A Prior-based Approach to 3D Face Reconstruction using Depth Images

Donny Tytgat, Sammy Lievens, and Erwin Six

Alcatel-Lucent Bell Labs,
Copernicuslaan 50, 2018 Antwerp, Belgium
{donny.tytgat,sammy.lievens,erwin.six}@alcatel-lucent.com

Abstract. We present a system for 3D face reconstruction that is based on live depth information combined with face priors. The system includes a stereo matching method that employs prior information for limiting the search space and introduces an offline scanned 3D model that is animated by means of 2D morphing techniques in order to match the live facial expressions. The resulting live- and synthetic depth images are combined and a live 3D mesh is generated. Computational aspects are taken into account for every part of the system, enabling a live face reconstruction system with the potential for real-time execution.

Keywords: Depth image, 3D, Face reconstruction, Real-time

1 Introduction

A live 3D face reconstruction system can offer new video communication solutions by decoupling the capturing device from the visualized data. Virtual cameras can be placed anywhere around the face, enabling a more compact and clear depiction of the conversation dynamics or establishing better gaze alignment. In addition, the reconstructed model can be visualized on 3D displays.

A system is presented that employs prior information, both generic and personalized, in order to create a 3D face reconstruction system that could be applied in a real-time video communication system. The system uses depth images as an intermediary format for data aggregation. A prior-aided stereo matching method is shown that uses a sparse set of correspondences in order to guide the dense stereo matching process. This is followed by a method that employs 2D morphing techniques for generating an approximation of an animated, personalized 3D model that is represented by multiple depth images. A surface reconstruction method is furthermore presented that produces a triangulated model from an arbitrary number of depth images. These three components are combined into a system that can produce a live 3D face mesh.

2 Related Work

3D face reconstruction is an active field in the domains of computer vision and computer graphics. Reconstruction from a single image is in its most generic form

an ill-posed problem. As such prior information is required in order to solve it. A survey of this field is discussed in [8]. Prior information such as light sources (shape from shading) and geometric/albedo models (analysis by synthesis) are often used. Results are however rarely satisfactory for realistic rendering due to the limited live information which poses a strict dependence on the priors.

An different approach involves the use of alternative inputs such as depth sensors in order to fit geometric models to the live data. A Kinect sensor is used in [7] for fitting a user-specific expression model to the data (their goal is expression transfer however, and not face reconstruction). This can offer satisfying results, however the approach is limited by the expressiveness of the model.

The use of more camera viewpoints can also increase the confidence of the results. Sparse stereo information is used in [6] for fitting a generic 3D model to the live data. Geometric details are limited due to the sparsity of the features in combination with a generic model. Models that are reconstructed using a multitude of cameras such as in [1] can offer excellent reconstruction quality at the expense of applicability and computational performance. Such methods generally require fewer priors and are thus less sensitive to relaxations of the contextual assumptions.

The hybrid approach that is presented here attempts to find a compromise between the convenient use of live data and flexible use of prior information.

3 Prior-aided Stereo Matching

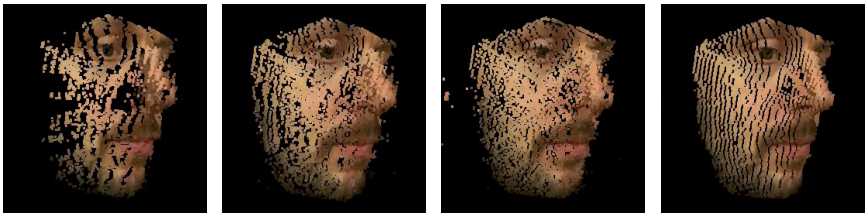
For the stereo matching process, prior information is ingested in the form of AAM (Active Appearance Models [3]) features along with a linear interpolation model on top of these features. AAM is applied to both rectified stereo images, thus introducing a sparse correspondence set S of features between the two images. In total 68 correspondences are used. These are triangulated in accordance with the face anatomy. The result is a lower- and upper bound (d_{min}, d_{max}) for the disparity search range in each triangle. In order to facilitate for small errors in the correspondence set, a fixed value δ is used to widen the search range.

The dense stereo matching algorithm is based on [9] and employs a locally adaptive window in order to accommodate to the local texture variations. Search ranges are limited using the local lower- and upper bounds. Integral images are used in order to accelerate the computation of the similarity score over variable rectangular window shapes.

The traditional winner-takes-all approach for selecting the final disparity value is replaced by an iterative approach where a per-pixel Gaussian Mixture Model (GMM) is used for including temporal- and neighborhood influences. First of all the temporal influence is modeled for each pixel by adding the corresponding GMM from the previous time instance. A Gaussian that models the stereo matching result is then added, and is followed by a number of iterations that consist out of combining each GMM with those of its neighbors and reducing these GMMs in order to limit the number of embedded Gaussians. GMM re-

duction is done closed-form by means of least-squares parabolic fitting in the log-domain.

Figure 1 illustrates some of the results with varying stages of the method in use. The effective resolution of the shown stereo-matched face is about 110x130 (bounding box around the AAM features). The method, which uses the stereo images along with the accompanying AAM features as input, runs at 22 frames per second for the shown results on an Intel Core i7 2820QM processor.



(a) (b) (c) (d)

Fig. 1: Incrementally improving stereo matching results by (a) restricting computation to the facial region, (b) additionaly constraining the disparity domain based on the sparse AAM correspondence set and increasing precision to 1/2 pixel resolution, (c) adding the temporal factor and (d) also including the neighborhood influence.

4 3D Model Animation by 2D Morphing

The live stereo matched depth image from the previous section is not enough to reconstruct the complete face due to occlusions and incomplete stereo matched data. To this purpose, a personalized 3D scan is available as a prior and needs to be adapted in order to reflect the actual facial expression state. The same AAM features as used in the previous section are applied here in order to transfer the live expressions to the 3D model. The resulting animated model consists out of a number of depth images rather than a mesh model. As will become clear in the next section, this is not necessarily a disadvantage.

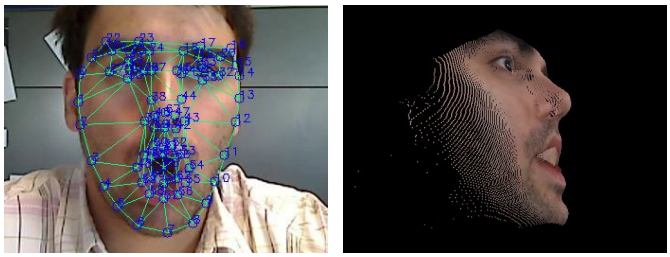
The model animation consists out of the following steps. First of all, the live AAM features need to be transferred to the 3D model; in other words the 3D coordinates of the AAM features are estimated. Once this is done, the 3D data is adapted in order to adhere to these new AAM coordinates. This is done in the 2D domain by projecting the 3D model a number of times, and doing 2D depth morphing from the original AAM coordinates to the new ones.

A common reference frame is constructed between the live- and offline data by means of matching the virtual camera to the live camera. The intrinsic camera parameters are directly transferred from the live camera whereas the extrinsic camera parameters are estimated by aligning three selected AAM points. Now the AAM features of the live data can be transferred by simply copying the 2D coordinates. The 3D coordinates of these transferred AAM points are then found by estimating the depth for the points, and back-projecting them back into 3D

space. In this case the depth of each AAM point is taken to be the same as that of the relevant original AAM point.

Now that the new 3D coordinates of the AAM features are known, we can transform the model by employing 2D morphing on an arbitrary number of generated depth images. The generated images should cover the complete face. The data is morphed by using a coarse mesh that spans the head and which is a superset of the AAM-based mesh. The AAM points are placed at their actual location, and the other points are linearly morphed in relation to the mesh triangle they belong to.

Figure 2 shows an input frame that has been augmented by AAM features and the triangular model, and a morphed 2D+depth map that has been back-projected. Note the coarse animation of the mouth (the mouth is closed in the source 3D model).



(a) AAM model on live data (b) Animated model (1 view)

Fig. 2: 3D model animation.

5 3D Surface Reconstruction

The previous two sections discussed ways of generating depth images of the model that is to be reconstructed. These are essentially point clouds (the associated calibration matrices are available), and can thus not be readily rendered in a flexible manner. The generation of a 3D surface model will remedy this.

The approach is based on [4] where an implicit surface function called the truncated signed distance function (TSDF) is constructed using a number of range images. This TSDF contains a signed distance metric d and a reliability value r for each sample. The distance metric indicates the nearest distance to the surface that is reconstructed; the sign defining whether it is located inside or outside of the object. In order to make the method scalable, an adaptive octree is used for sampling the implicit function. An additional iterative filter on the implicit surface function furthermore refines the results. This allows for a high quality surface reconstruction combined with real-time computational performance.

At first, the considered 3D space is uniformly sampled at a low sampling rate. This sampling rate is chosen in function of the minimum object size one wants to reconstruct. The diagonal of the cube in between samples cannot be

larger than the diagonal of the minimal inner sphere of the reconstructed object. The sampling space is then refined in an adaptive manner by evaluating the implicit surface function at each sample location. A cube is refined when it contains an edge in which the TSDF of the endpoints have opposite signs. This is done until the maximum octree depth is reached. As this process does not necessarily produce a uniform sampling frequency along the surface boundaries, an additional post-processing step is needed for refining the octree. Note that the maximum octree depth introduces an implicit scalability parameter for the method. By increasing the value one gets a higher detailed model at the expense of computational resources.

The TSDF at a certain 3D point is estimated by projecting the point to each depth image, and selecting the one with the best score. This score is based on the distance and its reliability. Note that the reliability of the measurement is decreased in relation with the distance *inside* the object. Indeed, when a measurement indicates that a point is behind the surface, it does not necessarily mean that the point is actually inside the object; the object can merely be occluding a free point in space.

A filtering phase is introduced on the TSDF in order to reduce the influence of noise or missing values in the depth images. The filter models the neighboring sample influence on the assumption that the implicit surface is a plane:

$$D'(p) = R(p)^{\alpha} D(p) + (1 - R(p)^{\alpha}) D_n(p) \tag{1}$$

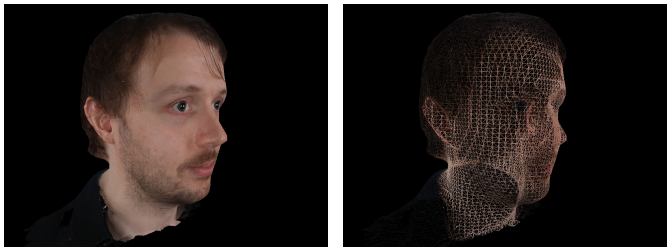
$$D_n(p) = \frac{\sum_{(x,y)\in N(p)} (R(x) + R(y)) \frac{(D(x) + D(y))}{2}}{\sum_{(x,y)\in N(p)} (R(x) + R(y))} \tag{2}$$

with $D(x)$ and $R(x)$ the distance and reliability components of the TSDF, α a parameter to control the neighbor influence and $N(x)$ the set of valid neighboring pairs in each dimension ($max(\|N(x)\|) = 3$). $R'(p)$ is calculated in a similar way with $R_n(p)$ being the average reliability of the valid neighbors.

The last step involves the generation of an explicit surface from the implicit model. This is done by generating a triangle mesh using the Marching Cubes algorithm [5].

Figure 3 shows a closed mesh model that was reconstructed using this method. 5 synthetically generated depth images have been used as an input and it consists out of 27.3k triangles. Table 1 contains the reconstruction errors and frame rate at 4 different maximum octree levels. No filtering was used for these measurements. Figure 4 shows the spatial distribution of the errors over the front of the model. Red indicates the smallest error; blue the largest (clipped to $2mm$). The error statistics and images were created using Metro [2].

Filtering performance is illustrated in figure 5 and considers two cases: gap filling and noise reduction. Gap filling is illustrated by adding unknowns to the depth values in the depth image. For demonstrating noise reduction, Gaussian



(a) Textured (b) Wireframe

Fig. 3: 3D Surface reconstruction.

Table 1: Surface reconstruction statistics

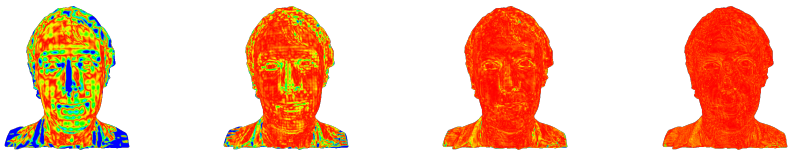
Model	Triangle count	Average error *mm (% of diag.)*	RMS *mm (% of diag.)*	Frames per second
Reference	876069	-	-	-
Level 5	6720	0.74692 (0.147%)	1.31329(0.258%)	20.78
Level 6	28212	0.25322 (0.050%)	0.42232 (0.083%)	9.15
Level 7	117024	0.11558 (0.023%)	0.17305 (0.034%)	3.02
Level 8	484072	0.08722 (0.017%)	0.11662 (0.023%)	0.93

white noise has been added to the depth image (the reliability noise is also modeled using Gaussian white noise on the reliability reduction that is proportional to the introduced error). A single depth image was used for reconstruction.

6 Live 3D Face Reconstruction

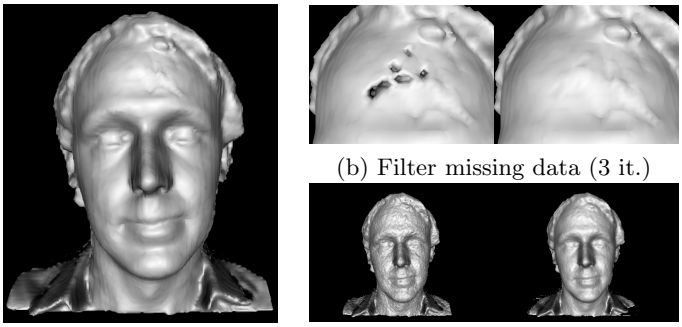
This final step brings the three previous sections together. As stated before, the goal involves the aggregation of live data and offline scanned data in order to produce a live 3D face mesh. The live depth images are produced by means of prior-aided stereo matching, the synthetic depth images are generated by 3D model animation using 2D morphing and all are aggregated using the 3D surface reconstruction method. Registration between the live- and synthetic depth images is done based on the 3D locations of the AAM points. One important aspect for the aggregation is the choice of reliability metric for the depth images. The metric that is used for the stereo matched depth images is related to the variance within each resulting GMM. A wide variance implies a less reliable value. For the synthetic depth images this metric is statically determined.

A result that uses one live stereo matched depth image combined with 5 animated depth images is shown in figure 6. Due to the unavailability of ground truth data, only a qualitative assessment can be given. The method performs well in regions where the animated depth images coarsely agree with the stereo matched data. The linear model that is used for animating the 3D model has



(a) Level 5 (b) Level 6 (c) Level 7 (d) Level 8

Fig. 4: Surface reconstruction error - ground truth annotation. Red: 0; blue: $\geq 2mm$.



(a) Original (b) Filter missing data (3 it.) (c) Filter Gaussian noise (10 it.)

Fig. 5: Implicit surface filter.

some deficiencies however, especially in the region of the cheeks. Despite this, the combination of a well chosen reliability metric and implicit surface filtering allows for a high quality 3D reconstruction using live data.

Real-time rates are not fully achieved for the complete system at the time of writing. This can be attributed to implementation details however, and not due to inherent computational constraints.

7 Conclusions

A method was presented that employs depth images from different sources in order to generate a live 3D mesh of a known face at potentially real-time rates. Prior information is used at different stages of the method in order to enhance quality and reinforce reliability. The generated mesh is the result of combining information from live stereo matched data and an offline scanned animated mesh in a way that is computationally tractable and resilient to errors.

Future work includes the replacement of the coarse animated mesh prior by a more accurate model; e.g. by means of a 3D morphable model in a face/expression space. We believe that the combination of such a model with real-time data could relieve this model from the disadvantages thereof. In addition, dynamically predicting the model reliability according to the current circumstances will also be investigated.

(a) Live input (stereo pair) (b) Textured mesh (c) Wireframe mesh

Fig. 6: Live 3D face reconstruction.

Acknowledgments. This research was carried out as part of the IBBT/IWT iCocoon project. This project is funded by the Interdisciplinary Research Institute IBBT and by the Agency for Innovation by Science and Technology (IWT-Flanders).

References

1. Beeler, T., Hahn, F., Bradley, D., Bickel, B., Beardsley, P., Gotsman, C., Sumner, R.W., Gross, M.: High-quality passive facial performance capture using anchor frames. ACM Trans. Graph. 30, 75:1–75:10 (August 2011)
2. Cignoni, P., Rocchini, C., Scopigno, R.: Metro: measuring error on simplified surfaces. Tech. rep., Paris, France, France (1996)
3. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: ECCV (2). pp. 484–498 (1998)
4. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. pp. 303–312. SIGGRAPH '96, ACM, New York, NY, USA (1996)
5. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. SIGGRAPH Comput. Graph. 21(4), 163–169 (Aug 1987)
6. Park, U., Jain, A.K.: 3d face reconstruction from stereo video. Computer and Robot Vision, Canadian Conference 0, 41 (2006)
7. Weise, T., Bouaziz, S., Li, H., Pauly, M.: Realtime performance-based facial animation. In: ACM SIGGRAPH 2011 papers. pp. 77:1–77:10. SIGGRAPH '11, ACM, New York, NY, USA (2011)
8. Widanagamaachchi, W., Dharmaratne, A.: 3d face reconstruction from 2d images. In: Computing: Techniques and Applications, 2008. DICTA '08.Digital Image. pp. 365 –371 (dec 2008)
9. Zhang, K., Lu, J., Lafruit, G.: Scalable stereo matching with locally adaptive polygon approximation. In: Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on. pp. 313 –316 (oct 2008)