

## ViSiCAST Milestone M4-3: Advanced system for physical motion modelling

|                          |   |
|--------------------------|---|
| <b>Project Number:</b>   | IST-1999-10500  |
| <b>Project Title:</b>    | ViSiCAST<br>(Virtual Signing: Capture, Animation, Storage,<br>Transmission) |
| <b>Deliverable Type:</b> | Report / Internal   |

|  |  |
|--|--|
| <b>Deliverable Number:</b>                           | M4-3   |
| <b>Contractual Date of Delivery:</b>                 | February 2002  |
| <b>Actual Date of Delivery:</b>                      | July 2002  |
| <b>Title of the Deliverable:</b>                     | Advanced system for physical motion modelling  |
| <b>Work-Package contributing to the Deliverable:</b> | WP4 (Animation & Modelling)  |
| <b>Nature of the Deliverable:</b>                    | Report / Internal  |
| <b>Author(s):</b>                                    | Nicolas ROUGON<br><br>ARTEMIS Project Unit<br>Institut National des Télécommunications,<br>9 rue Charles Fourier, 91011 EVRY CEDEX<br><br>Phone : +33 (0) 160764644<br>FAX : +33 (0) 160764381<br>Nicolas.Rougon@int-evry.fr |
| <b>Contributors:</b>                                 | Prof. Françoise PRETEUX<br>Dr. Eng. Nicolas ROUGON<br>Dr. Marius MALCIU<br>Eng. Marius PREDA   |

### Abstract:

This reports presents a markerless vision-based motion capture subsystem dedicated to face, which significantly reduces the invasiveness and calibration issues encountered by optical marker-based, camera-head mounted solutions. It relies on a robust model-based 3D/2D registration approach to rigid and non-rigid motion recovery in non calibrated, monocular image sequences, with arbitrary backgrounds and non stabilised lighting conditions. Due to built-in object representation compliance with the MPEG-4 SNHC standard, this subsystem integrates itself seamlessly in a network- and terminal-interoperable standardised software environment for distributed virtual character animation.

# Advanced system for physical motion modelling

## 1. Problem statement: simplifying motion capture using computer vision

Collecting salient data for animating virtual signing characters primarily relies on a motion capture system. The latter is responsible for grabbing the geometry and the dynamics of the various body parts involved in signing gesture, including hands, arms, trunk and face<sup>1</sup>. Compared to most virtual/augmented reality applications involving virtual characters (and notably, in the entertainment domain, to those connected to special effects and gaming industries), virtual signing requires a high degree of realism to guarantee the full restitution of semantic content, both in terms of legibility and emotion. This faithfulness requirement, which crucially conditions the cultural acceptability of virtual signing, sets hard constraints both on motion capture and animation technologies, with direct impact on data volumes to be processed in real-time and compressed for storage and transmission.

The currently operational solution, developed by TELEVIRTUAL, relies on a set of hardware sensors, dedicated to individual body parts, and driven by an integrated real-time software environment, called MASK-VR. The latter provides such core functionalities as physical layer interface, sensor calibration, signal processing, synchronisation and multiplexing, and data storage. MASK-VR is seamlessly linked to TELEVIRTUAL animation software, which allows real-time mapping of motion data onto a 3D avatar model. This enables interactive editing during capture session. The motion capture technologies involved in TELEVIRTUAL system comprise:

- For hand motion capture, data gloves, articulated over the first two phalanges;
- For body motion capture, a magnetic body suit consisting of a magnetic tunic and peripheral magnetic sensors placed on the head, arms, hands, knees and feet;
- For face motion capture, an optical subsystem consisting of reflective markers symmetrically placed at key points (around the mouth, and on the chin, nose, cheeks, forehead and eyebrows) and tracked by a CCD camera mounted on a helmet.

This set-up is representative of advanced first generation motion capture systems, in which hybrid and mostly hardware technologies are combined to ensure global capture of the motion of individual body parts.

Though providing good quality motion data, this system suffers from some intrinsic limitations which restricts its applicability to studio practice and limits its deployment into routine production contexts:

- *invasiveness*: once mounted on an expert signer, the complete system is cumbersome. In particular, connection cables may limit freedom of motion, and magnetic sensors may undergo spurious displacements in time, implying periodical re-adjustment.
- *calibration issues*: mapping motion data onto an avatar model requires implicitly aligning the human signer morphology onto the synthetic 3D model geometry. This so-called calibration step must be performed at the beginning of each capture session, and potentially repeated because of sensor derivation phenomena. In the current stage of development, calibration is performed interactively by an experienced user, thanks to the real-time editing functionality of the MASK-VR software.
- *reproducibility issues*: sensor positioning over face and body requires some expertise and can reveal both important intra- and inter-user variability. The subsequent lack of reproducibility is only partially compensated by the supervised calibration step.

In addition, the accuracy and realism of facial motion are degraded by the use of a sparse set of markers which deliver only pointwise measurements. Propagating motion information over the whole 3D head model requires non-local interpolation which is performed using an underlying bone set structure attached to the outer skin. Hence, a model-induced bias onto the scanned-from-life motion data.

---

<sup>1</sup> Legs can also be involved in some European sign languages, e.g. German Sign Language (GSL).

Considering the evolution of information technologies during the ViSiCAST Project course, some of these limitations have today received industrial answers or could be overcome in a near future. In particular:

- *wireless technologies* such as Bluetooth™ or EDGE are *de facto* candidate to reduce the invasiveness of data gloves and magnetic sensors devices.
- *vision-based motion capture technologies* have known a considerable development, resulting into reliable whole body, markerless and wireless capture systems. Most often, the latter make use of multiple camera (up to 24 for the most advanced ones) together with robust automated sensor calibration and motion tracking schemes supported by dedicated hardware, for achieving whole body non rigid tracking over the whole visual sphere with a high reproducibility degree. Vision-based capture performs well for recovering the motion of the largest body segments, including arms/forearms, legs/forelegs, trunk and head. However, despite intensive research, reliable and automated vision-based hand gesture and facial expression capture still remain open issues. This is mainly due to the large number of degrees of freedom and the important intra- and inter-individual variability to be dealt with when analysing such articulated / deformable objects. Moreover, vision-based hand gesture capture has to deal with an additional level of complexity originating from self-occlusion and contacts phenomena.

The current state-of-the-art in computer vision indicates that a reliable vision-driven hand gesture capture system is today out of reach. Consequently, data gloves are expected to remain the reference technology in the near future, with anticipated evolutions towards increased autonomy thanks to wireless technologies. Within the framework of ViSiCAST Work Package 4, efforts have therefore been focused on developing automated vision-based facial motion capture. This research has been conducted in the ARTEMIS Project Unit at the Institut National des Télécommunications (INT) in tight connection with the Ph.D. dissertation of Marius MALCIU, supervised by Prof. Françoise PRÊTEUX.

## 2. Model-based approaches for facial motion capture in computer vision

The following specifications have been retained:

- *Monocular framework using a single, non calibrated camera*: this aims at designing a cheap, head-free (bypassing calibration allows to avoid using a head-mounted, fixed camera), easy-to-set up and computationally efficient solution. Compared to multiple camera systems, reducing the available visual information motivates, however, an increased algorithmic complexity, which implies shifting from real-time to offline processing.
- *Arbitrary background and unknown, non stabilised lighting conditions*: this assumption is consistent with routine studio practice and avoids light source calibration;
- *Markerless approach*, with the objective of simplifying the system, reducing invasiveness while increasing measurement reproducibility and accuracy. Motion capture is consequently performed in a dense fashion over the whole visible part of the face, without requiring using a bone set structure whose role is confined to animation.
- *Large magnitude, complex motion management*, as a requirement to deal with head and face motion features in sign languages;
- *Robustness with respect to partial head occlusions*: this is necessary to deal with facial occlusions by hands occurring for certain gestures.

Complying with these requirements demands an important amount of *a priori* information about face geometry (global shape and regional features) and dynamics (admissible deformations associated with facial expressions and morphometric variability). This information enables object-oriented detection in dynamic scenes with arbitrary background complexity and lighting conditions, and object-based interpolation in case of partial occlusions. It also must be three-dimensional to allow depth inference from monocular images without calibration cues. The developed approach relies on prototypic geometric models, or *templates*, of head and facial components (e.g. eye, iris, mouth...), able to undergo rigid and non-rigid deformations to adapt to the geometry of image sequences. Template adaptation is driven by visual cues estimated from image data, including optical flow and texture. Model calibration and motion capture are then equivalently reformulated into a rigid/non rigid 3D/2D registration/tracking problem in image sequences.

More precisely, model-based facial image sequence analysis involves two stages with increasing complexity which are performed sequentially:

1. *Global head tracking*: it consists in adapting the scale and 3D pose of the head template using quasi-rigid (affine) transformations. Using a structured model incorporating regional information about facial features results into a rough localisation of facial components corresponding to neutral expression.
2. *Local shape measurement*: following quasi-rigid registration, the shape of the head model is non rigidly adapted by integrating the deformations of individual facial component templates which are estimated by independent registration onto their image counterparts.

In the sequel, these two steps are briefly reviewed. Extensive descriptions of the mathematical and algorithmic underlying the formalisms, as well as complete experimental validation can be found in the below-mentioned references.

### 3. Robust model-based head tracking and 3D pose estimation

For computational efficiency reasons, global head tracking is performed using a simplified 3D mesh template. The latter can be derived either as a mesh-optimised version of the head of the virtual signer model, or generated analytically as an interpolating surface from a collection of samples along the head silhouette viewed under three distinct incidences. For the sake of genericity, the second approach has been preferred. In practice, a 6 harmonics Fourier surface built from 180 sample points has been used.

Model registration is driven by integrating various visual cues derived from the input video sequence. Given local estimates of a cue over a sampling grid in a frame # $n$ , the 3D head template in its current pose is used to predict corresponding estimate values in frame # $n+1$ . Assuming some camera model (e.g. parallel projection) and given an estimate of the pose transform, the prediction process consists in (i) back-projecting the cue estimates in frame # $n$  onto the 3D head template in its reference pose, (ii) applying the estimated pose transform, and (iii) re-projecting the transformed estimates in frame # $n+1$ . Predicted estimates are then compared to actual ones in frame # $n+1$  according to some robust error metrics, yielding an error criterion. Linearly combining the error criteria associated to each cue results in a global error functional which is a (non linear) function of pose parameters. Iteratively minimising this functional with respect to transform parameters using a multiresolution downhill simplex technique yields an optimal 3D pose estimate. The overall method appears to be a non-variational parametric 3D/2D registration technique.

Specifically, the following cues and error criteria have been used:

- the optical flow between two consecutive frames, estimated using Quénot's dynamic programming-based method; it is locally compared to the displacement field induced by the currently estimated pose transform using the  $L^1$  norm.
- a local image texture attribute, simply chosen as luminance; its prediction from frame # $n$  is locally compared to luminance in frame # $n+1$  using the  $L^1$  norm.

In order to account for large translational motions, motion compensation using a block-matching scheme is performed prior to visual cue computation. Accurate pose estimates are obtained by using  $L^1$  norms weighted against a visibility index, locally defined as the orthogonal projection of the template normal onto the image plane. Moreover, robustness with respect to occlusions is achieved by restricting computations to regions jointly visible in the analysed two frames, identified by k-means clustering coupled with parametric modelling of the locally dominant motion.

This method has been extensively validated on a variety of sequences with varying complexity degree in terms of face morphologies, head motions, camera motions, backgrounds, lighting conditions and occlusions. Experiments on synthetic calibrated sequences have underlined its reliability: in 90% of the tests, parameter estimations are accurate up to  $1.5^\circ$  for in-plane rotation,  $4^\circ$  for azimuth angle,  $8^\circ$  for elevation angle, and 98% (resp. 97%) of head size for translation (resp. scaling) parameters. Experiments on natural sequences have confirmed these observations and demonstrated a very robust behaviour with respect to the above-mentioned scene properties.

## 4. Facial feature tracking using deformable templates

Following global quasi-rigid registration, the shape of the head model is non rigidly adapted by integrating the deformations of individual facial components. The latter are estimated by independent registration of *2D deformable templates* onto their image counterparts, starting from initial configurations induced by the globally registered (undeformed) 3D head model. For the first frame, template initialisation is achieved by projecting 3D facial features from the head model onto the image plane; for the remaining frames, the deformed templates optimally estimated in frame #n-1 are propagated in frame #n after quasi-rigid registration.

Deformable templates for facial features are designed to be compliant with the Face & Body Animation (FBA) part of the MPEG-4 SNHC (Synthetic and Natural Hybrid Coding) standard. Each template consists of a collection of MPEG-4 standardised *key points* connected by piecewise continuous curves. Here, interpolating B-splines have been chosen. For such composite templates as the mouth model, virtual linear connections between curvilinear components are also introduced to allow enforcing non local coupling constraints between individual template parts. Two templates are defined for the eyes to deal with iris visibility/occlusion in open/close configurations.

The state of a template is governed by an energy functional. It comprises an internal energy, describing the template intrinsic behavioural properties, and an external energy modelling interactions with the image content. In the developed approach, the internal energy functional combines material properties of the linear elasticity type along curvilinear components, with symmetry properties defined as normalised quadratic constraints along linear coupling units. The resulting template is then equivalent to a constrained mass-spring network. The external energy integrates visual cues derived from image data along the template, in such a way that its minima singularise salient stable registrations. Here, the external energy functional combines textural properties over the template support with feature-specific segmentation information. For the eye template, the latter is a contrast measure between iris and eye white in eye-open configuration. Eye configuration is automatically detected using a fast regularised, fuzzy classification scheme for chromatic Gaussian mixture models based on the Neighbourhood Expectation-Maximisation (NEM) algorithm. For the mouth template, segmentation information consists of the luminance contrast along outer lips boundaries, together with the membership function associated to the lips class resulting from NEM estimation.

For the sake of representation compactness and computational efficiency, the space of admissible transformations undergone by a template (or a template component, in the case of composite models) is chosen as a parametric function space. Specifically, second order bivariate polynomials have been considered. For a given template, the total energy is then a non linear function of deformation parameters, and is minimised iteratively using the downhill simplex method.

This method has been validated on test sequences relative to different people with a large variety of facial expressions. These experiments have allowed to assess the robustness of non rigid motion tracking and the accuracy of the resulting displacement measurements. An example of typical performances is provided on Figure 1.

## 5. MPEG-4 SNHC compliant facial motion capture and animation

Due to the choice of a MPEG-4 SNHC-compliant representation for facial features, the previously-detailed model-based approach for facial motion capture provides direct measurements of the so-called Face Animation Parameters (FAPs) from non calibrated video sequences. Together with the Body Animation Parameters (BAPs), which describe the motion of non facial body segments, FAPs are the primary inputs of the MPEG-4 SNHC suite developed by INT-ARTEMIS for interoperable distributed animation of virtual characters. This suite, which is extensively described in the ViSiCAST Deliverable D4-4 report, comprises a MPEG-4 SNHC Encoder, a MPEG-4 SNHC Decoder and a MPEG-4 SNHC Player. Letting aside compression issues to focus on motion capture aspects, the estimated FAPs can be directly plugged into the MPEG-4 SNHC Player for immediate avatar animation and rendering. In this way, a playback functionality is seamlessly integrated in the facial motion capture system. This analysis/synthesis loop is illustrated on Figure 2 where the underlying 3D structure of the face model, responsible for non local animation from local motion data, is also shown.



Figure 1 – Automatic eye and mouth motion capture in video sequences using deformable templates.

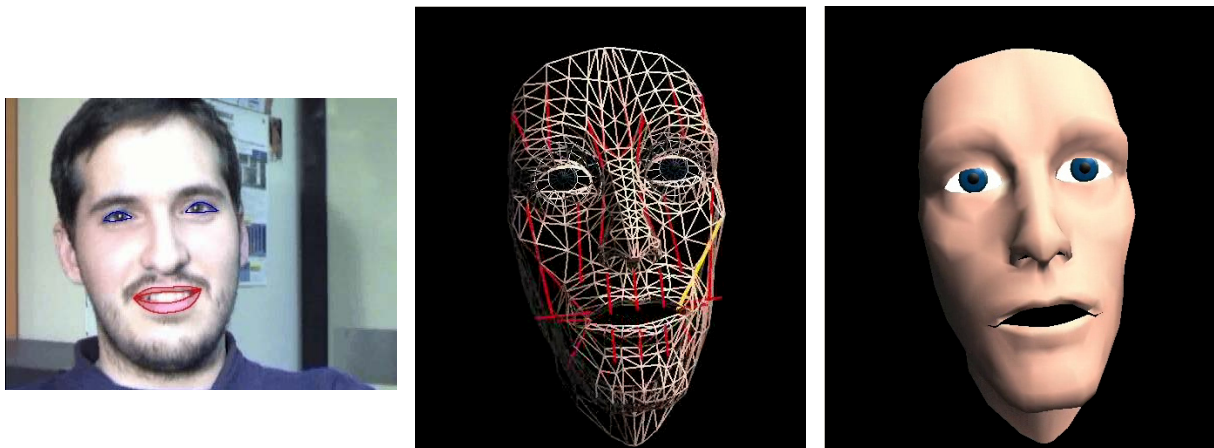


Figure 2 – MPEG-4 SNHC facial animation from motion data automatically captured in video sequences.

## 5. Concluding remarks

This report has presented a vision-based markerless motion capture subsystem dedicated to face, which has been developed by INT-ARTEMIS for automatically grabbing, creating and playing MPEG-4 compliant facial motion content.

This subsystem provides an offline, cheap and easy-to-set up solution for accurate and robust facial motion capture, which significantly reduces the invasiveness and calibration issues encountered by the currently used optical marker-based, camera head-mounted real-time system. Moreover, it integrates itself seamlessly in a network- and terminal-interoperable standardised environment for

distributed virtual character animation, developed by INT-ARTEMIS in the framework of Work Packages 1 and 4 of the ViSiCAST Project.

The original contributions underlying the vision-based facial motion capture methodology have been reported in details in this document. Moreover, they have been largely disseminated in the scientific community of Computer Vision, resulting into a successfully-defended Ph.D. Dissertation.

To conclude, the ViSiCAST requirements as stated in the proposal are completely and successfully achieved.

## 6. Related publications<sup>2</sup>

- Marius MALCIU  
*Approches orientées modèle pour la capture des mouvements du visage en vision par ordinateur*  
Ph.D. Dissertation Université Paris V – René Descartes, December 2001.
- Marius MALCIU, F. PRÊTEUX  
*MPEG-4 compliant tracking of facial features in video sequences*  
Proceedings EUROIMAGE International Conference on Augmented, Virtual Environments and 3D Imaging (ICAV3D'01), Mykhonos, Greece – pp.108-111, May 2001.
- F. PRÊTEUX, C. FETITA, M. MALCIU, M. PREDA  
*Advanced methods for 3D object representation and animation. Application to medical imaging and virtual humanoids*  
Proceedings IEEE International Conference on Communications 2000, Bucharest, Romania – pp. 38-49, December 2000.
- Marius MALCIU, F. PRÊTEUX  
*Tracking facial features in video sequences using a deformable model-based approach*  
Proceedings SPIE Conference on Mathematical Modeling, Estimation and Imaging, San Diego, CA – Vol. 4121, pp. 51-62, August 2000.
- Marius MALCIU, F. PRÊTEUX  
*A robust model-based approach for 3D head tracking in video sequences*  
Proceedings 4<sup>th</sup> IEEE International Conference on Automatic Face and Gesture Recognition (FG'2000), Grenoble, France – pp. 169-174, March 2000.
- Marius MALCIU, L-T. NESSI, F. PRÊTEUX  
*Pose 3D du visage dans des séquences vidéos : estimation robuste par modèle d'objet*  
Proceedings 12<sup>ème</sup> Congrès Francophone AFRIF-AFIA Reconnaissance des Formes et Intelligence Artificielle (RFIA'00), Paris, France – Vol. 1, pp. 24-26, February 2000.
- Marius MALCIU, F. PRÊTEUX, V. BUZULOIU  
*3D global head pose estimation: A robust approach*  
Proceedings International Workshop on Synthetic, Natural and Hybrid Coding and 3D Imaging (WSNHC3DI'99), Santorini, Greece – pp. 79-82, September 1999.
- F. PRÊTEUX, Marius MALCIU  
*Model-based head tracking and 3D pose estimation*  
Proceedings SPIE International Conference on Mathematical Modelling and Estimation Techniques in Computer Vision, San Diego, CA – Vol. 3457, pp. 94-108, July 1998.
- F. PRÊTEUX, Marius MALCIU, S. CURILA  
*Active 3D model-based registration*  
Proceedings Conference on Nonlinear Image Processing – IS&T/SPIE Symposium on Electronic Imaging, Science & Technology'98, San Jose, CA – Vol. 3304, pp. 186-196, February 1998.

---

<sup>2</sup> Available for download on : [www-artemis.int-evry.fr/Publications](http://www-artemis.int-evry.fr/Publications)