance shorter than 10% of intercenthal width, we used two-ring neighbourhood instead. If all vertices of subsequent N-ring neighbourhoods comply with these conditions, N-ring neighbourhoods were assumed. Our tests proved, that neighbourhoods greater than 5-ring were not introducing enough data to warrant computational complexity.

Having found Gaussian and mean curvatures of each vertex, we perform HK-sign surface categorization (as proposed in [6]) which allows us to find edges of sets of vertices that contain same HK-signature. Each group of vertices is treated as a separated segment for the purpose of further vertices selection.

Significance of range of motion

While analysis of single mesh obtained from corresponding frames of multiple video recordings will provide information about facial structure that can be sufficient for most animation purposes, one cannot benefit from high quality meshes while using only this data. To be able to predict which vertices can sufficiently represent given surface, there is a need to base our approach on dynamic data rather than static. Having numerous meshes obtained from each frame of video recording means that we effectively have an information about range of motion for facial structure for a given animation. Specific recordings designed for range of motion evaluation can be used instead of a recording of single facial expression in order to select vertices independently of specific mimicry.

Depending on technology used to obtain meshes, the correspondence of vertices between frames might or might not be included (in most cases, same vertex will be indexed by same value). In cases where correspondence was not given with construction of the mesh we used [7] to obtain such a correspondence data. It is worth to note, that due to the fact that facial expressions can radically impact construction process, one has to establish correspondence between consecutive frames rather than between each frame and starting, neutral one.

Due to the elasticity of the skin, segments established for first frame will be broken down into smaller subsets of vertices. However, one can expect, that with sufficient range of motion, the curvature in each vertex changes, so that vertices, which were previously located on the edge of the segment will move be further from it on consecutive frames. This effects in blurring of the segment edges. To compensate for such an effect, first DB SCAN (see [8]) algorithm is started from characteristic point, then three-dimensional curve is fitted to found clusters (see [9]) between characteristic points, which is in turn used to find the appropriate path with vertices closest to the curve. This approach proves to produce smaller segments which can then be used to select points that will carry the weight of animation.

Vertices selection

After above segmentation process, surface area will not be distributed evenly among the segments. We assume, that the number of feature points for facial animation is given as a parameter for algorithm. Characteristic points established as part of the input are included in this count, they will most probably lie on edges of segments, but this is not necessary. To select other vertices, first the surface and segments areas are calculated. This establishes how many points each segment should contain. Each segment is then analyzed from smallest one to the one with greatest area.

As part of segment analysis, first the deepest vertex is found. That is, for each vertex inside of the segment, the least number of connections that join it to segment edge is calculated. The vertex which has greatest number of necessary connections is established as deepest. In case of few vertices sharing this characteristic, the one closest to segment centroid is selected. Then, the furthest point on the edge of the segment is selected as second feature point. As long, as there are still vertices available for given segment, the distance between every vertex and every characteristic point on the segment is calculated, and the vertex with greatest distance is selected. Each subsequent segment is analyzed while taking into account the selected vertices on it's edge.

CONCLUSION AND FURTHER WORKS

The method presented in this paper aims to obtain an arbitrary number of feature points based on facial mesh obtained from video recordings. Apart from minimizing the data needed to store very high quality animations, this allows the use of algorithms destined for facial animation using only a subset of vertices (most notably, marker-based performance capture), such as [10]. This also eliminates the need to use very dense information to calculate animation with lower level of detail.

While presented approach proves to be working with meshes up to 3 million polygons, the complexity of the algorithm proves such calculations to be unfeasible. Further work is required in order to simplify the algorithm, especially in the area of segment edges detection.

ACKNOWLEDGMENTS

This project has been supported by the National Centre for Research and Development, Poland.(project INNOTECH In-Tech ID 182645 "Nowe technologie wysokorozdzielczej akwizycji i animacji mimiki twarzy." - "New technologies of high resolution acquisition and animation of facial mimicry.")

REFERENCES

1. D. Pęszor, A. Polański and K. Wojciechowski, "Preprocessing of 3D scanned images for facial animation on the basis of realistic acquisition," *AIP Conference Proceedings*, 2015, Vol. 1648 Issue 1
2. D. Pęszor, A. Polański and K. Wojciechowski, "Estimation of marker placement based on fiducial points for automatic facial animation," *AIP Conference Proceedings*, 2015, Vol. 1648 Issue 1
3. L. G. Farkas, M. J. Katic, C. R. Forrest, K. W. Alt, I. Bagic, G. Baltadjiev, E. Cunha, M. Cvicelová, S. Davies, I. Erasmus, R. Gillett-Netting, K. Hajnis, A. Kemkes-Grottenthaler, I. Khomyakova, A. Kumi, J. S. Kgamphe, N. Kayo-daigo, T. Le, A. Malinowski, M. Negasheva, S. Manolis, M. Ogetürk, R. Parvizrad, F. Rösing, P. Sahu, C. Sforza, S. Sivkov, N. Sultanova, T. Tomazo-Ravnik, G. Tóth, A. Uzun and E. Yahia, "International anthropometric study of facial morphology in various ethnic groups/races," *Journal of Craniofacial Surgery*, 08/2005, 16(4):615-46
4. T. D. Gatzke and C. M. Grimm, "Estimating curvature on triangular meshes," *International Journal of Shape Modeling*, 2006, Vol. 12, No. 01, pp. 1-28
5. M. Desbrun, M. Meyer and P. Alliez, "Intrinsic Parameterizations of Surface Meshes," *Computer Graphics Forum (Proc. Eurographics 2002)*, 2002, 21(3), pp. 209-218
6. P. J. Besl and R. C. Jain, "Invariant surface characteristics for 3D object recognition in range images," *Computer Vision, Graphics, and Image Processing*, 01/1986, Vol. 33, Issue 1, pp. 33-80
7. V. Jain, H. Zhang and O. van Kaick, "Non-rigid spectral correspondence of triangle meshes," *International Journal of Shape odeling*, 06/2007, Vol. 13, Issue 1
8. M. Ester, H. Kriegel, J. Sander and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, AAAI Press, pp.226-231, ISBN 1-57735-004-9
9. L. Fang and D. C. Gossard, "Multidimensional curve fitting to unorganized data points by nonlinear minimization," *Computer-Aided Design*, 1995, Elsevier Science Ltd. Vol. 27, No. 1, pp. 48-58
10. D. Pęszor, K. Wojciechowski, M. Wojciechowska, "Automatic Markers' Influence Calculation for Facial Animation Based on Performance Capture," *Lecture Notes in Computer Science*, 2015, Vol. 9012, pp. 287-296