

Applications of Face Analysis and Modeling in Media Production

Darren Cosker
University of Bath, UK

Peter Eisert
Humboldt University Berlin

Oliver Grau
Intel Visual Computing Institute

Peter J.B. Hancock
University of Stirling, UK

Jonathan McKinnell
BBC R&D, London

Eng-Jon Ong
Surrey University, UK

This article surveys automatic facial analysis and modeling methods using computer vision techniques and their applications for media production.

The visualization, interpretation, recognition, and perception of faces are important elements of our culture. Facial expressions carry important messages involved in communication and therefore play an important role in media. Faces provide essential information that allow us to identify individual people, perhaps even from just one frontal photograph. In the last few decades, computer vision techniques dealing with automatic processing of facial image information have become mature enough for many applications, including broadcasting and visual media production. The aim of this

article is to look at state-of-the-art computer vision techniques related to “faces.” This includes both cognitive methods (such as facial detection and recognition and 3D facial modeling) and applications of these methods in media production and access through digital services.

Broadcast and movie productions dedicate a great deal of know-how and technical equipment to the capture of people and their facial expressions. This starts with professional lighting with special care being taken in reproducing skin tones accurately in camera systems. Movie productions spend a lot of time and effort in planning and capturing the “perfect” setup that includes carefully staged acting and use of technical equipment. If necessary, the capture of a scene is repeated until the result is acceptable, with further improvements being achieved in postproduction. Budgets are usually much smaller in the broadcast industry, and postproduction is kept to a minimum or even is impossible if the production is broadcast live. In this context, there is a demand for automatic tools that support the production process.

Another important aspect of the production process is the generation of metadata. Currently, metadata is predominately generated by human operators who log key events, such as the approximate time when individual people appear in a shot and so forth. For denser and more accurate time-wise logging, automated methods are desirable. The ability to detect and identify individuals or actors by their faces will potentially play an important role in future logging systems. Related applications arise for archived video material. Professional users might want the ability to rapidly find footage of certain individual people. An end user might want to search for certain types of programs, their favorite actors, or a combination of both.

In this article, we give a brief overview of the psychology of face perception and then describe some of the applications of computer vision and pattern recognition applied to face recognition in media production. We also cover the automatic generation of face models, which are used in movie and TV productions for special effects in order to manipulate people’s faces or combine real actors with computer graphics.¹

The Psychology of Face Perception

This section outlines some of the key findings in face perception from psychology. Although we cannot cite all the original sources, fortunately there are three excellent recent surveys

of the field.²⁻⁴ A human's ability to recognize familiar faces is remarkably effortless, most of the time, although we are all aware of errors with less familiar people. However, for unknown faces, even matching two photographs is surprisingly inaccurate. Performance is typically around 70 percent correct, even with good quality images taken from the same viewpoint. Variations in viewpoint, expression, and lighting cause additional problems. Multistranded films, such as *Love Actually* (2003), rely on a cast of familiar actors to enable the audience to follow various characters.

There are two primary routes to recognizing a face: the individual features and the relationship between them, or "configuration." It is possible to recognize a face from isolated features and also from an image so blurred that the features are obscured, with only the configural pattern remaining. We perceive faces holistically, meaning that changes to one part affect the perception of others. Thus, it is easier to recognize an isolated top half of a face than a top half that it is aligned with an unrelated bottom half.

Color is surprisingly unimportant for face recognition; it is possible to invert the color palette of an image with little effect, although color can help if identification is uncertain. Unfamiliar face recognition (that is, recognizing someone you have seen before but do not know well) is strongly affected by "external features" such as hair, although there is at least anecdotal evidence that this is more relevant for European faces, where hair varies a lot, than, for example, East Asian faces. As we become more familiar with a face, we learn the internal features, especially becoming more sensitive to changes in the eye region. Nevertheless, we are strongly affected by our expectations, which became clear in the case of a famous image of Al Gore behind Bill Clinton where in fact the internal features of both faces were identical.

There are also clear "other race" effects pervading face perception that result in another layer of unfamiliarity—unfamiliar faces are hard, but those of another race are even harder. There appear to be multiple reasons for the deficit, including simple familiarity; thus greater contact with a race makes for better performance, but also more social cues, such as failing properly to process "outgroup" faces.

Our perceptual systems adapt rapidly to the environment. Thus, if you look at a waterfall for a few seconds and then at the bank, the

latter will appear to drift upward. The adaptation occurs with faces as well; if you look at a male face for a few seconds, then an ambiguous face will look female, whereas if you study a female face, then the same ambiguous face looks male. The effects also apply to identification, so it is possible to affect our memory of what someone looks like merely by studying a distorted version of their face or even by thinking about the person concerned.

An intriguing aspect of our familiar face recognition is the recognizability of caricatures. Cartoonists' drawings are often absurdly distorted and yet instantly recognizable. This is partly an iconic effect—a huge smile is Tony Blair if male or Julia Roberts if female. It is possible to generate caricatures automatically by accentuating deviations from an average face. Early work suggested that these caricatures may be more recognizable than veridical images but it seems there is little, if any effect for photographic images, except for short presentation times. There is some benefit, however, for facial composites of the kind made by a witness to a crime; caricaturing can make these images more identifiable.

Applications in Broadcasting Program Productions

Program creation is becoming increasingly demanding in terms of the amount of content, multiple platforms for delivery, shorter timescales, and ever tighter budgets. New methods of delivering programs to the viewer are resulting in greater choice both in terms of program genre and delivery format. Multiple delivery formats are often required, which presents technical as well as editorial challenges. Automatic generation of metadata is crucial to allowing the cost-effective production of broadcast programs and enabling new access possibilities, such as offering viewers specialized search functions.

Metadata can be created and used from production and ingest right through to the editing or nonlinear delivery of the content. By increasing the amount of metadata available in the editing process, content can be sorted and searched with greater speed, saving time and therefore money. New ways of editing, with multiple timelines based on the content within the scene (such as a timeline of all scenes containing a specific character) could facilitate greater speed while improving editorial storytelling. For a viewer to consume content in a

nonlinear fashion, metadata is crucial for creating a rich set of possible links and connections between different programs. It is desirable that this metadata does not have to be (re)entered at the end of the program creation workflow, which could be achieved by taking a selected set of the metadata logged or automatically tagged in the production (which has resources available for recording metadata) available on delivery.

In addition to attempting to increase the amount of metadata available in the edit and further downstream in the program creation lifecycle, metadata accuracy is important. For example, it can be qualitatively observed that repetitive tasks (such as repetitive time-based logging) can be inaccurately performed by humans, especially when they are distracted or working over long periods of time. Also, the detail and variety of the logs can decrease over time because of other constraints on the logger or production assistant's productivity. Ideally, the logger or production assistant (PA) would have their time freed up by using automated systems such that they can concentrate on the higher-level tasks such as logging a potential lack of journalistic integrity or the artistic quality of the shot. For ingest into the archive of previously broadcast content, the sheer enormous volume of the audio and video content, and therefore the fast throughput of the ingestion task, makes the general accuracy of human logs nonstandardized and nonuniform.

Identifying who and how many people are in a shot can aid in the categorization and searching of content, so it is the automation of these logs that is of most interest. A large amount of research has been undertaken lately to improve the automation of the face recognition task, and recently reasonable success for a variety of applications has been reported.

For test purposes, the application of face detection and recognition was applied to two different TV programs at the BBC: a business discussion program called "The Bottom Line," which involved a live "face" training process, and a drama called the "EastEnders." The large variety of different characters in the drama program, along with greater variations in pose, expression, and lighting, made recognition a difficult computer vision problem to solve. Specifically, a features-based learned boosted classifier detection algorithm and a 2D statistical eigen-space recognition algorithm were used because of their ease of implementation and

real-time performance. More state-of-the-art recognition methods could be employed for higher recognition accuracy in the future when the initial aim is the qualitative analysis of the possible use case of such methods in a production.

"The Bottom Line" is a business show for both television and radio with Evan Davis in which the guests (typically three or four) are entrepreneurs and leaders from the business community. In general, it is not possible to train the recognition algorithm before filming the program because the guests may not have had much previous media coverage and therefore a preexisting database of tagged faces of that individual may not be available. Therefore, it is desirable to use a training program for "live" training on the set. However, it must be easy to use and only take a minimal amount of logger/PA time.

For ease and rapidity of use during a production, a GUI was developed. The faces can be identified as being the same person by tracking the position of the face in the image. If the position of the detected face does not vary too much from frame to frame, it can be surmised that this is the same person's face. In this way, faces can be saved as sequences, and instead of presenting the user of the training GUI with individual faces to tag, the user can tag a sequence of faces all at once.

In a drama series such as the BBC's "The EastEnders," program training need not be done on each production because that is only necessary in principle for new characters. In this way, a training set for a series can be built up before a production and then the recognition algorithm can be run on a set without live training.

Drama, however, is intrinsically a more difficult face recognition problem than interview programs. Interview programs are more formalized, leading to a greater similarity of pose and lighting and often requiring recognition from a relatively small set of different people. Drama programs, on the other hand, have a greater variety of expressions, poses, and lighting. For example, character's faces can change over time because of aging or other factors requiring a retraining of the face recognition algorithm, and drama footage can occur indoors or outdoors, with a larger variety of shot types. In drama, the faces may be small in comparison to the size of the overall image, while this is rarely the case in an interview program.

In the trial use of facial recognition tools on BBC programs, qualitatively, the performance

during interview programs was found to be relatively good when compared to a human logger, and the training GUI was relatively easy to use during a production workflow. Further improvements could be made to the training process by, for example, automatically loading pretrained face sets for the presenter. Drama programs were found to be a greater challenge, however, because of the difficulties outlined here. In the future, by making use of recent advances in recognition algorithms, performance comparable to a human logger may be possible.


3D Facial Capture and Analysis

In recent years, there has been increasing interest in facial animation research from both academia and the entertainment industry. Visual effects and videogame companies both want to deliver new audience experiences—whether that is a hyper realistic human character, such as the reverse aging main character in *The Curious Case of Benjamin Button* (2008), or a fantasy creature driven by a human performer, like in *Avatar* (2009). Having more efficient ways of delivering high-quality animation, as well as increasing the visual realism of performances, has motivated much academic research in recent years.

Similarly, developments in academia both in computer vision and graphics, such as the detailed capture of moving surfaces in 3D and the dense nonrigid tracking of surfaces, have fed (and are still feeding into) movies and videogames. This section gives an overview of some of the key recent work in facial capture and animation. In this context, we will consider two avenues in recent work: static realism and capture as well as dynamic animation and capture.

Static Realism and Capture

Static facial realism is arguably at a level where humans now cannot distinguish real photographs from the best 3D facial models. Technology such as Light Stage allows the capture of highly detailed facial normal maps.⁵ Coupled with subsurface reflectance data,¹ facial models now display photorealistic likenesses to real faces. Such technology is now widely used in modern motion pictures—*Spider Man 2* (2004) being one high-profile use of Light Stage. In these circumstances, the high detail static scans are composited onto either a stunt actor's body or a digital double. This is

used when the area an actor is to be placed in a scene may not be practical or  (such as during an explosion). However, the actor's expression remains static in these situations, and close examination of such shots reveal a dead-like facial quality.

An early, high-profile example of facial replacement can be seen in the *Matrix* sequels,⁶ *The Matrix Reloaded* (2003) and *The Matrix Revolutions* (2003); passive facial scanning was used to obtain 3D faces with high detail facial texture. Fast subsurface scattering of the skin allowed shots to be rendered at a high throughput to meet the demands of the movie. An important aspect of the use of facial scans for movies and videogames is that faces must be renderable in a wide range of environments so that they can be convincingly composited into the overall scene. Therefore, the UV map (texture data) is typically diffuse albedo. Skin detail is enhanced through high-resolution normal maps⁵ or geometry,⁷ and rendering is enhanced through bidirectional reflectance distribution functions (BRDF). Acquisition of such BRDF detail has advanced significantly in recent years, resulting in highly detailed renderings.¹

Also important to the realism of synthetic humans is the realistic rendering of hair, which has made significant progress in the last few years.⁸ Similar to the BRDF modeling of skin, single hair fibers have been modeled as semi-transparent elliptical cylinders. By defining surface reflection as well as scattering inside the hair, the complex lighting characteristics of real hair with its view dependency, highlights, and color changes can be accurately reproduced with moderate rendering complexity. The possibilities of modeling several fibers up to a complete hairstyle range from nonuniform rational B-spline (NURBS) surfaces via thin shell volumes to strain-based modeling by parameterized clusters, fluid flow, or vector and motion fields. To simulate the complex lighting interaction between strands of hair, Tom Lokovic and Eric Veach proposed a deep shadow map that relates visibility to depth for each pixel, yielding realistic but computationally expensive self-shadowing.⁹ Only recently, approximation algorithms for making the simulation of multiple scatterings among hair fibers tractable have been proposed using methods such as photon mapping or spherical harmonics.

Whereas laser scanning technology was initially the most accurate way to derive static facial detail, passive scanning technology using



Figure 1. Static reconstruction of the head including hair from two images.

consumer hardware is now popular.⁷ Multiple consumer-level single-lens reflex (SLR) cameras are used to acquire high detail images. These provide strong features for stereo matching algorithms and can result in captured geometry with skin pore (mesoscopic) level facial details. One aspect to consider when using such data is practicality, as the meshes can contain millions of vertices. This approach differs from those methods currently considered in movies, where a low polygonal mesh is used along with high detail normal maps to display a facial mesostructure. There is therefore still a great deal of work to be done on practically using such technology for videogames and modern visual effects (VFX).

Many state-of-the-art stereo and multiview approaches are local in the sense that they reconstruct the 3D location and sometimes orientation of isolated image patches. Although this strategy is beneficial for parallelization, it requires a post-processing stage to generate a mesh: The reconstruction yields a point cloud with outliers that has to be filtered and meshed with appropriate algorithms such as Poisson meshing. Smoothness priors are often only considered at the meshing stage. Local reconstruction is difficult to combine with efficient interactive tools: Because each patch is unaware of its neighbors, the correction of a single mismatched patch by the user will not affect its neighbors, although they are likely to be erroneous as well. Therefore, the work of David Schneider and his colleagues¹⁰ follows an approach similar that of Thabo Beeler and his colleagues,⁷ but Schneider uses a mesh-based deformable image alignment for the reconstruction of high detail face geometry (including hair)

from two or more SLR cameras, as shown in Figure 1. Instead of iteratively matching small image patches along the epipolar line, an entire view is warped to target views in an uncalibrated framework incorporating a mesh-based deformation model. The additional connectivity information enables the incorporation of surface dependent smoothness priors and optionally user guidance for robust and interactive geometry estimation.

As previously mentioned, most digital face replacement in movies involves static face replacement, with the character having no facial expression. Although this might be satisfactory for a few frames, as soon as the face moves, or the shot continues for more than a few seconds, this illusion becomes hard to maintain. In the next section, we consider the movement of faces, especially with respect to maintaining an illusion of realism.

Dynamic Capture and Animation

The holy grail of facial animation research is the portrayal of characters indistinguishable from real humans. This is extremely difficult because humans are experts at detecting the slightest flaws in faces. To display a synthetic human that is truly life like, the movement of the face remains a major challenge.

Arguably, we are currently more successful when conveying dynamic realism in the playback of recorded dynamic performances than the authoring of new animation. To highlight this, we will first consider the acquisition of dynamic 3D facial sequences (which we call *4D*, for brevity).

There are now many commercial companies that market 4D facial capture systems—that is,

those that can obtain 3D mesh data at video recording rates (such as Dimensional Imaging and 3DMD). However, we concentrate here primarily on academic research in this area. One of the first compelling uses of dynamic facial capture in movies was in the *Matrix* sequels.⁶ A passive stereo capture system was constructed where 3D mesh data can be acquired from a face at a video rate along with a high-resolution texture. The recorded sequences are then composited onto the actors in key action sequences. An extension of this system called Universal Capture was later used in many Electronic Arts (EA) promotions and videogames (including *Tiger Woods Golf*). Here, the system was made more robust by adding markers to the actor's face. This could be used to stabilize and track a canonical mesh (a mesh with a known topology) through the captured sequence. Bernd Bickel and his colleagues adopted a similar approach with the addition of extra facial paint to appropriately capture wrinkles on the face.¹¹

The use of markers has overcome previous issues related to tracking such a mesh using optical flow. Such methods are notoriously susceptible to drift, caused primarily by fast facial changes such as during speech. The first approaches to avoid the drift in markerless tracking over longer sequences incorporated additional constraints from the silhouette or the use of an analysis-by-synthesis estimation. Both methods ensure that estimates are referred to a global reference and avoid error accumulation over time. This approach is also followed by Derek Bradley and his colleagues,¹² who proposed a multiview stereo capture system consisting of 14 HD cameras mosaiced together. This results in a highly detailed set of images upon which to apply optical flow for mesh tracking. Referencing the initial frames of the sequence results in improved mesh stabilization over time. Expanding further on this work, Thabo Beeler and his colleagues introduced the concept of anchor frames for stabilizing 4D passive facial capture.¹³ In their work, neutral frames in the sequences are searched for and then used to essentially reinitialize mesh tracking where possible. This also has the added benefit of offering robustness to certain facial self-occlusions (for example, caused by the lips). Although having perhaps a lower geometric resolution than previous passive static capture work,¹³ the extension to 4D including the impressive temporal mesh coherence is a high current benchmark.

In industry, 4D capture technology has often been described as “volumetric capture.” One highly successful recent demonstration of this is from the videogame *LA Noire*. Hundreds of hours of actor footage were recorded in a controlled lighting environment. Key 3D character scenes were then composited with the volumetric facial performances resulting in highly detailed and realistic models.

Another high-profile use of 3D technology for industrial use was the Digital Emily Project, a collaboration between Imagemetrics and the University of Southern California using Light Stage technology to capture high detail normal map and surface reflectance properties from an actor's face.¹⁴ A facial blendshape rig was constructed from captured 3D data and then matched to the performance of the actor using proprietary Imagemetrics markerless facial capture technology. Blendshapes are facial poses of different expressions—from stereotypical (happy, sad) to extremely subtle (narrow eyes). The term “rig” is used to describe the complete facial model with all of its control parameters. The facial rig's degrees of freedom depend on the number and complexity of the blendshapes, and new facial poses are created by combining blendshapes with different weights.

Although the *LA Noire* production is a high profile use of volumetric capture, it is essentially a playback of the captured 4D video. The Digital Emily Project, on the other hand, demonstrates a degree of performance-driven animation. In this type of animation, a performer animates a puppet via motion capture or speech (audio only or phonemes). A detailed review of these methods is beyond the scope of this article, but the interested reader is encouraged to read the excellent related course material.¹⁵ Our aim here is to make the distinction between direct playback of captured volumetric animation and the creation of realistic characters given some reference (such as actor performance).

The movie *The Curious Case of Benjamin Button* is a good example of successful human realistic performance-driven animation. MOVA (www.mova.com) performance-capture technology was used to collect 3D scans of Brad Pitt's face and used for blendshape rig construction. Animation was then carried out using markerless performance mapping from reference footage of the actor. *Avatar* also pushed forward the realism of facial performance. Although the characters were not human, the movie demonstrated the usability of modern

facial technology for productions requiring a large volume of high-quality performances. The production also used head-mounted cameras targeted at the actor's face and recording the movements of painted markers. These movements were transferred into a combination of blendshapes per frame, and the resulting animation is used as a first pass for artists who edit and enhance the performance with the aid of additional video reference.

This is an important point when considering performance-driven animation approaches. Although they can be used to animate a virtual target model, the result is not commonly used directly as an output on screen. For movie productions, as well as videogames, the performance may just be used to estimate performance timing. An artist will then tweak and add to the performance later. This may be for artistic reasons because the performance transfer is lacking some subtle details or contains errors.

Although marker-based motion capture techniques are widely popular, for example, using commercial optical capture systems or painted markers, markerless methods provide the potential to capture areas of the face where marker placement is too obtrusive. In addition, it raises the possibility of obtaining a dense capture field for the face, for example, based on skin pores. Where the facial rig is based on blendshapes,¹⁵ the aim is to optimize a set of weights that approximate the positions of the markers. In markerless systems, such markers might be located using image-based deformable tracking techniques such as active appearance models. Another alternative is to fit the blendshapes to 4D surface data. This latter method has also been shown to work with consumer 2.5 D capture devices such as the Microsoft Kinect system.¹⁶ However, whether practical use of this technology on a set might be hindered by uncontrolled environmental changes, or actor aversion to the active IR projected pattern, is unclear.

In all the examples so far, dynamics have been captured and replayed, often with considerable artistic manual intervention.¹ However, the concept of using such data to author entirely novel performances without reference footage remains a difficult challenge. The success of such methods still largely depends on artistic talent. However, advances in interactive facial models and new methods to create efficient rigs are promising avenues for improvement.

How do we create effective blendshapes? A standard approach in modern VFX is to select

them based on action units (AUs) from the Facial Action Coding System (FACS). This technology was used in the movies *Monsters House* (2006) and *Watchmen* (2009). Having a FACS basis can potentially provide a mapping between different facial rigs. This can be especially useful if one blendshape model is based on actor facial scans and fitted to an actor; then the weights are transferred onto a puppet model (perhaps of a creature). More recently, researchers considered creating blendshape rigs given only a few example expressions and a generic blendshape rig with a wider number of expressions.¹⁶ Such systems can potentially reduce the time required for an artist to manually sculpt blendshapes for rigs. Facial rigs in movies can also potentially become very large, with hundreds of blendshapes for "hero rigs."

Facial animation basis (or *blendshape basis*) are not restricted to artistically sculpted facial expressions or captured 3D scans. Principle component analysis (PCA) also offers a basis for animating faces. However, although this basis is orthogonal—meaning that each expression has a unique solution with respect to the basis—these are often not intuitive enough for artistic animation. To address this, J. Rafael Tena and his colleagues recently proposed a region-based PCA modeling approach that allows more intuitive direct manipulation of local facial regions.¹⁷ Their method also highlights how solving for expression weights locally can provide better approximation of motion capture data. Ultimately, however, an artist will desire a set of controls from the facial model that are both intuitive and orthogonal such that altering one expression does not interfere too much with others. To counter this, blendshape rigs become highly complex, with additional shapes included to counter these interference cases. This is not desirable for efficiency reasons, so future work is still required to address this core problem.

Automated Lip Reading

One area that makes important use of facial modeling and analysis is automated lip reading. An automated lip-reading system attempts to predict the content of a subject's speech based on an analysis of the lip movements. Potential applications would be to facilitate and improve speech recognition by combining and making use of both audio and visual information for recognizing the speech of a subject. Any attempt at automatic lip reading needs to address a number of demanding challenges.

The first challenge involves automatic facial feature tracking, which is a nontrivial problem because the face is a highly deformable object. For example, lips are highly deformable and can assume a variety of shapes. This difficulty is compounded by the potential appearance and disappearance of the teeth and tongue during speech, causing the inner lip's texture to change dramatically. Other parts of the face can contain extremely fast movements, for example, the eye shape can change from an open eye to a closed eye in the period of a single frame. There are also areas of the face that are challenging to track directly, such as points on a cheek where the texture can be homogeneous.

Accurate and robust facial feature tracking can be approached by means of a learned, person-specific, data-driven approach using only pixel intensities.¹⁸ A crucial component is the ability to automatically locate visual support that is optimal for tracking a particular point on the face (such as a mouth or eye corner). This allows us to potentially track any point on the face. Importantly, this includes points on regions where the visual complexity is high because of potential texture changes (for example, the inner lip) and facial features that are challenging because of the lack of texture (for example, points on a cheek). One commonly used method for tracking is the linear predictor (LP) floccs method. Each LP provides a mapping from sparse template differences to the displacement vector of a tracked facial feature. Multiple LPs can then be grouped into rigid floccs to track a single feature point with greater robustness and accuracy.

The next challenge involves dealing with the inherently temporal nature of the problem. It is not possible to simply locate a set of static visual features that can differentiate between two sets of spoken speech. Instead, it is crucial to model and use spatio-temporal information. However, other challenges arise from motion and appearance variations. The degree of movement of the mouth during speech also tends to be less than that of emotions and other typical forms of actions recognized, and variations are present across different individuals in terms of different mouth shapes, the possible presence of facial hair, and different styles of lip movements while speaking the same words. Using this tracking method, lip shape and appearance information can be extracted from the facial image.

Various machine learning approaches can then be used to learn classifiers for performing

lip reading. One popular method for classifying spatio-temporal data is the hidden Markov model (HMM), where temporal information is represented as a Markov model over a statistical distribution over lip appearance and shape features.¹⁹ Another method utilizes sequential patterns, an ordered sequence of feature subsets. Sequential patterns are used to form weak classifiers that are combined together into a strong spatio-temporal classifier using the boosting method.²⁰ However, open challenges still remain for the lip-reading task in terms of robustness to changing environmental conditions (such as lighting) and being able to deal with speech coarticulation, where the lip shape is affected by past, present, and future spoken words.

Conclusions

Techniques for automatic processing of faces in images have become mature as a result of progress in computer vision and image processing technology. Faces are an anchor point in communication and media production. Using recognition techniques for faces and or lip reading is attractive because it could save money in productions. Moreover, it enables new possibilities in the digital production flow and many of these we just start to see emerging.

Face modeling techniques are generally used in high-profile movie and videogame productions, mainly because of the costs involved. With the progress in techniques able to operate on consumer-level hardware, such as SLR cameras and depth sensors, the budget threshold will lower, enabling the use of these techniques in other areas, such TV production, and in end-user applications.

MM

Acknowledgments

This article covers topics presented at a workshop on faces at BBC R&D 2011, London.

References


1. C. Donner et al., "A Layered, Heterogeneous Reflectance Model for Acquiring and Rendering Human Skin," *ACM Trans. Graphics (Proc. SIGGRAPH Asia 2008)*, vol. 27, no. 5, 2008, pp. 140:1–140:12.
2. V. Bruce and A. Young, *Face Perception*, Psychology Press, 2011.
3. A. Calder et al., eds., *Oxford Handbook of Face Perception*, Oxford Univ. Press, 2011.

4. G. Hole and V. Bourne, *Face Processing: Psychological, Neuropsychological and Applied Perspectives*, Oxford Univ. Press, 2010.
 5. W. Ma et al., "Rapid Acquisition of Specular and Diffuse Normal Maps from Polarized Spherical Gradient Illumination," *Proc. Eurographics Symp. Rendering*, Eurographics Assoc., 2007, pp. 183–194.
 6. G. Borshukov et al., "Universal Capture: Image-Based Facial Animation for *The Matrix Reloaded*," *SIGGRAPH Course Notes*, ACM, 2005, article no. 16.
 7. T. Beeler et al., "High-Quality Single-Shot Capture of Facial Geometry," *ACM Trans. Graphics* (Proc. ACM SIGGRAPH 2010), vol. 29, no. 4, 2010, article no. 40.
 8. K. Ward et al., "A Survey on Hair Modeling: Styling, Simulation and Rendering," *IEEE Trans. Visualization and Computer Graphics*, vol. 13, no. 2, 2007, pp. 213–234.
 9. T. Lokovic and E. Veach, "Deep Shadow Maps," *Proc. 27th Ann. Conf. Computer Graphics and Interactive Techniques* (SIGGRAPH), ACM, 2000, pp. 385–392.
 10. D.C. Schneider et al., "Deformable Image Alignment as a Source of Stereo Correspondences on Portraits," *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition* (CVPR 2011), IEEE CS, 2011, pp. 41–52.
 11. B. Bickel et al., "Multi-scale Capture of Facial Geometry and Motion," *ACM Trans. Graphics* (Proc. ACM SIGGRAPH 2010), vol. 30, no. 3, 2011, article no. 33.
 12. D. Bradley et al., "High Resolution Passive Facial Performance Capture," *ACM Trans. Graphics* (Proc. ACM SIGGRAPH 2010), vol. 29, no. 4, 2011, article no. 41.
 13. T. Beeler et al., "High-Quality Passive Facial Performance Capture Using Anchor Frames," *ACM Trans. Graphics* (Proc. SIGGRAPH 2011), vol. 30, no. 3, 2011, article no. 75.
 14. O. Alexander et al., "The Digital Emily Project: Achieving a Photoreal Digital Actor," *IEEE Computer Graphics & Applications*, vol. 30, no. 4, 2010, pp. 20–31.
 15. P.F. Pighin and J.P. Lewis, "Performance Driven Facial Animation," *SIGGRAPH Course Notes*, ACM, 2006.
 16. T. Weise et al., "Realtime Performance Based Facial Animation," *ACM Trans. Graphics* (Proc. ACM SIGGRAPH 2011), vol. 30, no. 4, 2011, article no. 77.
 17. J. Tena, F. De la Torre, and I. Matthews, "Interactive Region-Based Linear 3D Face Models," *ACM Trans. Graphics* (Proc. ACM SIGGRAPH 2011), vol. 30, no. 4, 2011, article no. 76.
 18. E. Ong et al., "Robust Facial Feature Tracking Using Selected Multi-resolution Linear Predictors," *Proc. 12th Int'l Conf. Computer Vision*, IEEE CS, 2009, pp. 1483–1490.
 19. G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with Local Spatiotemporal Descriptors," *IEEE Trans. Multimedia*, vol. 11, no. 7, 2009, pp. 1254–1265.
 20. E. Ong and R. Bowden, "Learning Sequential Patterns for Lipreading," *Proc. British Machine Vision Conf. (BMVC)*, 2011, paper no. 55; www.bmva.org/bmvc/2011/proceedings/paper55/paper55.pdf.
- Daren Cosker** is a Royal Society Industry Fellow at the University of Bath and Double Negative Visual Effects. His research interests include applying computer vision and graphics to visual effects. Cosker has a PhD in computer science from Cardiff University, UK. He is a member of ACM SIGGRAPH. Contact him at dpc@cs.bath.ac.uk.
- Peter Eisert** is a professor of visual computing at the Humboldt University Berlin and head of the Computer Vision and Graphics Group at Fraunhofer HHI. His research interests include 3D image analysis and synthesis, face processing, image-based rendering, computer graphics, and 3D video processing. Eisert has a PhD in electrical engineering from the University of Erlangen, Germany. Contact him at peter.eisert@hhi.fraunhofer.de.
- Oliver Grau** is the associate director of operations of the Intel Visual Computing Institute, Germany. His research interests include innovative tools for visual media production and new user experiences, using computer vision and computer graphics techniques. Grau has a PhD in electrical engineering from the University of Hanover. Contact him at oliver.grau@intel.com.
- Peter J.B. Hancock** is a professor of psychology at the University of Stirling, UK. His research interests include the psychology of face perception, how our brains do it, and what sort of representations underlie our abilities. He conceived EvoFIT, a facial composite system currently used by police. Hancock has a PhD in computing science from the University of Stirling. He is a fellow of the British Psychological Society. Contact him at p.j.b.hancock@stir.ac.uk.
- Jonathan McKinnell** is a senior R&D engineer at the BBC R&D facility in London. His research interests include the research, development, and application of new and innovative ideas to produce television

programs. McKinnell has a PhD in Computational Mechanics from Imperial College London. Contact him at Jonathan.McKinnell@bbc.co.uk.

Eng-Jon Ong is a researcher at the Centre for Vision, Speech, and Signal Processing at Surrey University, UK, and is involved in an EPSRC project on automated lip-reading called LLiR. His research interests include computer vision (3D body tracking, hand

tracking, and facial features) as well as cognitive learning systems (COSPAL). Ong has a PhD in computer science from Queen Mary, University of London, UK. Contact him at e.ong@surrey.ac.uk.

 Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.