Applications of Face Analysis and Modelling in Media Production

Darren Cosker, Peter Eisert, Oliver Grau, Peter Hancock, Jonathan McKinnell, Eng-Jon Ong

Abstract

Facial expressions play an important role in day-by-day communication and more so in media. This article looks into automatic analysis and modelling of faces using computer vision techniques and their applications for media production.

Introduction

The visualisation, 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, for example in 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 on the one hand cognitive methods, i.e. face detection and recognition and 3D face modelling, and on the other hand applications of these methods in media production and access through digital services.

Broadcast and movie productions dedicate a lot 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' set-up 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 post-production. In the broadcast industry budgets are usually much smaller and post-production is kept to a minimum or even 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 that log key events, e.g. the approximate time when individual people appear in a shot, etc. 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 on 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 programme or their favourite actors or combination of both.

The first part of this article gives a brief overview of the psychology of face perception. The next sections describes some of the applications of computer vision and pattern recognition applied to face recognition in media production and the last part deals with automatic generation of face models. These are used in movie and TV-productions for 'special effects' in order to manipulate people's faces or to combine real actors with computer graphics[1].

[1] This article presents topics presented at a workshop on 'faces' at BBC R&D 2011, London, UK.

The psychology of face perception

This section outlines some of the key findings from psychology; we cannot cite all the original sources but fortunately there are three excellent recent surveys of the field [8], [9], [11]. Recognition of familiar faces is remarkably effortless, most of the time, though 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% correct, even with good quality images taken from the same viewpoint. Variations in viewpoint, expression and lighting cause additional problems. Multi-stranded films, such as 'Love Actually', rely on a cast of familiar actors to enable the audience to follow who is who.

There are two primary routes to recognising a face: the individual features and the relationship between them, or 'configuration'. It is possible to recognise 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 recognise an isolated top half of a face, than when it is aligned with an unrelated bottom half.

Colour is surprisingly unimportant for face recognition; it is possible to invert the colour palette of an image with little effect, although it can help where identification is uncertain. Unfamiliar face recognition (i.e., recognising someone you have seen before but do not know well) is strongly affected by 'external features' such as hair, though there is at least anecdotal evidence that this is more so 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, leading to a famous image of Al Gore behind Bill Clinton where in fact the internal features of both faces are identical.

There are clear 'other race' effects pervading face perception. It is as if there is another layer of unfamiliarity: unfamiliar faces are hard, 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, then at the bank, the latter will appear to drift upwards. This adaptation occurs with faces, too, so look at a male face for a few seconds and an ambiguous face will look female, study a female face and then the same ambiguous face looks male. The effects also apply to identification, so it is possible to affect our memory for 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 recognisability of caricatures. Cartoonists' drawings are often absurdly distorted and yet instantly recognisable. 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 these caricatures may be more recognisable than veridical images but it seems there is little, if any, effect for photographic images, except at very 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 of face recognition techniques in production of broadcasting programmes

Programme making is becoming increasingly demanding in terms of the amount of content, multiple platforms for delivery, shorter timescales and ever tighter budgets. New methods of delivery of programmes to the viewer are resulting in greater choice both in terms of genre of programme and delivery format. Multiple delivery formats are often required, which presents technical challenges as well as editorial ones. Automatic generation of metadata is crucial to allow cost effective production of broadcast programmes and enables new access possibilities, like specialised search functions for the viewer.

Metadata can be created and used from production and ingest right through to the editing or non-linear 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 upon the content within the scene (such as a timeline of all scenes containing a specific character) could facilitate greater speed whilst improving editorial storytelling. In order for a viewer to consume content in a non-linear fashion, metadata is crucial for creating a rich set of possible links and connections between different programmes. It is desirable that this metadata does not have to be (re-)entered at the end of the programme making workflow, which could be achieved by taking a selected set of the metadata logged or automatically tagged in the production (which has resource available for recording metadata) available on delivery.

In addition to attempting to increase the amount of metadata available in the edit and further downstream of the programme making lifecycle, accuracy of metadata is important. For example, it can be qualitatively observed that repetitive tasks (for example repetitive time based logging) can be inaccurately performed by humans, especially when distracted or over long periods of time. Also, the detail and variety of the logs can decrease over time due to other constraints on the logger or production assistant's productivity. Ideally the logger or production assistant (PA) would have their time freed up 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 non-standardised and non-uniform.

The number of people in shot and who is in shot can aid in the categorisation and searching of content and it is therefore the automation of these logs which 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 programmes at the BBC, the first a business discussion programme "The Bottom Line" which involved a live "face" training process and the second a drama "EastEnders". In a drama programme, the large variety of different characters to recognise, along with greater variations in pose, expression and lighting, make the recognition problem 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 due to their ease of implementation and real time performance. More state of the art recognition methods could be employed for higher recognition accuracy in future, where the initial aim was 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, where the

guests (typically three or four of them) are entrepreneurs and leaders from the business community. In general, it is not possible to train the recognition algorithm before filming the programme because the guests may not have had much previous media coverage and therefore a pre-existing 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 the logger / PA's 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 then 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 drama series such as BBC's "EastEnders" programme training need not be done on each production as 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 free run on set without live training.

Drama, however, is intrinsically a more difficult face recognition problem than interview programmes. Interview programmes are more formalised, leading to a greater similarity of pose and lighting often requiring recognition from a relatively small set of different people, whilst drama programmes have a greater variety of expression, pose and lighting. In drama, over time, people's faces can change due to ageing or other factors requiring a retraining of the face recognition algorithm, and drama can be indoor or outdoor, with a larger variety of shot types. In drama, the faces may be only small in comparison to the size of the overall image whilst in an interview programme this is rarely the case.

In summary, in a trial use of facial recognition tools on BBC programmes, qualitatively, the performance during interview programmes was found to be relatively good when compared to a human logger, whilst 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 pre-trained face sets for the presenter. Drama programmes were found to be a greater challenge, however, due to the difficulties outlined above. In future by making use of recent advances in recognition algorithms, comparable performance 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 video game companies both want to deliver new audience experiences – whether that is a hyper realistic human character (e.g. The Curious Case of Benjamin Button) or a fantasy creature driven by a human performer, like Avatar. 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 dense non-rigid tracking of surfaces, have fed - and are still feeding into - movies and video games. This Section gives an overview over 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 the Light-Stage [12] allows the capture of highly detailed facial normal maps. Coupled with sub-surface reflectance data [1] facial models now display photorealistic likenesses to real faces. Such technology is now widely used in modern

motion pictures, e.g. ‘Spider Man 2’ being one recent high-profile use of the Light-Stage. In these circumstances, the high detail static scans are composited onto either a stunt-actors body, or a digital double. This is where the actor is to be placed in situations that may not be practical or safe (such as explosions). However, it is still the case that the actor’s expression is static in these situations, and close examination of such shots reveal a dead-like facial quality. An early high-profile example of facial replacement was in the Matrix sequels [6], where passive facial scanning was used to obtain 3D faces with high detail facial texture. Fast sub-surface 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 video games is that faces must be renderable in a wide range of environments so that it 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 maps [12] or geometry [3], and rendering is enhanced through bidirectional reflectance distribution functions (BRDF). Acquisition of such BRDF detail has advanced widely over recent years resulting in highly detailed rendering [1].

Also very important to the realism of synthetic humans is the realistic rendering of hair, which has made a significant progress in the last years [16]. Similar to the BRDF modelling of skin, single hair fibers have been modelled 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 to model several fibres up to a complete hairstyle range from non-uniform rational B-Spline (NURBS) surfaces via thin shell volumes to strain-based modelling by parameterised clusters, fluid flow or vector and motion fields. In order to simulate the complex lighting interaction between strands of hair, Lokovich et al. propose a deep shadow map which relates visibility to depth for each pixel, yielding realistic but computationally expensive self-shadowing. Only recently, approximation algorithms for making the simulation of multiple scattering among hair fibres tractable have been proposed using methods like photon mapping or spherical harmonics.

Whereas laser scanning technology was initially the most accurate way to derive static facial detail, passive scanning technology using consumer hardware is now popular [3]. Multiple consumer level 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 is a different approach to those methods currently considered in for example movies, where a low polygonal mesh is used along with high detail normal maps to display facial meso-structure. There is therefore still a great deal of work to be done on practically using such technology for video games 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. While this strategy is beneficial for parallelization, it requires a post-processing stage to generate a mesh: The reconstruction yields a point cloud with outliers which 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: As each patch is unaware of its neighbours, the correction of a single mismatched patch by the user will not affect its neighbours, although they are likely to be erroneous as well. Therefore, [14] follows a similar approach as [3] but 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.

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

As previously mentioned, most digital face replacement in movies involves static face replacement, with the actor 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 since humans are experts in detecting the slightest flaws in faces. In the previous section we considered static faces. However, in order 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 play-back of recorded dynamic performances than the authoring of new animation. In order to highlight this we will first consider the acquisition of dynamic 3D facial sequences (termed here as 4D for brevity).

There are now many commercial companies that market 4D facial capture systems, i.e. those that can obtain 3D mesh data at video recording rate (e.g. Dimensional Imaging, 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 video rate along with high-resolution texture. The recorded sequences were 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 video games (e.g. Tiger Woods Golf, etc). Here, the system was made more robust by adding markers to the actors face. This could be used to stabilise and track a canonical mesh (i.e. mesh with a known topology) through the captured sequence. Bickel et al. [5] adopt a similar approach with the addition of extra facial paint to appropriately capture wrinkles on the face.

The use of markers has overcome previous issues related to tracking such a mesh using optical flow. Such methods are notorious to drift, caused primarily by fast facial changes such as during speech. First approaches to avoid the drift in markerless tracking over longer sequences are the incorporation of 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 Bradley et al. [7], who propose a multi-view stereo capture system comprising 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 stabilisation over time. Expanding further on this work, Beeler et al. [4] introduced the concept of anchor frames for stabilising 4D passive facial capture. In this work, neutral frames in the sequences are searched for and then used to essentially reinitialise mesh tracking where possible. This also has the added benefit of offering

robustness to certain facial self-occlusions (e.g. caused by the lips). Although having perhaps a lower geometric resolution than previous passive static capture work [4] –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 video game 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 by Alexander et al. [2]: The Digital Emily Project was a collaboration between Imagemetrics and USC using Light Stage technology to capture high detail normal map and surface reflectance properties from an actor’s face. 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 its control parameters. The degrees of freedom of the facial rig are dependent on the number and complexity of the blendshapes, and new facial poses are created by combining blendshapes with different weights.

While the LA Noire production is a high profile use of volumetric capture, it is essentially play back of the captured 4D video. On the other hand, the Digital Emily project 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). While we avoid a detailed review of methods in this area, the interested reader is encouraged to read the excellent course material of Havaldar et al. [10]. Our aim here is rather to make the distinction between direct playback of captured volumetric animation and the creation of realistic characters given some reference (e.g. actor performance).

The movie ‘The Curious Case of Benjamin Button’ is a good example of successful human realistic performance driven animation. MOVA (http://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 actors 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. While they can be used to animate a virtual target model, the result is not commonly used directly as an output on screen. Especially for productions of movies, but also for video games, 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.

While marker based motion capture techniques are widely popular e.g. 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 [10], the aim is to optimise a set of weights that approximate the positions of the markers. In marker-less 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 Kinect [17]. However, whether practical use of this technology ‘on-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. In the last part of this Section, we briefly consider some recent advances in blendshape rig construction that could help animators either approximate performance capture data more efficiently, or provide better artist tools.

One question we can ask is how to create effective blendshapes? A standard approach in modern VFX is to select these based on Action Units (AUs) from the Facial Action Coding System (FACS) (e.g. in 'Monsters House' and 'Watchmen'). 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 and then the weights are transferred onto a puppet model (perhaps of a creature). More recently, [17] considered creating blendshape rigs given only a few example expressions and a generic blendshape rig with a wider number of expressions. Such systems can potentially reduce artist time when manually sculpting 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 also 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. In order to address this, Tena et al. [15] recently proposed a region based PCA modelling approach that allows more intuitive direct manipulation of local facial regions. Their method also highlights how solving for expression weights locally can provide better approximation of motion capture data. Ultimately however, what an artist will desire of the facial model is a set of controls that are both intuitive and also orthogonal – such that altering one expression does that interfere too much with others. To counter this, blendshape rigs become highly complex, with additional shapes included to counter these interference cases. However, this is not entirely desirable for efficiency reasons, and future work is still required to address this core problem.

Automated Lip Reading

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

The first challenge involves the task of automatic facial feature tracking, a non-trivial problem, since the face is a highly deformable object. For example, the lip is highly deformable and can assume a large 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, in particular, points on the cheek where the texture can be homogeneous.
Accurate and robust facial feature tracking can be approached by means of a learnt, person-specific, data-driven approach using only pixel intensities [13]. A crucial component is the ability to automatically locate visual support that is optimal for tracking a particular point on the face (e.g.

mouth corner, eye corner, etc.). This allows us to potentially track any point on the face. Importantly, this includes points on regions where the visual complexity is high due to potential texture changes (e.g. inner lip) and facial features that are challenging due to the lack of texture (e.g. points on cheek). One commonly used method for tracking is the method of linear predictor flocks. Each Linear Predictor (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 flocks 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 find some 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 due to speech also tends to be less than that of emotions and other typical forms of actions recognised and variations are present across different individuals in terms of different mouth shapes, possible presence of facial hair and different styles of lip movements whilst speaking the same words. Using the above 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 [18]. Another method utilises 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 method of Boosting [19]. However, there remain open challenges to the task of lip reading, in terms of robustness to changing environmental conditions (e.g. lighting) and being able to deal with speech co-articulation, where the lip shape is affected by past, present and future words that is spoken.

Conclusions

Techniques for automatic processing of faces in images become mature due to progress in computer vision and image processing technology. Faces are anchor point in communication and media production. Using recognition techniques for faces and or lip reading is attractive as 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 modelling techniques are currently mainly used in high-profile movie and video games production, mainly because of the cost involved. With the progress in techniques able to operate on consumer level hardware, like DSLR cameras and depth sensors the budget threshold will be lowered and this will enable use of these techniques in other areas, like TV-production and on the end user side.

References

[1] C. Donner, T. Weyrich, E. d'Eon, R. Ramamoorthi, S. Rusinkiewicz , A Layered, Heterogeneous Reflectance Model for Acquiring and Rendering Human Skin, ACM Transactions on Graphics (Proc. SIGGRAPH Asia 2008), Vol. 27, No. 5, pp. 140:1–140:12, Singapore, 2008.

[2] O. Alexander, M. Rogers, W. Lambeth, J.-Y. Chiang, W.-C. Ma, C.-C. Wang, P. Debevec, The Digital Emily Project: Achieving a Photoreal Digital Actor, IEEE Computer Graphics and Applications July/August 2010.

[3] T. Beeler, B. Bickel, P. Beardsley, R. Sumner, M. Gross , High-Quality Single-Shot Capture of Facial Geometry, Proceedings of ACM SIGGRAPH, Los Angeles, USA, July 25-29, 2010.

[4] T. Beeler, F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, R. Sumner, M. Gross, High-Quality Passive Facial Performance Capture using Anchor Frames, ACM Transactions on Graphics (Proc. SIGGRAPH 2011), vol. 30, no. 3, August 2011.

[5] B. Bickel, M. Botsch, R. Angst, W. Matusik, M. Otaduy, H. Pfister and M. Gross, Multi-Scale Capture of Facial Geometry and Motion, ACM Transactions on Graphics, 26, 3, 2007.

[6] G. Borshukov, D. Piponi, O. Larsen, J. Lewis, C. Tempelaar-Lietz, Universal capture - image-based facial animation for "The Matrix Reloaded", SIGGRAPH Course Notes, 2005.

[7] D. Bradley, W. Heidrich, T. Popa, A. Sheffer, High Resolution Passive Facial Performance Capture, ACM Transactions on Graphics (Proceedings of SIGGRAPH 2010).

[8] V. Bruce and A. Young. "Face Perception", Psychology Press, 2011

[9] A. Calder, G. Rhodes, M. Johnstone and J. Haxby (Eds), "Oxford Handbook of Face Perception", Oxford University Press, 2011

[10] P. Havaldar, F. Pighin and J. P. Lewis, Performance Driven Facial Animation, SIGGRAPH Course Notes 2006.

[11] G. Hole and V. Bourne, "Face Processing: Psychological, Neuropsychological and Applied Perspectives", Oxford University Press, 2010

[12] W. Ma, T. Hawkins, P. Peers, C. Chabert, M. Weiss and Paul Debevec, Rapid Acquisition of Specular and Diffuse Normal Maps from Polarized Spherical Gradient Illumination, Eurographics Symposium on Rendering, 183-194, 2007.

[13] E. Ong, Y. Lan, B. Theobald, R. Harvey, R. Bowden. Robust facial feature tracking using selected multi-resolution linear predictors. In Proceedings of the 12th International Conference on Computer Vision, 2009.

[14] D. C. Schneider, M. Kettern, A. Hilsmann, P. Eisert, Deformable Image Alignment as a Source of Stereo Correspondences on Portraits, CVPR 2011 Workshop NORDIA, Colorado Springs, USA, June 2011.

[15] J. Tena, F. De la Torre, I. Matthews, Interactive Region-Based Linear 3D Face Models, ACM Transactions on Graphics, 30,4, 2011.

[16] K. Ward, F. Bertails, T.-Y. Kim, S. R. Marschner, M.-P. Cani, M- Lin, A survey on hair modelling: Styling, simulation and rendering, IEEE Transactions on Visualization and Computer Graphics TVCG, 13, 2, 2007.

[17] T. Weise, S. Bouaziz, H. Li and M. Pauly, Realtime Performance Based Facial Animation, ACM Transactions on Graphics, 30, 4, 2011.

[18] G. Zhao, M. Barnard, M. Pietikainen. Lipreading with local spatiotemporal descriptors. IEEE Transactions on Multimedia, 11 (7), 2009.

[19] E. Ong, R. Bowden. Learning Sequential Patterns for Lipreading, Proceedings of BMVC 2011.

Bios

Dr Daren Cosker is a Royal Society Industry Fellow at the University of Bath and Double Negative Visual Effects. His interests are in applying computer vision and graphics to visual effects. Between 2007 and 2012, Dr Cosker was a RAEng/EPSRC Research Fellow. He received a BSc and Ph.D degrees from Cardiff University in 2001 and 2006. He is a member of ACM SIGGRAPH.

Peter Eisert is Professor for Visual Computing at the Humboldt University Berlin and heading the Computer Vision & Graphics Group at Fraunhofer HHI. He has published more than 100 papers, is Associate Editor of IJIVP and in the Editorial Board of JVCIR. His research interests include 3D image analysis and synthesis, face processing, image-based rendering, computer / graphics, as well as 3D video processing.

Dr. Oliver Grau received his Diploma and PhD from the University of Hanover. In 2012 he joined as Associate Director of Operations of the Intel Visual Computing Institute. Previously, he worked for the BBC R&D department in the UK. His research interests are in innovative tools for visual media production and new user experiences, using computer vision and computer graphics techniques.

Peter Hancock, Professor of Psychology at University of Stirling. MA Chemistry, Oxford; MSc Intelligent Systems, Brunel; PhD Computing Science, Stirling. Fellow of the British Psychological Society. Research interests in the psychology of face perception, how our brains do it, what sort of representations underlie our abilities. Conceived EvoFIT, a facial composite system currently used by police.

Dr. Jonathan McKinnell is a Senior R&D Engineer at the BBC R&D facility in London. Jonathan is involved in the research, development and application of new and innovative ideas to produce television programmes. Previously, Jonathan worked at Filmlight Ltd in Soho, making use of computer vision algorithms to aid colour grading in the post production movie industry.

Eng-Jon Ong is a researcher at the Centre for Vision, Speech and Signal Processing at Surrey University, UK. He has been involved in research in computer vision (3D body tracking, hand tracking, facial features, etc..), as well as other interesting projects including cognitive learning systems (COSPAL). Currently he is involved in an EPSRC project on automated lip-reading called LILiR.

Contact information

Daren Cosker
Department of Computer Science
University of Bath
Bath, BA2 7AY
UK

Phone: 01225 385356
email: dpc@cs.bath.ac.uk

Peter Eisert, Prof. Dr.-Ing.

Einsteinufer 37,
10587 Berlin, Germany
Phone: +49-(0)30-31002-614
Fax: +49-(0)30-31002-190
URL: http://iphome.hhi.de/eisert

Oliver Grau
Intel Visual Computing Institute
Saarland University
Campus E2 1/1.13
66123 Saarbrücken
Germany

Peter Hancock
Professor,
Deputy Head of Psychology,
School of Natural Sciences
University of Stirling
FK9 4LA, UK

phone 01786 467675
fax 01786 467641
pjbh1@stir.ac.uk
http://www.psychology.stir.ac.uk/staff/staff-profiles/academic-staff/peter-hancock

---

Jonathan McKinnell
BBC R&D, Centre House,
56 Wood Lane, London,
W12 7SB
UK

Email: Jonathan.McKinnell@bbc.co.uk

Eng-Jon Ong
Centre for Vision, Speech and Signal Processing
School of Electronics and Physical Sciences
University of Surrey,

Guildford, GU2 7XH,
United Kingdom

e.ong@surrey.ac.uk