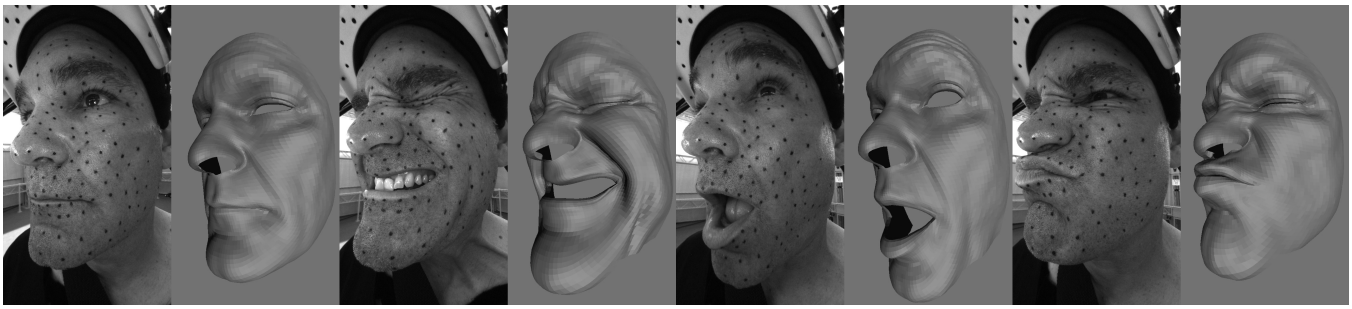
Masquerade: Fine-scale details for head-mounted camera motion capture data

Lucio Moser
Digital Domain
lmoser@d2.com

Darren Hendler
Digital Domain
darren@d2.com

Doug Roble
Digital Domain
doug@d2.com



ABSTRACT

We present Masquerade, a novel modular and expandable tool for adding fine-scale details to facial motion capture data from head-mounted cameras. After studying two important related works we developed a framework to reproduce the original approaches as well as to test equally promising alternatives. This framework has been vital for understanding the limitations of previous approaches and to explore ways to improve the results. Our final solution was a combination of algorithms and data representations that produced better results than previous works when tested with our evaluation data. Since then, Masquerade is being actively used in production for enhancing marker data with fine-scale details.

CCS CONCEPTS

•Computing methodologies →Motion capture;

KEYWORDS

Motion capture, head-mounted cameras, data-driven upsampling.

ACM Reference format:
Lucio Moser, Darren Hendler, and Doug Roble. 2017. Masquerade: Fine-scale details for head-mounted camera motion capture data. In *Proceedings of SIGGRAPH '17 Talks, Los Angeles, CA, USA, July 30 - August 03, 2017,* 2 pages.
DOI: http://dx.doi.org/10.1145/3084363.3085086

1 OVERVIEW

Two main techniques are used for facial motion capture in visual effects: (1) high-resolution capture using a fixed camera rig around the seated actor; and (2) low-resolution capture using head-mounted cameras (HMC). In general, HMC capture is the preferred method, since the actors can move freely on set. Perhaps the biggest limitation of HMC capture is that it will always lag behind in terms of quality of reconstruction compared to the seated captures, which have higher resolution, controlled lighting, and more cameras for wider coverage. Previous works have unified these two approaches, using data-driven techniques to create high-resolution results from low-resolution captures. Masquerade builds on the previous work, producing results with lower reconstruction errors and using less memory during the computations.

2 RELATED WORK

There are several data-driven approaches to HMC motion capture. The method in [Wu et al. 2016] uses markers as well as optical flow for improved tracking. It requires anatomical constraints to be added to the model and uses a global solver to fit both large and fine deformations at once. The method in [McDonagh et al. 2016] is based on markerless tracking and it requires building synthetic training data from photo-real, actor-specific renders in a range of poses and lighting situations that are used in a regression model. Both works seem to produce very good results but their added requirements complicate the workflow and implementation.

In [Bickel et al. 2008] the large-scale deformations and fine-scale deformations are separated and processed independently. The former is done by warping the face based on the tracked markers and the latter is solved by interpolating fine-scale details from a database built from the high-resolution capture and facial pose descriptions.

The approach in [Bermano et al. 2014] has the same main components as [Bickel et al. 2008] but is designed for supporting different types of motion capture data, working with the same high-resolution database of fine-scale details. This is an interesting advantage and it can be used when the marker positions change between HMC recording sessions (markers are reapplied every day and may shift locations).

SIGGRAPH '17 Talks, Los Angeles, CA, USA
 978-1-4503-5008-2/17/07…$15.00
DOI: http://dx.doi.org/10.1145/3084363.3085086

3 IMPLEMENTATION

Our work started by examining the approaches in [Bermano et al. 2014] and [Bickel et al. 2008], abstracting the main building blocks that they share, and implementing their algorithms to verify the original approaches. To measure the error of reconstruction in a perfect data situation, we acquired high-resolution performance captures. We generated artificial low-resolution captures from the high-resolution data by attaching virtual markers to the high-resolution meshes. A performance that explored a wide range of expressions was used as training data and the emotional dialogue performances were used as evaluation data.

Using the low-resolution data, the reconstruction error for the training and evaluation data was measured and the visual quality of the results was evaluated. The goal was to find the highest quality of reconstruction with the lowest memory requirements. Interestingly, we identified benefits and shortcomings of both original works, which guided us to our final solution.

The approach taken by [Bickel et al. 2008] resulted in low reconstruction errors in our data sets. We noticed artifacts on areas that were not directly covered by the markers, such as the lips. This may be because the edge-strain representation used for the pose ignores the surface bending. Therefore, different lip configurations are represented similarly and result in blending unrelated training samples. Further, the method is memory intensive as each vertex has its own radial basis function (RBF) weight matrix, potentially limiting either the resolution of the geometry or the number of examples used for training.

We tested the approach in [Bermano et al. 2014] with our motion capture data and found it was significantly slower due to the use of Quadratic Programming solvers. We also obtained larger reconstruction errors than with [Bickel et al. 2008]. This may be due to the face being segmented into four large regions and each one using a large multi-dimensional value for representing its pose. High-dimensional data analysis can suffer from the "Curse of Dimensionality" which results in similar distances to several training samples, leading to blending too many samples. In addition, because the large and fine scale deformations are solved independently, there is no guarantee that the combined deformation gradients will recreate the scale of the actor's face. A slight error in the scale components of deformation gradients can result in global changes in the scale of the face.

Our final implementation resulted in a combination of the two approaches, as follows.

- To get around the artifacts in the lips seen in [Bickel et al. 2008] experiments, we utilized deformation gradients (like [Bermano et al. 2014]) to represent the pose of the face as opposed to edge strains. Deformation gradients represent the deformation relative to the rest pose, including the bending that the surface undergoes.
- Following the strategy in [Bermano et al. 2014], the system can reuse training data for different marker sets because the geometry used as the pose representation is generated from the large-scale deformations.
- Similar to [Bickel et al. 2008], we use local vertex offsets to encode the fine-scale details. It avoids the scaling issues from deformation gradients and is an improvement over world-space vertex offsets, which don't adapt to the large-scale deformations.
- Due to the long computation time required for Quadratic Programming solvers, we opted for using RBF interpolation with a biharmonic kernel, as in [Bickel et al. 2008]. But instead of building one solver per high-resolution mesh vertex, we found that better quality of reconstructions are achieved by segmenting the face into small regions, one per marker, and applying one RBF per region. That dramatically reduced the number of solvers and made it independent on the output resolution. There is some redundancy added since the regions have soft boundaries and their results are blended using geodesic weights for each vertex, but it still is significantly less than the memory cost of the per-vertex solvers. We believe this strategy works because the blending weights of the fine-scale details vary smoothly in the high-resolution face and we took advantage of that, reducing the number of computations and amount of memory required.
- Additionally, we adopted a similar greedy training strategy as used in [Bermano et al. 2014] and [Bickel et al. 2008]. For most of the tests we concluded that using 80 frames for training (out of 320 training frames) was sufficient for high reconstruction quality. Since we are using one RBF per marker, we can further reduce the memory required by training the solver of each region with their own set of best frames. We found that using the 40 best frames for each region achieves similar reconstruction errors. It can be further reduced to 20 and still produce visually comparable results while reducing the memory requirements by a factor of four.

Our optimal solution has been successfully applied to real production data. Consistently, it has produced lower reconstruction error with less computation and requires less memory than [Bickel et al. 2008].

REFERENCES

Amit H. Bermano, Derek Bradley, Thabo Beeler, Fabio Zund, Derek Nowrouzezahrai, Ilya Baran, Olga Sorkine-Hornung, Hanspeter Pfister, Robert W. Sumner, Bernd Bickel, and Markus Gross. 2014. Facial Performance Enhancement Using Dynamic Shape Space Analysis. *ACM Trans. Graph.* 33, 2, Article 13 (April 2014), 12 pages. DOI: http://dx.doi.org/10.1145/2546276

Bernd Bickel, Manuel Lang, Mario Botsch, Miguel A. Otaduy, and Markus Gross. 2008. Pose-space Animation and Transfer of Facial Details. In *Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '08)*. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 57–66. http://dl.acm.org/citation.cfm?id=1632592.1632602

Steven McDonagh, Martin Klaudiny, Derek Bradley, Thabo Beeler, Iain Matthews, Kenny Mitchell, undefined, undefined, undefined, and undefined. 2016. Synthetic Prior Design for Real-Time Face Tracking. *2016 Fourth International Conference on 3D Vision (3DV)* 00 (2016), 639–648. DOI: http://dx.doi.org/doi.ieeecomputersociety.org/10.1109/3DV.2016.72

Chenglei Wu, Derek Bradley, Markus Gross, and Thabo Beeler. 2016. An Anatomically-constrained Local Deformation Model for Monocular Face Capture. *ACM Trans. Graph.* 35, 4, Article 115 (July 2016), 12 pages. DOI: http://dx.doi.org/10.1145/2897824.2925882