

Realtime Performance-Based Facial Avatars for Immersive Gameplay

Mark Pauly*

Computer Graphics and Geometry Laboratory
EPFL

Abstract

This paper discusses how realtime face tracking and performance-based animation can facilitate new ways of interaction for computer gaming and other online applications. We identify a number of requirements for realtime face capture and animation systems, and show how recent progress in tracking algorithms advances towards satisfying these requirements. Performance-based animation has the potential to significantly enrich the emotional experience of in-game communication through live avatars that are controlled directly by the facial expressions of a recorded user. We briefly outline other potential use scenarios of realtime face tracking systems and conclude with a discussion of future research challenges.

CR Categories: I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction techniques; I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Animation;

Keywords: performance-based animation, avatars, realtime face tracking, immersive gameplay

Links: [DL](#) [PDF](#)

1 Introduction

A number of studies have shown that role play is crucial for children in the development of their cognitive skills [Bergen 2002]. Character enactment is an essential means of communication and story-telling, for example in theatrical performances and movies. Acting and role play are often facilitated by some form of physical face masks that also play an important role in cultural or religious rituals throughout the world [Pernet 1992]. Masks are often employed to disguise the bearers identity, for example in the Venice carnival, where the disguise enables a partial suspension of social norms [Johnson 2011].

The above examples suggest an innate human desire to impersonate somebody else, or to alter or augment one's own personality or appearance. This powerful idea is also fundamental in virtual worlds, and in particular in computer gaming, where role play and impersonation of game characters are important concepts. Current games, however, often restrict the direction of avatars to spatial actions, such as walking, jumping, or grasping, but provide little control over facial expressions. Yet facial expressions, when directly controlled by the human player, offer the potential for much richer



Figure 1: A digital avatar controlled by the user's own facial expression using realtime face tracking and retargeting offers exciting opportunities for interactive gameplay.

interactions that capture the emotional discourse, and might even enable entirely new forms of performance-driven gameplay.

This paper discusses several research projects at the EPFL Computer Graphics and Geometry Laboratory (lgg.epfl.ch) as a complement to an invited talk by the author on performance-based facial animation at the 2013 ACM SIGGRAPH conference on *Motion in Games*. For conciseness of exposition, we do not provide a discussion of related work here, but refer instead to the cited papers for a review of the existing literature.

In Section 2 we specify a number of requirements that need to be met to successfully implement a realtime face tracking system for interactive games. Section 3 summarizes and compares three prototype solutions that we have developed over the past five years, highlighting some of the essential concepts and algorithms necessary to meet the identified requirements. We hope that this analysis provides guidance for the design of future performance-based animation systems suitable for large-scale deployment in computer gaming and other online applications. Section 4 highlights a number of potential use-cases of the presented research, including non-gaming applications. Section 5 outlines several research challenges for future work.

2 System Requirements

Integrating face tracking technology into games necessitates that a number of requirements are satisfied to achieve convincing gameplay and support wide-spread adoption by gamers. We identify the following core features that affect the design of suitable algorithms for performance-based facial animation:

- **Accuracy.** The reconstructed facial expressions and rigid head pose need to be sufficiently accurate to match the facial dynamics of the user. In particular for speech, facial expressions need to closely follow the typical lip motions associated with phonemes in order to avoid lip syncing artifacts that often degrade the perception of the animated character.

*e-mail: mark.pauly@epfl.ch

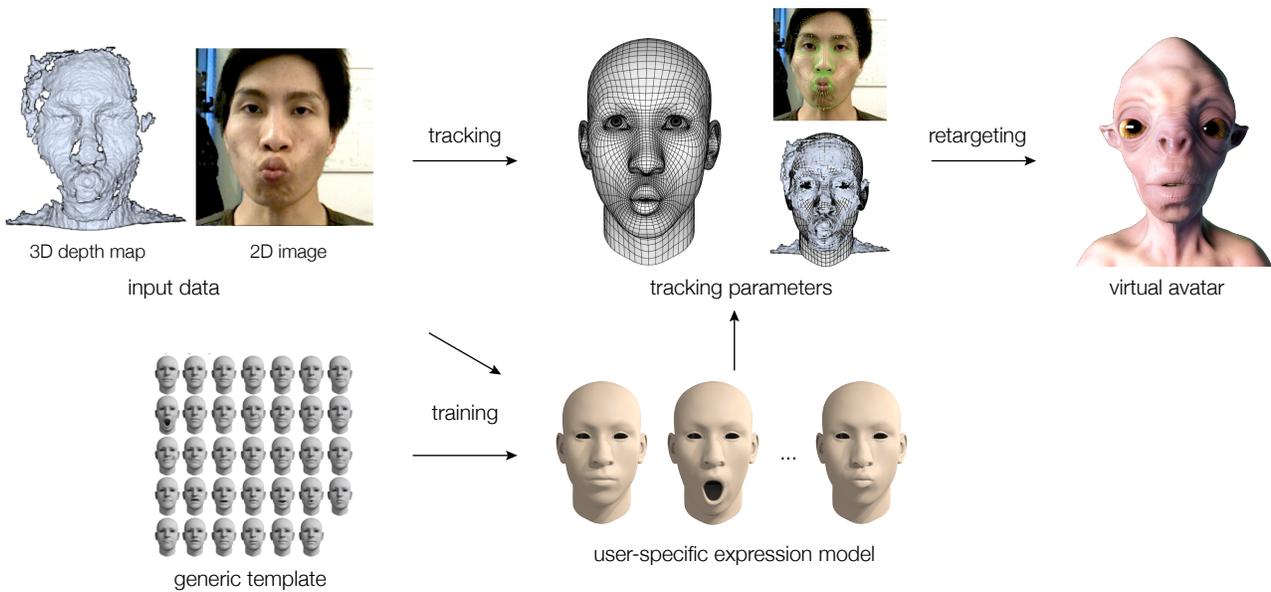


Figure 2: Performance-based facial animation pipeline. Input data from the sensor is processed to determine suitable tracking parameters, such as rigid head pose and expression coefficients. These parameters are then used to retarget the performance onto a virtual avatar that mimics the facial expressions of the recorded user.

- **Performance.** Engaging interactions mandate a sufficient framerate and low latency. This requires fast sensing and puts strong constraints on the computational efficiency of the tracking and retargeting algorithms. In particular for a gaming environment where substantial computation resources are needed for other components of the game, time and memory efficient algorithms are of paramount importance.
- **Robustness.** To maintain an immersive game-play, the face tracking and retargeting algorithms need to be robust and yield plausible results even for difficult tracking configurations. These can be caused by bad lighting, fast motions, or occlusions, for example.
- **Usability.** To facilitate wide-spread acceptance among users, both tracking hard- and software need to be easy to use. This requires the sensing to be non-invasive; face markers or visible active illumination that simplify tracking are not suitable. Furthermore, the system should as much as possible be *calibration-free*. User-specific training and system calibration need to be kept to a minimum.

The listed requirements are often in conflict. For example, performance considerations might necessitate algorithmic simplification that reduce accuracy. High usability requires non-intrusive sensing devices which potentially degrades robustness. For interactive applications robustness is often favored over accuracy, i.e. plausible expressions are more important than geometrically accurate ones. These tradeoffs define a complex design space for realtime performance-based animation systems.

3 Performance-based facial animation

The above system constraints have been drivers in a series of research projects in realtime performance-based facial animation that we have conducted over the past few years [Weise et al. 2009; Weise et al. 2011; Bouaziz et al. 2013a]. This section provides an overview of the common elements of these systems and outlines their differences. The goal of this analysis is to identify fundamen-

tal concepts and algorithms that are essential to successfully deploy performance-based animation in computer games.

Figure 2 shows a high-level view of the performance-based facial animation pipeline. Common to all our system is input sensing data in the form of a color image and 3D depth map of the user’s face for each frame of the recorded sequence. Our first system [Weise et al. 2009] made use of a custom-built structured light scanner, while subsequent systems [Weise et al. 2011; Bouaziz et al. 2013a] employed commodity sensors, such as the Microsoft Kinect or the Primesense Carmine (Figure 3). All of these scanners provide real-time sensing at 25-30 frames per second with a latency in the order of 100ms. Based on the per-frame input, facial tracking parameters are estimated and fed into a retargeting stage that drives the animation of a selected virtual avatar. Face tracking is facilitated by a user-specific dynamic expression model (DEM) that defines a parameterization of the expression space of the recorded user. The DEM is built in a separate training stage that is either performed offline or directly integrated into the tracking optimization.

A core hypothesis in our research is that 3D depth information of the observed face can significantly improve the tracking quality. We build on this hypothesis by formulating tracking as a geometric registration problem that aligns the user-specific DEM with the observed geometry and image data. This registration is facilitated by representing the DEM in a reduced basis that is constructed in a training stage. The following paragraphs provide more detail on the individual stages of the pipeline and explain the commonalities and differences between the three presented systems.

Dynamic expression model. We represent a specific facial expression as a 3D surface mesh with fixed connectivity (see Figure 2). Tracking then amounts to computing the 3D position of each mesh vertex to best match the observed depth map. However, treating each vertex position as an optimization variable is not suitable for a realtime system. In the context of facial tracking we can make use of the specific structure of the tracked geometry to improve efficiency and robustness. Human facial expressions are highly coupled spatially and temporally on the vertex level. This



Figure 3: *The structured light scanner used for the realtime performance-based animation system of [Weise et al. 2009] (left) yields superior data quality compared to the commodity sensors employed in [Weise et al. 2011; Bouaziz et al. 2013a] (right). However, this custom-built system is expensive and uses intrusive light projections, making it unsuitable for consumer applications.*

means that vertices do not move independently as the recorded face moves or changes expression. This strong correlation can be exploited to build parameterizations of the dynamic facial expression space based on dimensionality reduction.

Our first system presented in [Weise et al. 2009] uses principal component analysis (PCA) to build a basis representation, while our latter systems [Weise et al. 2011] and [Bouaziz et al. 2013a] use a blendshape basis to model the facial expression space. With both approaches, a specific facial expression is represented as a weighted sum of a neutral pose (PCA mean resp. neutral expression) and a set of basis vectors (principal components resp. blendshape expressions). These linear representations allow designing efficient solvers for the tracking optimization as detailed in the respective papers.

A main advantage of PCA is that the basis vectors are orthogonal, which implies that each facial expression is represented by exactly one corresponding coefficient vector. Blendshape models are typically not orthogonal and often redundant in the sense that the same expression can be represented by different coefficient vectors. This indeterminacy can complicate the tracking, as a smooth transition in expressions can lead to discontinuous coefficient vectors, which might impair temporal smoothing and complicate retargeting. On the other hand, a blendshape basis can be freely designed, which offers a number of advantages. While PCA is essentially "blind" to high-level semantics of the input data, blendshapes allow incorporating external knowledge about face geometry and dynamics. Building a PCA model requires sufficient expression coverage during training, while the generic blendshape basis already defines a full set of facial expressions, which greatly reduces the number of required training poses (see paragraph on training below). Overall, we found that the benefits of blendshape models outweigh the advantages of PCA for realtime face tracking.

Registration. We view 3D tracking as a template-based non-rigid geometry registration problem. General methods, such as [Pauly et al. 2005; Li et al. 2008; Li et al. 2009; Li et al. 2012] directly optimize for the vertex coordinates of the 3D mesh or employ some generic deformation model to align the template with the 3D scans. In our context, the DEM provides a parameterization of the dynamic expression space of the user that uses significantly fewer unknowns (typically in the order of 40 to 60) than the number of mesh coordinates. Formulating the registration as an energy minimization problem over the DEM coefficients allows solving the resulting non-linear optimization in realtime.

Our systems are initialized using a face detector to obtain a rough positional estimate [Viola and Jones 2001]. During tracking, we re-

duce the number of necessary iterations by initializing the optimization with the coefficients of the previous frame, thus exploiting the temporal coherence of the recorded facial performance. A geometric alignment of the DEM with the recorded depth map is achieved using point correspondences. In our systems, we replace the commonly used closest point queries for establishing correspondences with a simple orthogonal projection lookup, which only requires constant time computation and does not affect the registration accuracy significantly. We found that splitting tracking into a separate rigid motion estimation and a subsequent computation of the expression parameters simplifies the optimization and leads to more robust results.

For conciseness we omit the specific registration energies here and refer to [Bouaziz and Pauly 2013] for a detailed description of geometry and image registration methods. This tutorial also describes how texture correspondences can be integrated into the registration, which provides additional cues for tracking. A promising approach to improve the robustness of the rigid pose estimation has recently been proposed in [Bouaziz et al. 2013b]. This method employs sparsity inducing norms to automatically reject bad correspondences, but needs to be further optimized to be suitable for realtime processing.

Training. The PCA model of [Weise et al. 2009] is built by scanning a large number of facial expressions for each user in an offline pre-process. First a neutral expression is scanned by having the user rotate her head in front of the scanner while keeping the neutral pose fixed. Using incremental aggregation, all recorded depth maps are fused into a single 3D scan. This scan is subsequently registered with a generic face template to obtain a user-specific neutral pose with the same mesh structure as the template. Additional poses are created by warping this template to the recorded expression scans using non-rigid registration based on a membrane model. The membrane model at the time was chosen for simplicity. Since performance is not a crucial factor during the offline processing, we suggest using more sophisticated deformation models. We experimentally compared several such models [Botsch et al. 2006a; Sumner et al. 2007; Botsch et al. 2007; Eigensatz and Pauly 2009; Bouaziz et al. 2012] and currently recommend the projection-based optimization of [Bouaziz et al. 2012] that offers a good tradeoff between deformation quality and algorithmic complexity.

One difficulty of the PCA approach is to determine when sufficiently many expressions have been reconstructed to adequately cover the facial expression space of the user. This problem can be largely avoided when using a blendshape model that provides a complete generic template basis [Weise et al. 2011]. The training process is then greatly simplified by reducing the number of required training poses. However, it is not practical to ask the user to directly match the blendshape expressions during training, because these might be difficult to assume. Instead, the system requires the user to perform a small set of pre-defined expressions such as smile, mouth open, squint, etc., that are then mapped to the blendshape basis using example-based facial rigging [Li et al. 2010]. This method performs a gradient space optimization (see also [Botsch et al. 2006b]) to adapt the generic blendshape basis to best match the recorded expressions.

While training the blendshape model is relatively easy, it still requires about 15-20 minutes, which is unsuitable for many interactive or gaming scenarios. The main innovation of [Bouaziz et al. 2013a] is to completely dispense with the offline preprocess and directly incorporate the training into the tracking optimization. This means that the DEM is adapted on the fly to the specific facial features of the user. Since *no* prior training is necessary, this system fully satisfies the usability requirement.



Figure 4: *Non-gaming interactive applications. From left to right: Facial expression coefficients control a music synthesizer; a digital clone of the host of a TV show; augmented reality with synthesized eye glasses.*

Animation prior. A fundamental issue when processing acquired data is sensing noise. In particular for commodity sensors based on infrared projection, data quality can be fairly low (see Figure 3). In order to achieve robust tracking, we need to provide appropriate priors that avoid over-fitting to noise and outliers in the recorded depth map. The reduced model of the DEM already provides a strong regularization by coupling vertices in the linear model. However, not every set of coefficients yields a realistic or plausible facial expression, hence additional regularization is required when dealing with noisy data. Explicit bounds on the values of coefficients provide a first means of delineating the expression space, but typically cannot capture its complex structure. A more sophisticated solution defines the expression space based on examples [Weise et al. 2011]. Tracking is formulated as a maximum a posteriori optimization that solves for the most probable tracking coefficients based on the observed data and a statistical prior that is derived from example animation sequences. In our experiments we found this approach to yield good results, but more research is needed to fully explore these statistical priors. In particular, the dependency of the results on the extent and quality of the training data has not been studied systematically yet. Obtaining high-quality training data that sufficiently covers the dynamic expression space is far from trivial, which can limit the practical applicability of this approach.

Summary. All three systems can robustly track the user and animate a virtual avatar in realtime. In our current implementations, however, the computational overhead is still fairly substantial, requiring additional research and engineering efforts to achieve performance levels acceptable for a gaming application.

Our first system [Weise et al. 2009] used active structured light sensing that is not suitable for consumer applications. The later systems employ RGB-D sensors specifically developed for computer gaming and do not impose any relevant restrictions in terms of hardware usability. Software usability is significantly improved by the online learning algorithm of [Bouaziz et al. 2013a] that requires no prior training by the user, thus rendering the system fully operational from the beginning of the interaction. This improvement comes at the cost of a slight reduction in tracking accuracy and increased computational cost, however.

We found that geometry and image registration offer important complementary benefits. Image information can significantly improve tracking in regions with sufficient texture detail, such as the eyes or the mouth. The 3D depth map helps with the rigid pose, but also with subtle motions like moving the chin forward and backward, which cannot be handled well with a single video camera.

All systems have proven fairly robust and can handle common prob-

lems like minor occlusions or bad lighting. Since our later systems use active infrared sensing, tracking even works in complete darkness. However, texture-based components, such as eye tracking, degrade with reduced light intensity.

4 Applications

We briefly outline some potential applications made possible by realtime face tracking and animation systems. With foreseeable improvements of both hard- and software, we expect to see many more applications than the small set of examples discussed here. Yet even this limited set provides an indication of the profound impact that this technology can potentially have in computer gaming and other domains (see also Figure 4).

Controlling the facial expressions of avatars in realtime allows engaging interactions in games and other online applications. For example, in-game communication can benefit from "face-to-face" encounters among game characters controlled by different players. This kind of emotional avatars with the full expressivity of facial features can even be at the center of new forms of game-play. Beyond gaming, other forms of online communication can greatly benefit from the technology outlined in this paper. Virtual 3D communication spaces can be populated with avatars that either closely match the true appearance of the tracked person, e.g. in a professional video conference call, or are selected by the user to his or her liking, e.g. in a chat room. For such applications, the physical face masks mentioned in the introduction are replaced by virtual masks in the form of digital avatars.

Avatar-based live interaction can also become an interesting communication means in television or other broadcast media. Customer support applications can enable richer interactions compared to mere voice assistance and create a unified corporate interface by using a single avatar driven by all customer support agents. Realtime performance-driven facial animation also show promise in medical applications, where new forms of interactive therapy might become possible. For example, avatar-based training sessions can be designed for people with autism or other disorders of neural development. Finally, we believe that fascinating new art installations and interactive exhibits can be built upon the discussed technology as illustrated in Figure 5.

5 Challenges

Most of the above mentioned applications are not a reality yet. While the research progress outlined in this paper brings performance-based facial animation within reach of consumer level



Figure 5: Different versions of the Mimicry installation. When an observer steps in front of the picture frame, the character depicted in the virtual painting starts mimicking the person’s facial expression in realtime. The tracking sensor is embedded in the frame.

applications, a number of fundamental challenges remain. Beyond necessary improvements in computational performance, we see a need for further research in the following areas:

Retargeting. The discussed systems provide very basic retargeting that requires the animated avatar to essentially match the semantics of the tracking model. In [Weise et al. 2009], a retargeting method based on deformation transfer is used, while [Weise et al. 2011; Bouaziz et al. 2013a] support retargeting by directly transferring blendshape coefficients. Both methods are rather restrictive in that they require the animated avatar to closely match the tracking model semantics. Establishing a mapping between tracking coefficients and animation controls of a given avatar can be difficult, because expressions might not be compatible when the geometry or motion space of the target avatar differs strongly from that of the tracked user. More general retargeting methods that can bridge this semantic gap are needed.

Avatar creation. Creating compelling virtual avatars is currently a highly complex process that requires expert skill and significant manual work. While 3D face scanning can be considered a mature technology, creating a fully rigged dynamic face model is difficult to achieve with scanning alone. Avatars might also be desired that create a blend of the scanned person with an existing digital character. For example, creating a personalized fantasy creature with recognizable facial features of the user poses challenging research questions in dynamic morphing and texture blending.

Realism. The well-known uncanny valley effect poses a significant challenge when aiming for realistic digital doubles of the user or other realistic human avatars. Subtle facial dynamics and micro-expressions are very difficult to track reliably and animate correctly. Simply adding more PCA coefficients or blendshapes to the DEM is not a viable solution, because of an increased sensitivity to acquisition noise.

Secondary effects. The extracted expressions of the observed user are only one aspect of a facial performance. The systems described in this paper do not consider tracking of hair, teeth, tongue, nor do they simulated the corresponding dynamic geometry on the synthesized avatar. Non-trivial hairstyles are still very challenging to model, simulate, and render. No robust solutions currently exist to realistically track and synthesize long hair in realtime.

6 Conclusion

The current state-of-the-art in realtime performance-based face animation shows that compelling virtual interactions are possible in an online context. This technology offers numerous opportunities for in-game communication and avatar control, but also enables new forms of human communication that yet need to be explored. A number of open problems remain that offer numerous challenges for future research.

Acknowledgments. Thanks to Brian Amberg and Thibaut Weise of faceshift AG and to Emilie Tappolet and Raphaël Muñoz and the HEAD of Geneva for the design of the Mimicry installation. Thanks also to Ivo Diependaal for creating the Alien model. The main projects described in this paper have been done in collaboration with Sofien Bouaziz, Hao Li, Yangang Wang, and Thibaut Weise. This research is supported by the Swiss National Science Foundation grant 20PA21L_129607 by the Swiss Commission for Technology and Innovation.

References

- BERGEN, D. 2002. The role of pretend play in children’s cognitive development. *ECRP* 4, 1.
- BOTSCH, M., PAULY, M., GROSS, M., AND KOBELT, L. 2006. Primo: coupled prisms for intuitive surface modeling. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, SGP ’06, 11–20.
- BOTSCH, M., SUMNER, R., PAULY, M., AND GROSS, M. 2006. Deformation transfer for detail-preserving surface editing. In *Vision, Modeling, Visualization 2006*, 357–364.
- BOTSCH, M., PAULY, M., WICKE, M., AND GROSS, M. H. 2007. Adaptive space deformations based on rigid cells. *Comput. Graph. Forum* 26, 3, 339–347.
- BOUAZIZ, S., AND PAULY, M. 2013. Dynamic 2d/3d registration for the kinect. In *ACM SIGGRAPH 2013 Courses*, ACM, New York, NY, USA, SIGGRAPH ’13, 21:1–21:14.
- BOUAZIZ, S., DEUSS, M., SCHWARTZBURG, Y., WEISE, T., AND PAULY, M. 2012. Shape-up: Shaping discrete geometry with projections. *Comp. Graph. Forum* 31, 5 (Aug.), 1657–1667.
- BOUAZIZ, S., WANG, Y., AND PAULY, M. 2013. Online modeling for realtime facial animation. *ACM Trans. Graph.* 32.

- BOUAZIZ, S., TAGLIASACCHI, A., AND PAULY, M. 2013. Sparse iterative closest point. *Computer Graphics Forum (Symposium on Geometry Processing)* 32, 5, 1–11.
- EIGENSATZ, M., AND PAULY, M. 2009. Positional, metric, and curvature control for constraint-based surface deformation. *Comput. Graph. Forum* 28, 2, 551–558.
- JOHNSON, J. H. 2011. *Venice Incognito: Masks in the Serene Republic*. University of California Press.
- LI, H., SUMNER, R. W., AND PAULY, M. 2008. Global correspondence optimization for non-rigid registration of depth scans. *SGP* 27.
- LI, H., ADAMS, B., GUIBAS, L. J., AND PAULY, M. 2009. Robust single-view geometry and motion reconstruction. *ACM Trans. Graph.* 28, 175:1–175:10.
- LI, H., WEISE, T., AND PAULY, M. 2010. Example-based facial rigging. *ACM Trans. Graph.* 29, 32:1–32:6.
- LI, H., LUO, L., VLASIC, D., PEERS, P., POPOVIĆ, J., PAULY, M., AND RUSINKIEWICZ, S. 2012. Temporally coherent completion of dynamic shapes. *ACM Transactions on Graphics* 31, 1 (Jan.).
- PAULY, M., MITRA, N. J., GIESEN, J., GROSS, M., AND GUIBAS, L. J. 2005. Example-based 3d scan completion. In *SGP*.
- PERNET, H. 1992. *Ritual Masks: Deceptions and revelations*. University of South Carolina Press.
- SUMNER, R. W., SCHMID, J., AND PAULY, M. 2007. Embedded deformation for shape manipulation. *ACM Trans. Graph.* 26, 80.
- VIOLA, P., AND JONES, M. 2001. Rapid object detection using a boosted cascade of simple features. In *CVPR*.
- WEISE, T., LI, H., VAN GOOL, L., AND PAULY, M. 2009. Face/off: Live facial puppetry. In *Proc. Symposium on Computer Animation*, ACM, 7–16.
- WEISE, T., BOUAZIZ, S., LI, H., AND PAULY, M. 2011. Real-time performance-based facial animation. *ACM Trans. Graph.* 30, 77:1–77:10.