http://dx.doi.org/10.12785/amis/paper

# Estimation of marker placement on scanned model based on fiducial points for automatic facial animation

*Damian Pęszor*[1,2], *Andrzej Polański*[1] *and Konrad Wojciechowski*[1]

[1]Polish-Japanese Academy of Information Technology, Koszykowa 86, 02-008 Warsaw, Poland
[2]Institute of Informatics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland

Received: ..., Revised: ..., Accepted: ...
Published online: ...

**Abstract:** Facial animation is presently achieved using performance capture techniques, most of which is based on the idea of tracking reflective markers located on facial surface. Available systems either provide information about artificial skeleton which is used to animate character or recognize actions which are used to calculate transformations for animation. Finding a way to tie surface of animated facial model to surface on which marker is placed on actor's face with different facial features is required to create a fully automated system for facial animation that will apply actor's facial expressions to animated mesh without need for adjustment.
Presented in this paper is a way of estimating marker arrangement on realistic facial model acquired using structured light-based 3D scanner.Since such registration is prone to errors related to specifity of facial surface, the preprocessing needed to use model acquired this way is described. The process of aligning preprocessed facial model to the mesh of neutral human face with manually selected marker's positions is presented. Finally, paper describes finding markers estimates on preprocessed mesh using anthropometric features represented through fiducial points and aforementioned neutral model by using position of harder to find markers with relation to those described by anthropometric features.

**Keywords:** facial animation, marker placement, fiducial points, performance capture

## 1 Introduction

Performance capture is a set of technologies used to acquire information about deformation of surface of human face during facial expressions. Most technologies track reflective markers placed on actor's skin in order to find actions or transformations of virtual bones. This can then be used by animator as inspiration, as a source to propel abstract skeleton system that controls facial expressions or to detect actions that can be used to determine transformations of surface. Alternatively, this can serve as a basis for semi-automatic facial animation based on surface transformations. This technique, however, is still refined so it could became fully automatic.

Due to the complexity of human facial structure and number of ways mimicry can affect it, appropriate arrangement of reflective markers on facial surface is one of most important parts of successful performance capture. Those markers should be placed in a way that can effectively represent possible deformations of face's surface while facial expressions are performed. Oftentimes this potentially crucial step is neglected and markers are placed in possibly dense grid, which is thought to best represent entire surface of face. Practice shows, however, that there are a number of factors that result in issues for even best performance capture systems, and certainly those with lower parameters. Each such system must struggle to distinguish small markers used for performance capture from other reflections on face which are result of noise and non-perfect lighting conditions. Also, there is a need to distuingish one marker from another if they are close to each other, which limits the density of arrangement. Fast facial expressions as well as overall head movement result in system being unable to track marker and will recognize it as entire new one or even mistake it for another thus giving significant errors. Some facial expressions may occlude some markers, which again results in tracking problems. All this leads to need of manual repair of recorded data, which in turn becomes less and less reliable as density of the grid grows.

* Corresponding author e-mail: dpeszor@pja.edu.pl

Unlike many other surfaces animated using motion capture techniques, face has structure which restricts possible expressions thus allowing the use of more sophisticated marker placement methods than simple grid. This results in finer details, better modeling of distortion, lower number of markers and finally less time spent on correcting errors due to more analytical approach to facial animation.

It is typical that animated character resembles actor whom animates it. While there can be a lot of difference in terms of texture and details, the anthropometrical structure of face, however, is often quite similar. This is because change in antropometrical distances would require scaling and moving some of triangles. Using fiducial points as a basis for marker placement provides model-independent data by assuming positions of anthropometrical markers instead of unreliable scaling based on grid arrangement.

While marker placement based on characteristic of human face is not uncommon, it is mostly based on intuition and not automatically transferrable to another model leading to further manual labor related to applying distortion found using marker trajectories. The aim of this work is to present a way of finding fiducial points on scanned face, which can be used to place markers for performance capture that can be independent from structure of actor's face, not be dense enough to cause recognition errors, while provide meaningful data.

## 2 Materials and Methods

### *2.1 Model acquisition*

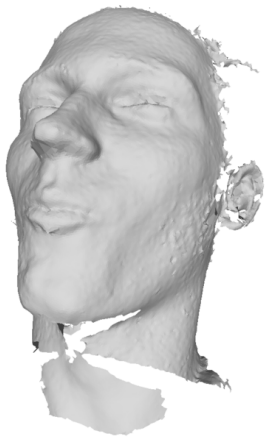


**Fig. 1:** Sample untextured mesh as acquired by 3dMDface System

To achieve the most realistic results it is essential to use an appropriate model of human face. Although there are artists capable of creating a very realistic model of the face, using a model acquired from real person is still undeniably an approach that guarantees the highest degree of realism. Additionally, in case of performance capture being done by same actor as the one used to obtain the model, his mimicry will be always best represented by his own face. It is also worth to note, that while facial animation can be used to propel many abstract faces, it is impractical to present such algorithms on anything different than real face, since human could possibly take an error in animation for a feature of abstract anatomy. Also, some fiducial points might be detected incorrectly on an abstract face. In case of real face, however, a slight error is clearly visible and anthropometrical data applies.

Obtaining an accurate model of human face is mostly done using 3D scanning techniques employed by 3D scanners, depth cameras or similar devices. The technology used to acquire mesh of scanned object is not perfect and neither are the conditions under which acquisition is performed, therefore this approach is prone to some specific errors that are either non present or negligible in case of different type of surface. Humans can easily notice any error of face representing mesh, especially if it changes over time due to animation, therefore special care needs to be employed to remove erroneous data from acquired model. This work presents approach used for scanned mesh correction and rationale behind it based on year-long study of facial animation in Polish-Japanese Academy of Information Technology.

Models of human face used in this work were obtained by three-dimensional scanning using structured light based scanner, namely 3dMDface System at The Institute of Theoretical and Applied Informatics of Polish Academy of Sciences in Gliwice. Eight different actors were scanned for a total of forty models representing neutral facial expression and some emotions, samples of which are presented in Fig. 1 and Fig. 2.

### *2.2 Mesh preprocessing*

Due to specifity of structured light-based scanning, few preprocessing steps are needed in order to improve quality of the model. Number of different errors is clearly visible as well as parts of scene that are not strictly related to human face mimicry and for purposes of animations are effectively just a noise. This is different problem from having noisy, but consistent structure of mesh (as solved by e.g. [1]) It is worth to note that most of below methods can be used together to minimize amount of traversing through vertices or facets. They are, however, presented sequentially for simplicity.

Meshes acquired using 3D scanning technologies are most often represented in one of common formats used in three-dimensional computer graphics. In most cases, such model is composed of vertices' positions, triangles described by three vertices' indices as well as other data,

**Fig. 2:** Sample textured mesh as acquired by 3dMDface System.

i.e. texture coordinates, normals to the triangle in each of vertices, etc.

Depending on scanner used, it is possible that acquired meshes might be composed of vertices that are not necessary part of model's surface, and therefore their indices are not present in any of triangle's definition. Those vertices should be identified and removed so they won't affect further calculations, e.g. finding centroid of the mesh. Some of described steps of preprocessing might change the model so that some vertices will become isolated. In our approach, those were deleted, however it is a matter of quality of the model whether it is actually worthwhile to do so.

In some cases, scanner might produce duplicate triangles, that is; multiple triangles composed of same vertices, in either same or different sequence. While the number of such occurences seem to be low, those easy to detect duplicated triangles can be troublesome in further removal of non-manifolds. It is therefore worthwhile to remove those duplicates before attempting further steps of preprocessing.

A major issue with any retrieved model is it's topological continuity. The specificity of 3D mesh acquisition for the purpose of facial animation indicates that face is central object on the scene and is not occluded by anything different than actor's hair. Based on that, it is clear that major part of acquired mesh will represent face itself, while smaller fragments will be either erroneous or will represent hair or ear, which should not be affected by performance capture. Any algorithm, either one related to preprocessing or animation itself, will be affected through decrease of quality of results or performance in case of topologically discontinued model having additional parts which will not be animated.

Next step of preparing acquired model for the purpose of facial animation is removal of surfaces not continuous with main part of mesh. Topologically continuous surface $S_i$ is obtained using any of mesh-traversing algorithms, such as breadth-first search (BFS). The traversing algorithm starts with any triangle $t_j$ of the mesh and traverses through adjacent facets while grouping them into a set representing continuous surface. When no adjacent facets are present, algorithm should start again for any non-grouped facet and continue to do so until every facet of the mesh is grouped. After the grouping is done, triangle count $n(S_i)$ in each group will determine which group is major part of acquired facial model and therefore should be kept while other should be removed, as seen in eq. 1.

$$\arg\max_i f(i) := n(S_i) \tag{1}$$

In case of scanned mesh, regions like eyes, eyelashes, eyebrows, mustache, beard or hair are prone to generate non-manifold edges. This type of error can significantly impair any mesh-based algorithm. Non-manifold edge can be found by traversing through each facet of the model, while counting every facet adjacent to every edge. Different number of adjacent facet's count has different consequences:

- 1 adjacent facet means that edge of the facet is also edge of entire continuous surface.
- 2 adjacent facets indicate correct edge on the inside of continuous surface.
- More adjacent facets mean that analyzed edge is non-manifold.

BFS algorithm is used to obtain sets $A$ of facets connected to each facet of non-manifold edge. Only facets from which sets indexed as shown in eq. 2 are retrieved, should be preserved.

$$\arg\max_i f(i) := n(A_i), \arg\max_j f(j \neq i) := n(A_j) \tag{2}$$

In some cases it is possible, that this strip of triangles will again be connected to main surface. If on both ends of this strip there are non-manifold edges, while other triangles adjacent to original non-manifold edge are correctly adjacent to another triangles at same depth of search, consistency of this strip with model is highly doubtful and entire strip should be removed. Another technique is to check the difference between normals of facets around non-manifold edge. In most cases, two of them will have little angle between each other, while high angle compared to normal from erroneous facet. As a last resort, one should remove facet adjacent to non-manifold edge to which adjacent path is shorter until it can not find another triangle that was not visited earlier, which will prevent long strip of hair being attached to face.

Even with facets related to non-manifold edges removed from the model, some artifacts will still be present. Some can be removed by removing non-manifold vertices. In most cases, this error will be seen as a pyramid representing few recognized vertices of hair strip

attached to main surface of the model. In this case it is reasonable to remove set of facets that has smaller area. In case of areas being comparable - set with smaller area of polygon connecting outer edges of it's triangle will need to be cut, so smoother side of non-manifold vertex will be preserved.

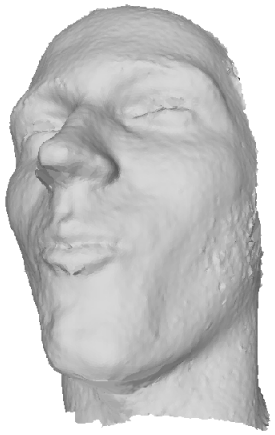


**Fig. 3:** Sample untextured mesh after removal of erroneous data.

Some noisy data around edges of the model will remain intact after removing non-manifolds. Those strips of triangles representing hair or parts of ear might be continuous part of mesh with topologically correct edges and therefore are difficult to find. First method to do so is based on the breadth of connection to main surface. For this purpose the edges of the model are found (see below) and number of vertices each edge is composed of is calculated. For every vertex of an edge, Euclidean distance to every another in given range is calculated. Experiments show that square root of number of vertices is a good range of search for second vertex. Vertices that have smaller distance to base vertex than those with fewer vertices in between, is likely to be a starting point for unwanted strip. Starting from the most distant (in terms of vertices in between) vertex having this characteristic, the shortest path through facets' edges is calculated. Then, the part of mesh that contains edge in between selected vertices can be removed.

Another type of error is related to single vertex being placed with an offset to original model (this is mostly reason of either a reflection or correctly estimated position of hair surface, which was not affecting surrounding facial surface due to it's small size) which creates a triangle connected to others under large angle. To prevent that, one is bound to calculate normals for triangles near the edge of the model (with correction methods explained above, a depth of 3 triangles proves to be sufficient) and compare them. An estimation of curvature can be calculated or simply for each step of depth, average of normals can be used to extrapolate next depth's normal. Our experiments prove that triangles with normals being off from estimate by more than $30^\circ$ can be considered erroneous and removed.

It is quite common that scanned model contains some gaps due to light dispersion, device's error recognition or part of actor's head not being clearly visible by camera. Simple filling this region with triangles will be perceived as artificial. Many algorithms were developed that deal with this issue, of which [4] proved to provide great results in case of facial hair and chin area, which due to it's size is most demanding. It is important to note, that this algorithm assumes continuous, manifold surface, therefore previous error removal is necessary. [4] assumes that every identified hole is going to be filled, this is not really the case when model is to be used for facial animation. Any additional data that is not strictly related to mimicry will only harm further attempt of animating face and glueing it to skeleton-based model. Therefore after identifying all gaps of the model, one should calculate the length of edge of each gap and resign from filling the hole which has longest border. Next, [4] refers to 3D adaptation of advancing front mesh algorithm [5] which creates a basic patch of triangles to fill the gap, however it's results do not preserve shape of surrounding region. Next, the main part of [4] contribution takes place as desirable normals are computed, triangles are rotated and patch is reconstructed based on Poisson equation. Authors suggest using Harmonic-based desirable normal computing for more planar surfaces since it is faster than geodesic-based approach, in case of relatively high curvature of chin and not so planar curvature of chin-neck area, it is preferable to use geodesic-based normal computing, which will not be as expensive in case of other holes present in the model due to it's small sizes. After successful reconstruction of vertices positions, it is worthwhile to calculate texture coordinates for those vertices as well because assuming any colour not based on texture will result in unrealistic appearance. This can be done using spherical interpolation based on linear combination of vertices surrounding the hole.

It is typical for any 3D model to contain data about normals, however one needs to remember, that those normals are based on same data that led to development of erroneous mesh. It is therefore rewarding to estimate normals for parts of model affected by error removal. There are number of algorithms for normal estimation, many of them focused on preserving sharp features. In case of face, however, proper selection of algorithm for normal estimation should focus on minimizing error in low noise environment, since most of noise is already removed from model, and high-noise methods could damage the quality of fine details like wrinkles.

Depending on quality of acquired mesh, removal of erroneous facets have resulted in up to 13% of vertices and facets of original mesh being deleted of which none were proper parts of face's expressive area. Sample

meshes obtained using described approach are shown in Fig. 3 and in Fig. 4.

**Fig. 4:** Sample textured mesh after removal of erroneous data.

## *2.3 Transformation to canonical pose*

Most existing algorithms based on anthropometrical data are used to solve the issue of facial recognition rather than automatic animation. One of most promising algorithms that can be applied to animation, called Anthroface 3D, was developed by Gupta et al. [6]. A basic step of bringing human face to canonical pose is assumed, since scanned face could be rotated impairing proper determination of directions used for finding fiducial points. This step also reduces the area searched for fiducial point, due to assumption of anthropometrical restrictions. This enables to find fiducial points using features which are not globally unique, but are strong enough to be found if search area will be reduced.

[6] suggests aligning mesh to one in canonical pose, since differences between faces are not big enough to prevent that. While Gupta et al. propose iterative closest point (ICP) [7], it had been improved for robustness, since it can easily find local minimum of it's error function instead of global one. For a case of face alignment, it is rewarding to use more robust approach of Trimmed ICP [8]. This approach minimizes the error in case of two aligned models not being exactly same, but sharing overall shape.

As reported by it's developers, TrICP works properly for initial relative rotations of up to 30°, therefore to fully automatize the process, there is a need to cover the rotation space with enough density. In our approach, this problem was solved using standard icosphere algorithm with level of refinement being a parameter. Each vector from icosphere's centre in (0,0,0) to vertex is used as a basis to find rotation used to rotate model before running TrICP algorithm. In case of residual error being not small enough, part of icosphere with biggest errors is rejected, while other part is refined. This process is repeated until acceptable residual error is achieved. TrICP seems to focus on case of two models being partial representations of same object and therefore have same scale, in case of face alignment it is rewarding to also include scale aspect. Both TrICP and ICP use unit quaternion method [9] to obtain absolute rotation and translation between transformed model and original one. Apart from rotation and translation, [9] also includes a way to calculate scale. Using scale improves quality of consequent iterations of TrICP. Although Horn suggests that in case of one model being more accurate than other, either scale calculated using rotation matrix or it's inverse should be used. In this case, however, symmetrical scale have yielded better results. TrICP might cause some problems in case of models being of different resolution because of finding correspondences which in one model represent small part of surface, while in the other majority of vertices is used. Neutral mesh should therefore be prepared with resolution comparable to the one achieved by source of models to be animated.

## *2.4 Fiducial points estimation*

Few changes are proposed to [6]. Manual detection of noise tip in neutral model was extended to every fiducial point for which estimation has been implemented as well as further points in between. Search in initial region of 96x96 mm about ICP estimate of nose tip seems to be based on view of canonical pose of human face projected to 2D creating sort of depth map rather than 3D model. This approach can be misleading in case of bulbous, rounded or pinched nasal tip (see [10]) which might make curvature assumptions invalid, especialy in case of high resolution, when curvature estimated on basis of one-ring neighbourhood can be much lower than expected. Different approach was therefore assumed. Region searched for nose tip was determined by placing sphere of radius $r$ with center in position of neutral face's nose tip $p_{nt}$ in aligned mesh composed of set of vertices $V = (v_1, v_2, ..., v_n)$. The radius is calculated as:

$$r = \|p_{nt} - (\sum_{i=1}^{n} v_i)/n\|/2 \tag{3}$$

Similarly, all regions used in searching for fiducial points were defined as a spheres. Next, a gradient based method was used to traverse from vertices closest to centroid of the model to farthest. This allowed to differentiate nose tip (farthest from centroid) from nose bridge and sides. Due to model's centroid position depending on quality of TrICP as well as bordering vertices of the model, it is necessary to establish smaller (5 times, which corresponds to radius of ca. 1cm) sphere in newly found nose tip. Nose tip estimation based on curvature (as

suggested in [6]) is then much less prone to errors. Still, however, it is rewarding to use curvature estimation based on more than one-ring neighbourhood. Among methods presented in [11] paraboidal fitting was used to obtain curvatures with two-ring neighbourhood. Obtaining further fiducial points follows similar procedures. Once all points were found using curvature-based or other methods (a range of two-dimensional methods employing texture information as well as three-dimensional algorithms for facial recognition have already been developed and can be used here), neutral model is once again used.

Depending on quality of performance capture system, one might be able to use space between already established fiducial points to increase sampling of face surface thus increasing the quality of animation. It is important to note, that those additional fiducial points need to be distant enough from other points, so system could distinguish markers placed on those points without a chance of recognizing one as the other. Fiducial points on animated mesh should be now aligned to their known correspondences on neutral model using ICP (there is no need for TrICP due to low number of points). This serves to correct the initial pose. After this, neutral model should be warped, so each fiducial point would occupy same coordinates as it's correspondent point on second model. All additional marker points on neutral mesh should be warped so they would remain in their position relative to fiducial points. For every marker point $m_j$ a normal $N$ can be calculated from plane composed of it's neighbouring fiducial points. Marker placement on animated model can be estimated as a vertex $v_i$ such that:

$$\arg\min_i \frac{\|v_i - m_j\|((v_i - m_j)\cdot N)}{\|v_i - m_j\|\|N\|} \tag{4}$$

## 3 Results

Marker placement was estimated on face acquired using structured light-based scanning by 3dMDface System. In presented case, positions of 58 markers were estimated, with average distance from manually selected points being equal to 1,3% of the distance from mesh centroid to manually selected point and maximum one being as big as 2,7%. On models acquired during this research, an average error of 2,2% was achieved in case of 58 points. Figure 5 presents found estimates as n-ring neighbourhood of faces with area big enough to be noticeable.

## 4 Discussion

Estimates of marker placement found using described approach seems to indicate, that the overall structure of the method shows promise of great results, which seem to be better the more points is used, since the search region becomes smaller estimate becomes more precise. In this study, information about colour as given by texture coordinates for every vertex was not used, however given the number of developments about facial recognition based on images, it is reasonable to assume that smaller errors could be achieved if that knowledge was employed. Future study about deformation based on found characteristic points is also needed to fully utilize results of this method. This approach might however be used as an initial step toward non-skeleton and non-action based fully automatic facial animation.

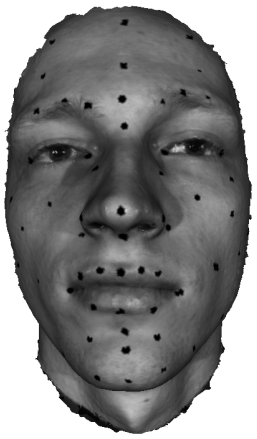


**Fig. 5:** Sample textured mesh with estimated marker placement

## Acknowledgement

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.")
The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

## References

[1] Fleishman S., Drori I., Cohen-Or D. "Bilateral mesh denoising" in *ACM SIGGRAPH 2003 Papers*, 2003, pp. 950-953.
[2] Dey T.K., Li G., Sun J. "Normal Estimation for Point Clouds: A Comparison Study for a Voronoi Based Method" in *Proceedings of the Second Eurographics / IEEE VGTC conference on Point-Based Graphics*, Eurographics Association Aire-la-Ville, Switzerland, Switzerland, 2005, pp. 39-46.
[3] Pauly M., Keiser R., Kobbelt L., Gross M. "Shape Modeling with Point-Sampled Geometry" in *Proceedings of ACM SIGGRAPH 2003*, ACM Press, 2003, pp. 641-650.

[4] Zhao W., Gao S., Lin H. "A Robust Hole-Filling Algorithm for Triangular Mesh" in *The Visual Computer: International Journal of Computer Graphics*, Springer-Verlag New York, Inc. Secaucus, USA, Vol. 23 Issue 12, November 2007, pp. 987-997.

[5] George L.P., Seveno E. "The advancing-front mesh generation method revisited" in *International Journal for Numerical Methods in Engineering*, Volume 37 Issue 21, November 1994, pp. 3605-3619.

[6] Shalini Gupta, Mia K. Markey and Alan C. Bovik "Anthropometric 3D Face Recognition" in *International Journal of Computer Vision*, Springer US, 2010, Volume 90, Issue 3, pp. 331-349.

[7] P. J. Besl, H. D. McKay "A method for registration of 3-D shapes" in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, Volume 14 Issue 2, February 1992, pp. 239-256.

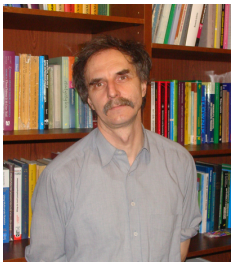
[8] D. Chetverikov, D. Svirko, D. Stepanov, P. Krsek "The Trimmed Iterative Closest Point algorithm" in *Proceedings of 16th International Conference on Pattern Recognition*, Volume 3, 2002, pp. 545-548.

[9] Berthold K. P. Horn "Closed-form solution of absolute orientation using unit quaternions" in *Journal of the Optical Society of America A*, Volume 4, Issue 4, 1987, pp. 629-642.

[10] Dean M. Toriumi, Mark A. Checcone "New Concepts in Nasal Tip Contouring" in *Facial Plastic Surgery Clinics of North America*, Volume 17, Issue 1, 2006, pp. 55-90.

[11] T. Surazhsky, E. Magid, O. Soldea, G. Elber, E. Rivlin "A comparison of Gaussian and mean curvatures estimation methods on triangular meshes" in *Proceedings of IEEE International Conference on Robotics and Automation*, Vol. 1, 2003, pp. 1021-1026.

**Damian Pęszor** obtained his BSc degree in Information Technology in 2011 from Bytom Faculty of Information Technology of the Polish-Japanese Academy of Information Technology and his MSc degree in Computer Science in 2012 from the Silesian University of Technology in Gliwice, Poland. He is currently a PhD student at Silesian University of Technology in Gliwice, Poland. His research is currently centered on the development of performance capture-based facial animation and motion analysis.



**Andrzej Polański** received his MSc degree in Electronic Engineering and his PhD and DSc degrees in Automatic Control from the Silesian University of Technology in 1982, 1990, 2000, respectively. He obtained his professorship in technical sciences in 2009. He worked as Post Doctoral Fellow at the University of Texas, Human Genetics Center, Houston, USA (1996-1997) and as a Visiting Professor at Rice University, Houston, USA (2001-2003). Currently, he is Professor at the Silesian University of Technology. His research interests are in bioinformatics, biomedical modeling and control, modern control, systems theory and optimization theory.

**Konrad Wojciechowski** received his MSc diploma in Electrical Engineering from the Academy of Mining and Metallurgy, Cracow, Poland in 1967, and the PhD and DSc degrees in Control Theory from the Silesian University of Technology, Gliwice, Poland, in 1976 and 1991, respectively. At present he is Professor at the Silesian University of Technology and Polish-Japanese Academy of Information Technology. He received his professoral title in 1999. His area of scientific activity includes linear and nonlinear control theory, neural nets, image processing and pattern recognition, computer vision, computer graphics, animation and games.