Digital Clones and Virtual Celebrities

Facial Tracking, Gesture Recognition and Animation for the Movie Industry

Barnabás Takács[1], Thomas Fromherz,
and Steve Tice[2]
Virtual Celebrity Productions
3679 Motor Ave., Suite 200
Los Angeles, CA, 90034, U.S.A.

Dimitris Metaxas
University of Pennsylvania
Dept. of Computer and Information Science
200 South 33rd St.
Philadelphia, PA, 19104-6389, U.S.A.

Keywords: Facial tracking, Gesture recognition**,** Markerless**,** Production system, Real-time, Special effects

Correspondence:

Barnabás Takács, Director of Research and Development
Virtual Celebrity Productions, Los Angeles.
Tel: +1 310 253-5131 / 212
Email: TakacsB@virtualceleb.com

---

[1] Corresponding author - TakacsB@virtualceleb.com
[2] Currently with QuantumWorks Corporation.

Abstract

We describe a high performance facial tracking and animation solution, called the Digital Cloning System™, specifically designed to meet the needs of the entertainment industry. Special and digital effects are now being used in most feature films, thus representing an important new application domain for facial and body tracking technologies. Virtual Celebrity's system consists of (i) a real-time module responsible for pre-visualization of the tracked data on the face of an animated celebrity, and (ii) an off-line component that suitably meets the needs of high quality special effects and film production. We briefly review the state-of-the-art and present a comparative analysis of needs, methods, and existing solutions in film production vs. Internet and video telephony applications. Experimental results and demonstrations prove the effectiveness of our method.

1. Introduction

Automated analysis and synthesis of facial expressions have been in the focus of face recognition research for the last decade. Experimental systems have been developed and tested for model-based coding techniques that achieve visual communication at very low data rates[1,2] and high compression of 3D animated data. More recently, the advent of Internet-driven avatar technology has created a strong need among users to assume a particular identity in virtual spaces and to freely interact with each other. For both applications, the principal front end is a real-time facial motion tracker that follows the motion of a small set of characteristic points on the face (typically less then 20), and uses gesture recognition to analyze, transmit, and re-animate events related to expressions and emotions[3-6].

We describe a new domain for facial feature tracking, expression analysis and animation which, although similar in formulation, poses a set of completely different technical problems. The digital effects industry (in film production) frequently faces the challenge of having to seamlessly integrate animated 3D models into 2D live-action footage. Such was the case in recent Block Buster films like *Titanic*, *Batman and Robin* and *Face-Off*[7]. In this paper, we present a brief overview of our achievements in advancing the state-of-the-art in facial feature tracking, gesture recognition, and photo-realistic facial animation for this particular application domain.

Virtual Celebrity Productions has designed and built a proprietary *Digital Cloning System™* (DCS) which utilizes a dual tracker architecture and advanced facial animation tools to create photo-realistic/'seamless' human characters[8]. The two major components of our tracker module are: (i) a real-time tracking & animation system called *Geppetto*, and (ii) a very high quality off-line

tracking/rendering pipeline. The DCS module accepts a sequence of film resolution images (4K by 4K pixels, 16 bit / color) and outputs tracking data and high quality 3D animation ready for film production. A detailed description of the *DCS* can be found in [8]. The specific system incorporates:

- A real-time facial tracker and gesture recognizer coupled with an animation software used to pre-visualize the effects of current tracking data.
- A set of specialized facial feature trackers that extract all the necessary information relevant for creating a photo-real *Digital Clone*™.
- A camera tracker that derives the 3D motion of the camera as needed for special effects.
- A highly accurate off-line facial feature tracker that is capable of tracking up to 500 points without markers on an actor's face.
- A deformable model-based 3D face tracker based on the integration of optical flow and edges extracted from the given image sequence[9].
- Interfaces to the major computer animation packages allowing for direct control of any muscle or surface deformation-based facial animation model.

2. High Accuracy Face Tracker and Facial Animation System

The overall architecture of the *VCP Face Tracker* is presented in Figure 1. The basic API consists of multiple layers starting from a set of basic tracking algorithms to 2D/3D and camera motion trackers. On the top of this general purpose API, a set of application specific routines were developed that extract highly detailed 3D data of head motion (yaw, pitch, roll, 3D position) and facial features such as eyes (pupils, eyelids), mouth corners, lips, nose, etc. The user can also specify to track a number of occasionals including wrinkles and miscellaneous points. Once all subtle facial movement is extracted the tracker proceeds to generate animation data, first by rendering a real-time version using VCP's *Geppetto* system, and subsequently creating direct animation data for high end tools such as SoftImage, Houdini, and/or Maya.

We now describe the basic building blocks of the system in detail. The real-time module of the tracking system was developed for producing low-cost animation for Internet and TV. For feature films, the high quality tracking is an off-line process. However, the user can still utilize the power of the real-time mode, as a pre-visualization engine to show what the final animation will look like. In the

off-line mode, any frame in a sequence can be processed at any time and as many times as needed. The processing time is not of major importance; accuracy, however, is. A typical input image would have a 4096x4096 pixel resolution and 16 bits per color channel. When formulating the tracking algorithm, we needed to take into consideration that the presence of highlights, shadows, overlaps, and multiple (often colored) light sources inherently limit the performance of any point-tracker algorithm. Due to these limitations of the imaging process and the total lack of control over the environment, one is required to design an integrated tracking solution comprising of multiple basic algorithms. Table 1. summarizes the differences and similarities of facial tracking and gesture recognition in the entertainment industry vs. the classical domain, i.e. video telephony, avatars, and human-computer interfaces. As is easily seen, the technical requirements of these two seemingly similar problems often require fundamentally different solutions.

Real-time Facial Tracking and Gesture Recognition: As an important component of the *Digital Cloning System*^{TM}, a real-time gesture recognizer and animator module serves to pre-visualize the quality and integrity of current tracking data. It is based on a performance animation system we have developed, called the *Geppetto*^{TM} Animation System, which is able to produce *Interactive Digital Performers*. Unlike other cartoon-like character systems, it captures the subtleties of minute actor eye movements, and while adding subtle animation behaviors, produces a realistic human quality in our characters. The motion capture module was built from off-the-shelf components[10].

Although *Geppetto* has its own built-in real-time tracker currently tracking 12 points on a face, it can also be fed with data from the off-line module of the *VCP Face Tracker*. The points are connected to an underlying muscle-expression system. From this tracking data a mixture of 19 basic muscles and multitudes of expressions are recognized and rendered. Gesture recognition is a two-stage process consisting of (i) an off-line calibration and (ii) real-time tracking & animation. During calibration, a sequence of example facial gestures like 'right side smile' or 'jaw down' are processed, building a relationship between tracked data positions and person specific gesture values. A learning algorithm is responsible for adapting the particular point configuration on the actor's face to the generic muscle model of the celebrity face model. In the animation stage, the pre-visualization module receives tracking data from the *VCP Face Tracker* and translates it into gesture values. This translation is supported by a rule table built up in the preceding learning step, thus making the translation robust and

less ambiguous. The resulting gesture values consist of morph targets, which are applied to a corresponding 3D model of the animated character.

Basic Tracking Algorithms: We implemented a set of basic tracking algorithms that deliver subpixel accuracy. Specific algorithms include region-based tracking algorithms, such as normalized correlation, texture-based trackers (Gabor jets), and optical flow. We found that each of these solutions exhibit serious limitations in a real-world production environment. Figure 2 shows a typical frame from a TV commercial in which the actor acts in front of a green screen later to be replaced by a background image using chroma-key methods. The over 350 points (marked in the figure) were automatically tracked by our system. Note the dark chin line and the low-contrast of the lip contour. Although the face itself is well lit, there are several image regions where highlights and dark shadows might cause the above algorithms to lose points. To overcome the problems associated with each of the above techniques, we designed a hybrid tracking algorithm that combines the advantages of these methods while minimizing the sensitivity to noise and imaging condition. The *VCP Point Tracker* adaptively selects the best method for each tracking point in a given frame. The image is then processed multiple times and the results are integrated, minimizing a cost function defined over all tracking points. Since this is the first layer in the API, we did not put much emphasis on achieving 100% accuracy. Instead, we designed higher level 2D/3D trackers which impose a geometric model on the tracking process and used this model to correct for inaccuracies from lower API layers. This geometric correction layer operates on top of region-based, Gabor jet, and optical flow tracking algorithms.

The second layer of our architecture implements a set of s*tructural trackers* that, instead of following individual points in the image plane, take advantage of an underlying model geometry to which these (individual) points belong. Thus, one can track point groups, 2D structures, and/or splines as well as true 3D objects, imported in the form of a geometry description file. The 3D tracker can also be set to maintain rigid shape or accommodate non-rigid deformations.

Real-time Facial Animation: The *Geppetto* system mentioned above is used as a real-time visualization tool to display how the tracking data modulates the expressions of the animated celebrity. As described earlier, this sub-system uses gesture recognition to estimate muscle parameters from a

small number of tracked points in a form similar to Facial Action Coding (FAC) units, and utilizing morph-targets translates these values to animation data. These morph-targets are built automatically from the high resolution models which our animators create for each film project.

Film-quality Facial Animation (SoftImage/Houdini/Maya): For the film-quality animation, the *VCP Face Tracker* was interfaced with high performance animation packages, such as SoftImage, Houdini, and Maya. The interface controls an iterative off-line process that converts the tracking data into animation data ready for further production stages. The algorithm iteratively deforms the high resolution facial surface using a set of effectors; muscles or channels. First, it estimates the intrinsic camera parameters used to shoot the footage, then using these settings proceeds to find the best muscle configurations responsible for a particular expression on the actor's face.

In the present version of the software, deformations of the facial surface are controlled through a total of 23 channels, 19 of which correspond to the mouth and cheek areas and 4 control the eye movements. While some channels directly relate to anatomical muscles found on the human face, others are linked into virtual muscle groups to help create all possible facial expressions. Each muscle value is estimated to an accuracy of 0.5% on a full range of 0-100%. In addition, we also allow for over-stretching (up to 200%) if the valid range of a particular muscle does not accommodate the tracking data. This feature was found to be very useful in creating, updating, and improving our 3D models. Furthermore, by means of this off-line process, we create a very detailed and accurate data set supporting the learning module of the gesture recognizer above[3].

In summary, we have developed a dual animation architecture bringing the high speed and robustness of gesture recognition together with precise parameter estimation in the model space. The gesture recognition engine (*Geppetto*) offers real-time pre-visualization of the tracked data on the animated person's face using only a limited set of 12 points and low resolution (subsampled) images. It also generates seed expressions for the subsequent film-quality process. The high performance interface (SoftImage, Houdini, Maya), on the other hand, works in an off-line manner, typically processing a given film sequence overnight. The module accepts high resolution images (max. 4Kx4K), and presents its output in an environment suitable for film production professionals. Thus, the

[3] Values of 23 muscles creating facial expressions within 0.5% accuracy are stored together with 2D tracked coordinates and 3D geometry.

conversion of 2D markerless film footage to 3D animation happens in a fully automated way, thus saving time and effort of the visual artists working on the project.

3. Experiments & Results

We have tested the performance and accuracy of our facial tracking and animation solution in the framework of VCP's *Digital Cloning System*^{TM}. We have conducted a series of tests using a variety of cameras and performance scenarios. The real-time module tracks 12 markers and calculates 23 basic muscles and a multitude of expressions at 30 frames per second. The tracking speed of the off-line production system is 1-30 seconds per 4Kx4K frame depending on the number of tracked points (currently we allow a maximum of 500 points), and, most importantly, depending on the required accuracy of the calculated data. Specifically, these processing times correspond to subpixel accuracy tracking in the image domain, 0.1 degree accuracy of global head rotation (yaw, pitch, roll), and 10 mm accuracy in global head and camera motion along the Z axis. Once tracking is complete, the high performance animation module applies the extracted motion and fills the animation channels at a speed of 1-10 minutes per frame.

Figure 3 (left) shows frames from a typical input sequence used in a TV commercial. The actor is placed in front of a green screen to help later compositing. Key facial points are tracked and applied to the high resolution model of another person. The final results corresponding to the input are shown on the right. Figure 4 shows an example of one of Virtual Celebrity Production's digital celebrities. Marlene Dietrich is a German-born actress who became a global icon in the 30's. Two of her original photographs (above) and the corresponding digital model is shown (below). We are currently working on a new project potentially featuring her in a TV commercial. For more information, please visit our websites at http://www.virtualceleb.com and http://www.globalicons.com .

5. Conclusions

In this paper, we described a high performance facial tracking and animation solution, called the *Digital Cloning System*^{TM}, specifically designed to meet the needs of the entertainment industry. The system consists of the real-time component *Geppetto,* responsible for pre-visualization of the tracked data on the face of an animated celebrity, and an *off-line* module that suitably meets the needs of high quality special effects and film production. We developed this dual animation architecture after

reviewing in detail the technical advantages and disadvantages offered to the tracking problem by various solutions. Our experimental results and demonstrations prove the usability of our technique in the targeted application environment. We believe this technology is the first step towards a new age in making films in Hollywood.

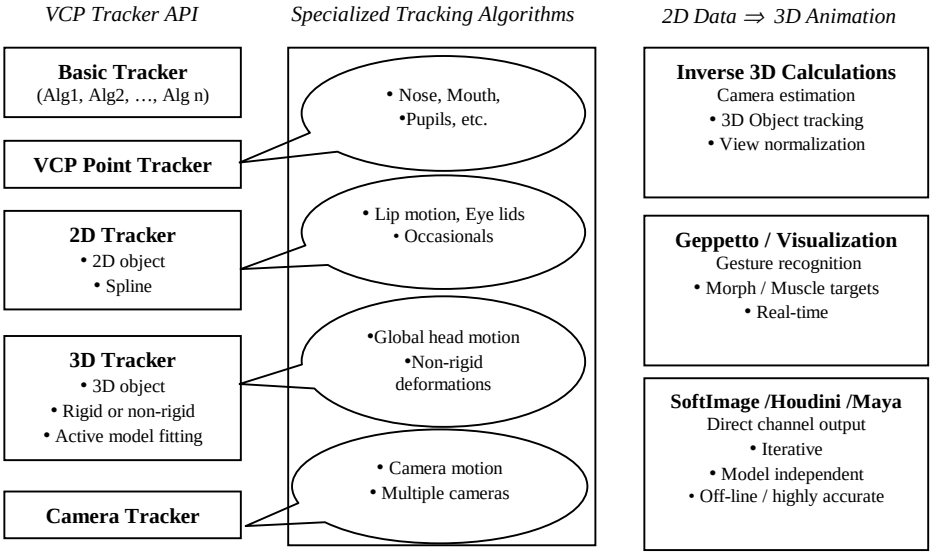
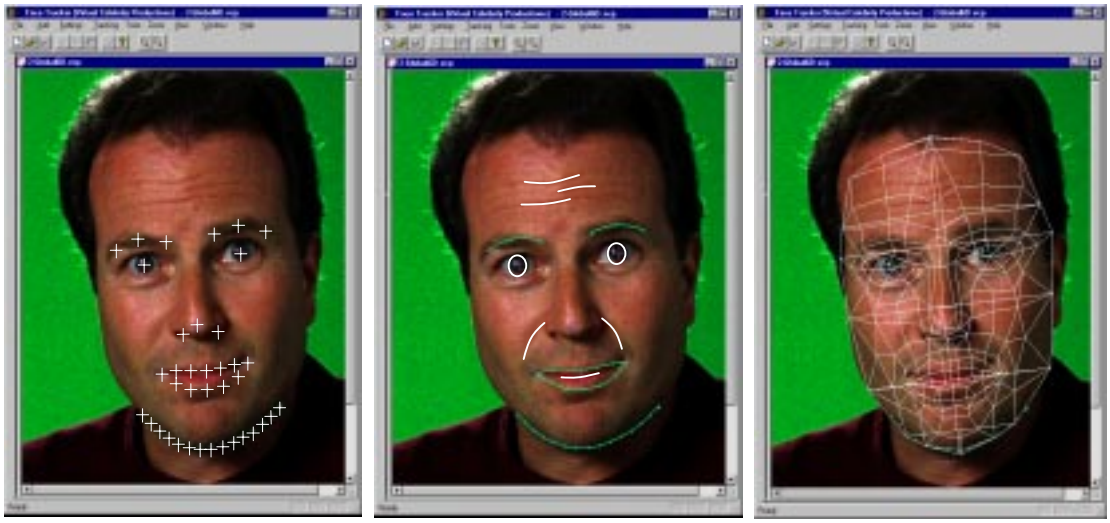


Figure 1: Overview of Virtual Celebrity Production's face tracker architecture.



Eye centers & Pupils (4 pts)
Mouth corners (2 pts)
Nose & Nostrills (3 pts)

Inner/Outer lip contour (40 pts)
Jaw line (15 pts)
Eyebrows (12 pts)

Face mask (200 pts)
Misc. points (64 pts)
Wrinkles & Occasionals (64 pts)

Figure 2: 350+ tracked facial points, fitted splines and global head model used by the VCP tracker.

Figure 3: Typical input (left) and rendered output (right) sequence of a talking head used in a TV commercial.

Figure 4: Original photographs of Marlene Dietrich (above), and her Digital Clone™ at a younger age (below).

	Photo-realistic Film Production	Video Telephony & Virtual Telepresence
GOAL:	Convey acting, emotions, personality	Communication, anonimity & fun
IMAGING:		
# cameras	Single or multiple	Single
Camera Type	Unknown / High performance professional	Unknown / Low-cost commercial
Camera properties	Uncalibrated (mostly unknown)	Uncalibrated (don't care)
Camera motion	YES	NO (stationary)
Image size	Very high resolution	Low to mid-resolution
	4Kx4K / 16bits * 4 color	128x128 or 256x256
Color	YES (principle movie camera)	YES
	NO (high speed & resolution tracking cameras)	
Capture speed	24-30 fps for film	10-30 fps
	30-120 fps for additional tracking cameras	
ENVIRNOMENT:		
Conditions	Uncontrolled	Uncontrolled
Illumination	Changing / multiple light sources, different color	Fairly constant
TRACKING:		
Face Finding:	Semi-automatic / automatic	Automatic
Camera motion	YES - high accuracy 3D tracking	NO
Light source estimation	YES	NO
Markers	Markerless or with some make-up	NO
# tracked points	100-500	5-20
Face Tracking:	Automatic / subpixel accuracy	Automatic / coarse
Global head motion	0.1 degrees accuracy	1-5 degs accuracy for user interaction
Facial features	Highly detailed with subpixel accuracy	Coarse to medium accuracy
	Needs to track unstructred facial regions	Mostly tracks well defined fedutial points
	Pupils, eyelids, gaze	Eyes, blink detection
	Mouth region, inner-outer lips	Mouth corners, lip tracking
	Facial surface / exact 3D data estimation	NO - expression detection only
	Wrinkles & occasionals	NO
Speed:	Real-time and/or off-line	Real-time (15-30 fps)
Processing	Multiple passes, bi-directional	Single-pass, unidirectional
ANIMATION:		
Model construction	Manual / pre-production	Automatic, off-line
Model adaptation	Semi-automatic	Automatic
	Animated individual is a different person	Animated individual is the same person
Gesture recognition	YES (previsualization only)	YES
Advanced Animation	YES (Houdini,Maya,SoftImage,etc.)	NO
Photo-quality	YES	Currently NO, but required in the future
Compositing	YES / semi-automatic	YES / automatic
Postproduction	YES	NO
MISC.:		
Voice data:	Not always available	YES
Multi-modal	YES (visual & voice)	YES (visual & voice)

Table 1: Comparison of facial tracking specifications for the movie vs. the communication / Internet industry.

7. References

[1] Pearson, D.E. and J.A. Robinson (1983) , Visual Communication at Very Low Data Rates, *Proc. IEEE,* **93**(4), 795-812.

[2] Choi, C.S, K. Aizawa, H. Harashima, and T. Takebe (1994), Analysis and Synthesis of Facial Image Sequences in Model-Based Image Coding, *IEEE Trans. On Circuits and Systems for Video Technology* **4**(3), 257-275.

[3] D. Terzopoulos, F. Parke, D. Sweetland, K. Waters, M. M. Cohen (1997), SIGGRAPH 97 Panel on Facial Animation: Past, Present and Future, http://mambo.ucsc.edu/psl/sig97/siggraph97-panel.html, Los Angeles.

[4] K. Waters, T. Levergood (1995), DECface: A System for Synthetic Face Applications, *Multimedia Tools and Applications*, **1**,349-366.

[5] K. Waters, J. Rehg, M. Loughlin, S. B. Kang, D. Terzopoulos (1996), Visual Sensing of Humans for Active Public Interfaces, *Digital Cambridge Research Lab TR 95/6*.

[6] Y. Yacoob, L. Davis (1994), *Computer Vision and pattern Recognition Conference*, chapter Computing Spatio-Temporal Representations of Human Faces, *IEEE Computer Society*,70-75.

[7] Robertson, B (1998), Right on Track, *Computer Graphics World,* **5,** 54-66.

[8] Takács, B. and T. Fromherz (1999), VCP Face Tracker and Animation Software, *Tech. Rep. TR-VCP99-01*, Virtual Celebrity Productions, Los Angeles, California.

[9] DeCarlo,D. and D. Metaxas (1996), Deformable Model-Based Face Shape and Motion Estimation, *ICAFGR*, 146-150.

[10] Tice, S.E. and M.J Fusco (1993), Digital Content Creation Systems, *Character Motion Systems Course #1, SIGGRAPH '93*.