Supplemental Document: A Versatile Scene Model with Differentiable Visibility Applied to Generative Pose Estimation

Helge Rhodin[1] Nadia Robertini[1, 2] Christian Richardt[1, 2] Hans-Peter Seidel[1] Christian Theobalt[1]

[1] MPI Informatik [2] Intel Visual Computing Institute

1. Introduction

In our paper, we introduce a new visibility and image formation model that is differentiable everywhere. Applied to generative pose estimation, it improves convergence of numerical optimization. In this supplemental document, we explain in more detail, and generality, the following parts of our main paper:

- Full details on the scene and model parameters of all our experiments (Section 2).
- Application of the visibility model to geometry and appearance estimation (Section 2.1).
- The qualitative comparison to Stoll *et al*. on the *Walker* sequence [2] (Section 2.2).
- Impact of varying the image resolution (Section 2.3).
- Analytic gradients of visibility and introduced objective functions (Section 3).

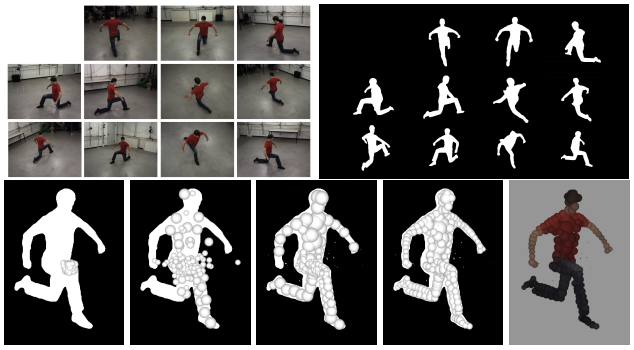

Figure 1. Shape and appearance estimation using the photo-consistency F_{pc}. Top: Input RGB images and extracted silhouettes. Bottom: reconstruction process from an unused camera view. From left to right: initialization, after 100, 300, 10000 iterations, and color back-projection. Each Gaussian is represented by a sphere of radius equal to its standard deviation.

2. Results

Details on all input sequences, complexity of the utilized models, and optimization results are given in Table 1.

2.1. Shape optimization

The goal of this experiment is to evaluate the applicability of our image formation model to geometry and appearance estimation from multiple RGB images. Input to the method are 11 RGB images from calibrated cameras and corresponding silhouettes obtained by background subtraction (see Figure 1). We initialize the model with 200 small white Gaussians positioned randomly around the center of the capture volume. Subsequently, the position $\boldsymbol{\mu}_q$ and size σ_q of each Gaussian G_q is optimized for photo-consistency F_{pc} between silhouette images and model density.

Figure 1 shows the reconstructed shape after 100, 300 and 10000 gradient iterations from a 12th camera which is not used for optimization. This verifies that the geometry of an object can accurately be estimated in the same manner as the object pose. Note that a few Gaussians are pushed outside of the silhouette (presumably due to too large gradient steps) and vanish in size. These could be removed in a post-process.

The color of each Gaussian G_q is inferred by the weighted average over the pixel colors of all pixels (u, v) in all RGB input images, weighted by the corresponding Gaussian visibility $\mathcal{V}_q((u, v), \boldsymbol{\gamma})$. This shape estimation of a human actor is a special instance of the actor model creation step proposed by Stoll *et al*. [2], who optimize a parametric actor model that consists of Gaussians constrained to move with a skeleton. We show that our method is versatile enough to approximate the actor's shape without positional constraints between blobs.

Sequence	*Soccer* (two actors)		*Soccer* (one actor)		*Marker*	*Walker*	Shape estimate	Rigid objects
Published by	Elhayek *et al.* [1]					Stoll *et al.* [2]		Our
Number of cameras	3				2	4	11	1
Number of frames	300				500	522	1	
Frame rate	23.8				25	25	n/a	
Camera type	mobile phone *(HTC One X)*				PhaseSpace Vision Camera			simulated
Raw image resolution	1280×720				1296×972			256×256
Environment	outdoor, uncontrolled background and lighting				studio, uncontrolled background			synthetic
Tracked subjects	2		1		1			2
Number of joints	118		59		61	66	0	
Number of parameters	84		42		44	43	800	9
Number of Gaussians	182		72		72	77	200	28
Input image resolution	320×180	640×360	320×360	640×360	324×243			128×128
Input pixels per frame	≈12,000	≈202,000	≈7,000	≈122,000	≈10,000	≈25,000	≈80,000	≈16,000
Ground truth	Manual annotation, 3D triangulation				Marker	12 cam	n/a	constructed
Average error [cm]	4.81	4.69	5.88	4.70	3.79	2.55	n/a	
Timing [iterations/s]	3.33	0.23	10.0	0.86	8.1	5.01	0.68	2.14

Table 1. A table describing each scene and the relevant parameters, such as number, type and resolution of cameras, pose parameter, run time per gradient iteration, and reconstruction error (average Euclidean 3D distance over all joints and frames to the ground truth).

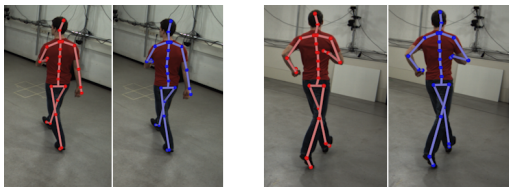

Figure 2. Evaluation of our method on the *Walker* sequence [2]. With only four cameras, our method (blue) is able to accurately track the whole sequence containing walking (left) and jogging (right) motions. Here compared to 12 camera tracking with the method of Stoll *et al.* (red).

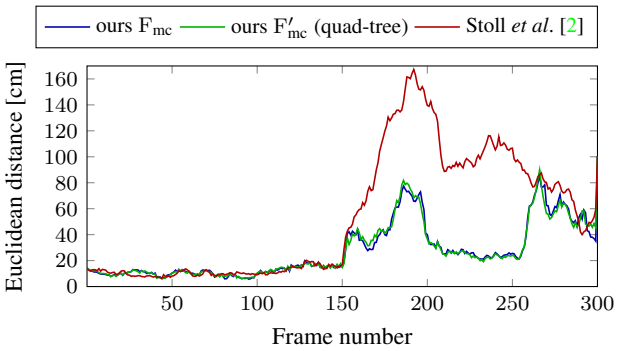

Figure 3. Comparison of F_{mc} and F'_{mc}, based on the Euclidean 3D joint position error of all limb joints with respect to marker based ground truth. The impact of using a hierarchical representation is much smaller than the improvement on Stoll *et al.*

2.2. Additional tracking Experiment

We further evaluate the proposed visibility model on Stoll *et al.*'s *Walker* sequence [2], for which no ground truth is available. Instead, a reimplementation of the method of Stoll *et al.* using 12 cameras is used as reference. The same skeleton with 77 pose parameters and 43 Gaussians and image resolution 324×243 is used for both methods. The approach of Stoll *et al.* applied on the full available camera acquisition setup, consisting of 12 cameras, produces qualitatively comparable results to our approach applied on 4 cameras only, see Figure 2. The average Euclidean 3D joint position distance between both results is only 2.55 cm. At the point of very fast arm-jogging motions with strong occlusions, small errors are visible, however, our method recovers quickly. Please watch the supplemental video for tracking results over the whole sequence.

2.3. Input image resolution and thresholding

The method of Stoll *et al.* operates on a hierarchical image representation, that clusters regions of similar color in a quad-tree, whereas our method operates on pixels. To verify that the gained improvements are primarily due to our introduced visibility model, and not due to different image resolutions, we run our algorithm on the studio sequence with the same quad-tree representation that models squared areas of similar color by a single pixel of corresponding size. To make our energy model applicable to representations with varying pixel size, we construct the energy F'_{mc}, which is equivalent to F_{mc}, but weights each pixel by its area. The influence of the hierarchical representation on the reconstruction quality is much smaller than the improvement on Stoll *et al.*, as shown in Figure 3.

Moreover, we validate that the error due to excluding model blobs with negligible contribution (Section 3.3 in the main paper) is vanishingly small; the average error across the first 100 frames of the *Marker* sequence is increased by only 0.0038 cm (0.1% of the total error).

3. Analytic gradients

In this section, we provide the analytic gradients of the *Gaussian visibility* $\mathcal{V}_q$ and the objective functions D_{mc} and D_{pc} with respect to the Gaussian parameters $\boldsymbol{\gamma}$. The gradients of D_{mc} and D_{pc} follow immediately from the visibility gradient:

$$\frac{\partial \mathrm{D}_{\mathrm{mc}}(\boldsymbol{\gamma}, \mathbf{I})}{\partial \gamma} = \sum_{(u,v)} \sum_{q} d(\mathbf{I}(u,v), a_q) \frac{\partial \mathcal{V}_q((u,v), \boldsymbol{\gamma})}{\partial \gamma}, \tag{1}$$

and

$$\frac{\partial \mathrm{D}_{\mathrm{pc}}(\boldsymbol{\gamma}, \mathbf{I})}{\partial \gamma} = 2 \sum_{(u.v)\in I} \sum_{i} \left(\hat{L}_i((u,v), \boldsymbol{\gamma}) - \mathbf{I}_i(u,v) \right) \frac{\partial \hat{L}_i((u,v), \boldsymbol{\gamma})}{\partial \gamma}, \tag{2}$$

where $i \in \{R, G, B\}$ ranges over the different color channels, and the radiance gradient is

$$\frac{\partial \hat{L}(\mathbf{o}, \mathbf{n}, \boldsymbol{\gamma})}{\partial \gamma} = \sum_{q} \mathbf{a}_q \frac{\partial \mathcal{V}_q}{\partial \gamma}. \tag{3}$$

The gradient of visibility of Gaussian G_q is

$$\frac{\partial \mathcal{V}_q(\mathbf{o}, \mathbf{n}, \boldsymbol{\gamma})}{\partial \gamma} = \sum_{s \in S_q} \frac{\partial \lambda_q}{\partial \gamma} T(\mathbf{o}, \mathbf{n}, s, \boldsymbol{\gamma}) G_q(\mathbf{o} + s\mathbf{n}) + \lambda_q \frac{\partial T(\mathbf{o}, \mathbf{n}, s, \boldsymbol{\gamma})}{\partial \gamma} G_q(\mathbf{o} + s\mathbf{n}) + \lambda_q T(\mathbf{o}, \mathbf{n}, s, \boldsymbol{\gamma}) \frac{\partial G_q(\mathbf{o} + s\mathbf{n})}{\partial \gamma}, \tag{4}$$

with $\lambda_q = \ell \sigma_q$, for some fixed ℓ, it holds

$$\frac{\partial \lambda_q}{\partial \bar{\sigma}_q} = \ell, \frac{\partial \lambda_q}{\partial \bar{\mu}_q} = 0. \tag{5}$$

The gradient of Gaussian density is

$$\frac{\partial G_q(\mathbf{o} + s\mathbf{n})}{\partial \gamma} = \frac{\partial \exp\left(-\frac{(s-\bar{\mu}_q)^2}{2\bar{\sigma}_q^2} \right) \bar{c}_q}{\partial \gamma}. \tag{6}$$

In the following, we derive derivatives for all sub-terms with respect to the individual Gaussian parameters. The sampling location s depends on $\bar{\sigma}_q$ and $\bar{\mu}_q$ of the Gaussian G_q for which the visibility is inferred. In this case, $\frac{s-\bar{\mu}_q}{\bar{\sigma}_q} = \frac{\bar{\mu}_q + k\ell\bar{\sigma}_q - \bar{\mu}_q}{\bar{\sigma}_q} = k\ell$ and

$$\frac{\partial \bar{c}_q \exp\left(-\frac{(k\ell)^2}{2}\right)}{\partial \bar{\mu}_q} = 0, \quad \frac{\partial \bar{c}_q \exp\left(-\frac{(k\ell)^2}{2}\right)}{\partial \bar{c}_q} = \exp\left(-\frac{(k\ell)^2}{2}\right), \text{ and } \frac{\partial \bar{c}_q \exp\left(-\frac{(k\ell)^2}{2}\right)}{\partial \bar{\sigma}_q} = 0, \tag{7}$$

for parameters $\bar{\sigma}_p, \bar{\mu}_p$ with $p \neq q$, the gradient $\frac{\partial G_q(\mathbf{o}+s\mathbf{n})}{\partial \gamma}$ is zero.

The gradient of transmission T is

$$\frac{\partial T(\mathbf{o},\mathbf{n},s,\boldsymbol{\gamma})}{\partial \gamma} = T(\mathbf{o},\mathbf{n},s,\boldsymbol{\gamma}) \frac{\partial \sum_p \frac{\bar{\sigma}_p \bar{c}_p}{\sqrt{\frac{2}{\pi}}} \left(\operatorname{erf}\left(\frac{-\bar{\mu}_p}{\sqrt{2}\bar{\sigma}_p}\right) - \operatorname{erf}\left(\frac{s-\bar{\mu}_p}{\sqrt{2}\bar{\sigma}_p}\right)\right)}{\partial \gamma}. \tag{8}$$

For the case of constant sampling location, i.e. $\frac{\partial s}{\partial \bar{\mu}} = \frac{\partial s}{\partial \bar{\sigma}} = 0$, it holds

$$\begin{aligned}
-\sqrt{\frac{\pi}{2}} \frac{\partial \bar{\sigma}_p \bar{c}_p \left(\operatorname{erf}\left(\frac{s-\bar{\mu}_p}{\sqrt{2}\bar{\sigma}_p}\right) - \operatorname{erf}\left(\frac{-\bar{\mu}_p}{\sqrt{2}\bar{\sigma}_p}\right)\right)}{\partial \bar{\mu}_p} &= \bar{c}_p \left(\exp\left(-\frac{(s-\bar{\mu}_p)^2}{2\bar{\sigma}_p^2}\right) - \exp\left(-\frac{(-\bar{\mu}_p)^2}{2\bar{\sigma}_p^2}\right)\right), \\
-\sqrt{\frac{\pi}{2}} \frac{\partial \bar{\sigma}_p \bar{c}_p \left(\operatorname{erf}\left(\frac{s-\bar{\mu}_p}{\sqrt{2}\bar{\sigma}_p}\right) - \operatorname{erf}\left(\frac{-\bar{\mu}_p}{\sqrt{2}\bar{\sigma}_p}\right)\right)}{\partial \bar{c}_p} &= -\sqrt{\frac{\pi}{2}} \bar{\sigma}_p \left(\operatorname{erf}\left(\frac{s-\bar{\mu}_p}{\sqrt{2}\bar{\sigma}_p}\right) - \operatorname{erf}\left(\frac{-\bar{\mu}_p}{\sqrt{2}\bar{\sigma}_p}\right)\right), \\
-\sqrt{\frac{\pi}{2}} \frac{\partial \bar{\sigma}_p \bar{c}_p \left(\operatorname{erf}\left(\frac{s-\bar{\mu}_p}{\sqrt{2}\bar{\sigma}_p}\right) - \operatorname{erf}\left(\frac{-\bar{\mu}_p}{\sqrt{2}\bar{\sigma}_p}\right)\right)}{\partial \bar{\sigma}_p} &= -\sqrt{\frac{\pi}{2}} \bar{c}_p \left(\operatorname{erf}\left(\frac{s-\bar{\mu}_p}{\sqrt{2}\bar{\sigma}_p}\right) - \operatorname{erf}\left(\frac{-\bar{\mu}_p}{\sqrt{2}\bar{\sigma}_p}\right)\right) \\
&\quad + \frac{\bar{c}_p (s-\bar{\mu}_p)}{\bar{\sigma}_p} \exp\left(-\frac{(s-\bar{\mu}_p)^2}{2\bar{\sigma}_p^2}\right) + \frac{\bar{c}_p (-\bar{\mu}_p)}{\bar{\sigma}_p} \exp\left(-\frac{(-\bar{\mu}_p)^2}{2\bar{\sigma}_p^2}\right).
\end{aligned} \tag{9}$$

And for the case of sampling locations s depending on G_q, i.e. $\frac{\partial s}{\partial \bar{\mu}} \neq 0$ and $\frac{\partial s}{\partial \bar{\sigma}} \neq 0$, we substitute $\frac{s-\bar{\mu}_q}{\bar{\sigma}_q} = \frac{\bar{\mu}_q + k\ell\bar{\sigma}_q - \bar{\mu}_q}{\bar{\sigma}_q} = k\ell$ as before and consider the special cases

$$\begin{aligned}
-\sqrt{\frac{\pi}{2}} \frac{\partial \bar{\sigma}_q \bar{c}_q \operatorname{erf}\left(\frac{k\ell}{\sqrt{2}}\right)}{\partial \bar{\mu}_q} &= 0, \\
-\sqrt{\frac{\pi}{2}} \frac{\partial \bar{\sigma}_q \bar{c}_q \operatorname{erf}\left(\frac{k\ell}{\sqrt{2}}\right)}{\partial \bar{c}_q} &= -\sqrt{\frac{\pi}{2}} \bar{\sigma}_q \operatorname{erf}\left(\frac{s-\bar{\mu}_q}{\sqrt{2}\bar{\sigma}_q}\right), \\
-\sqrt{\frac{\pi}{2}} \frac{\partial \bar{\sigma}_q \bar{c}_q \operatorname{erf}\left(\frac{k\ell}{\sqrt{2}}\right)}{\partial \bar{\sigma}_q} &= -\sqrt{\frac{\pi}{2}} \bar{c}_q \operatorname{erf}\left(\frac{s-\bar{\mu}_q}{\sqrt{2}\bar{\sigma}_q}\right), \\
-\sqrt{\frac{\pi}{2}} \bar{\sigma}_p \bar{c}_p \frac{\partial \operatorname{erf}\left(\frac{\bar{\mu}_q + k\ell\bar{\sigma}_q - \bar{\mu}_p}{\sqrt{2}\bar{\sigma}_p}\right)}{\partial \bar{\mu}_q} &= -\bar{c}_q \exp\left(-\frac{(s-\bar{\mu}_q)^2}{2\bar{\sigma}_q}\right) \\
-\sqrt{\frac{\pi}{2}} \bar{\sigma}_p \bar{c}_p \frac{\partial \operatorname{erf}\left(\frac{\bar{\mu}_q + k\ell\bar{\sigma}_q - \bar{\mu}_p}{\sqrt{2}\bar{\sigma}_p}\right)}{\partial \bar{\sigma}_q} &= -\bar{c}_q k\ell \exp\left(-\frac{(s-\bar{\mu}_q)^2}{2\bar{\sigma}_q}\right).
\end{aligned} \tag{10}$$

Finally the derivatives of the ray density of Gaussian G_q with respect to their 3D parameters are

$$\begin{aligned}
&\frac{\partial \bar{\sigma}_q}{\partial \sigma_q} = 1, \frac{\partial \bar{\sigma}_q}{\partial \mu_q} = 0, \\
&\frac{\partial \bar{\mu}_q}{\partial \sigma_q} = 0, \frac{\partial \bar{\mu}_q}{\partial \mu_q} = n, \\
&\frac{\partial \bar{c}_q}{\partial \sigma_q} = \bar{c}_q \frac{(\mathbf{o}-\boldsymbol{\mu}_q)^\top (\mathbf{o}-\boldsymbol{\mu}_q) - \bar{\mu}_q^2}{\bar{\sigma}_q^3} + \frac{\bar{c}_q}{c_q} \frac{\partial c_q}{\partial \sigma_q}, \text{ and} \\
&\frac{\partial \bar{c}_q}{\partial \boldsymbol{\mu}_q} = \bar{c}_q \frac{(\mathbf{o}-\boldsymbol{\mu}_q) + \bar{\mu}_q \mathbf{n}}{\bar{\sigma}_q^2}.
\end{aligned} \tag{11}$$

References

[1] A. Elhayek, E. Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In *CVPR*, 2015.

[2] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of Gaussians body model. In *ICCV*, pages 951–958, 2011.