max planck institut informatik

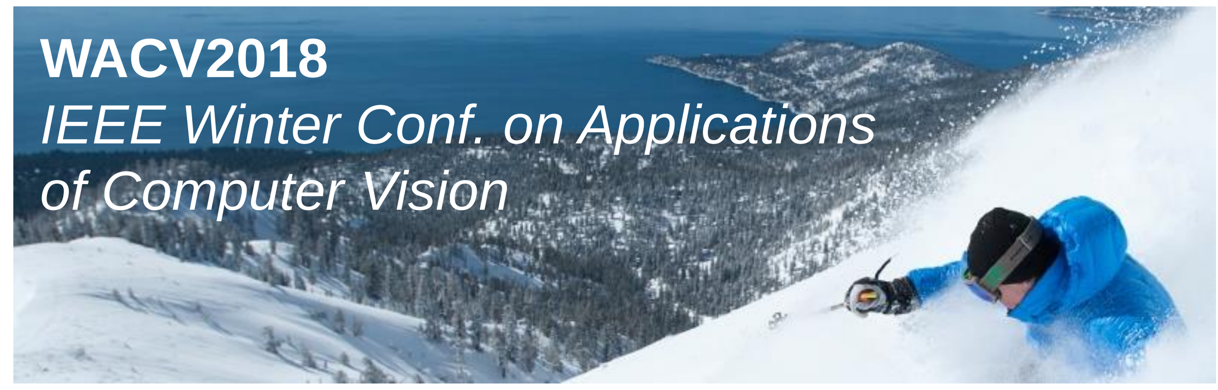
Graphics Vision & Video

intel VISUAL COMPUTING INSTITUTE

Illumination-invariant Robust Multiview 3D Human Motion Capture

Nadia Robertini, Florian Bernard, Weipeng Xu, Christian Theobalt

MPI for Informatics, Intel VCI (Saarbruecken, Germany)

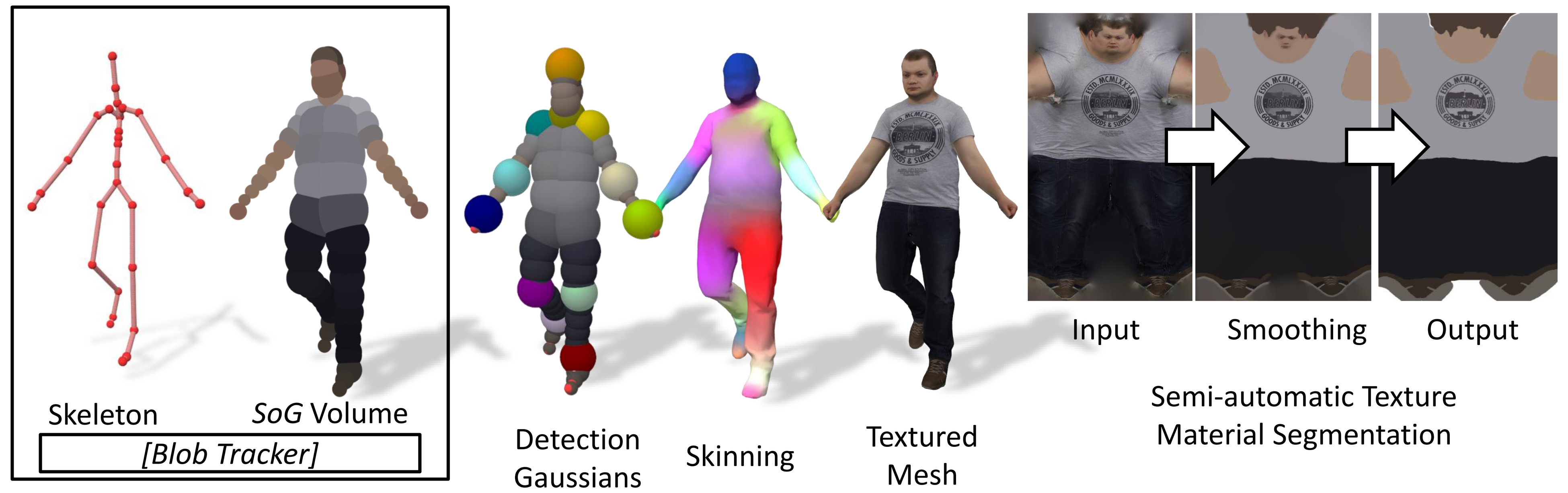


Introduction:

- Our goal is to capture human body motion under changing lighting conditions in a multiview setup.

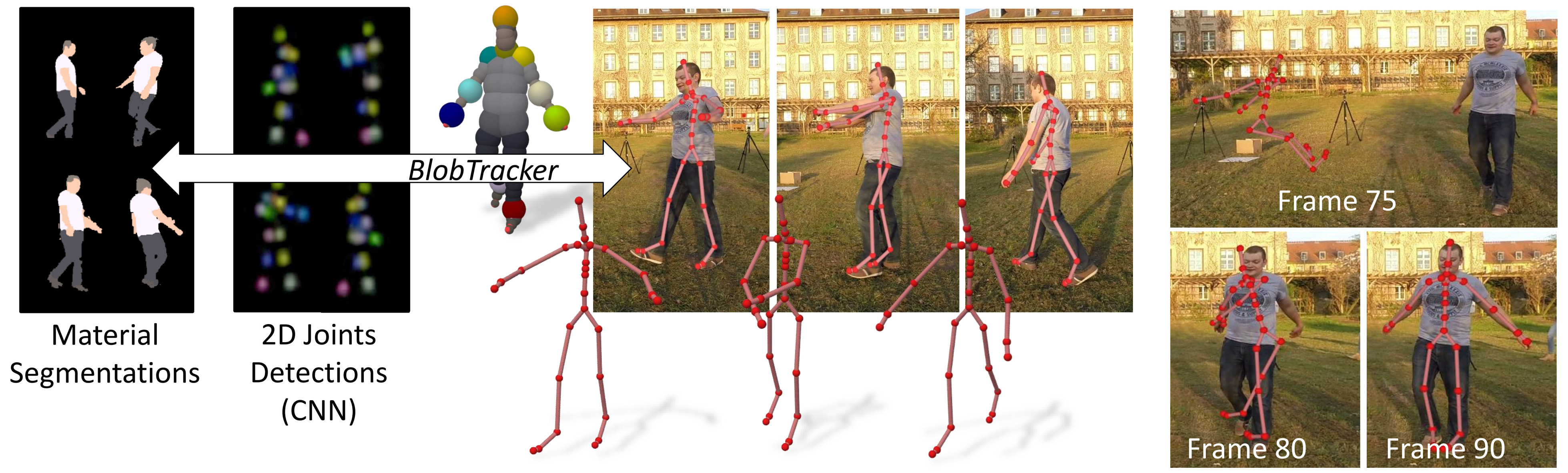
Actor Model:

- We augment the highly simplified *BlobTracker* human model introduced by *[Stoll et al.]* with a textured mesh (automatically skinned to the skeleton) with labeled materials.



Pose Tracking:

- **Goal**: Run our augmented *BlobTracker* approach taking as input optimal illumination-invariant **material segmentations** (with influence w_s) as well as **2D joint detections** (w_d) to robustly estimate the body motion.

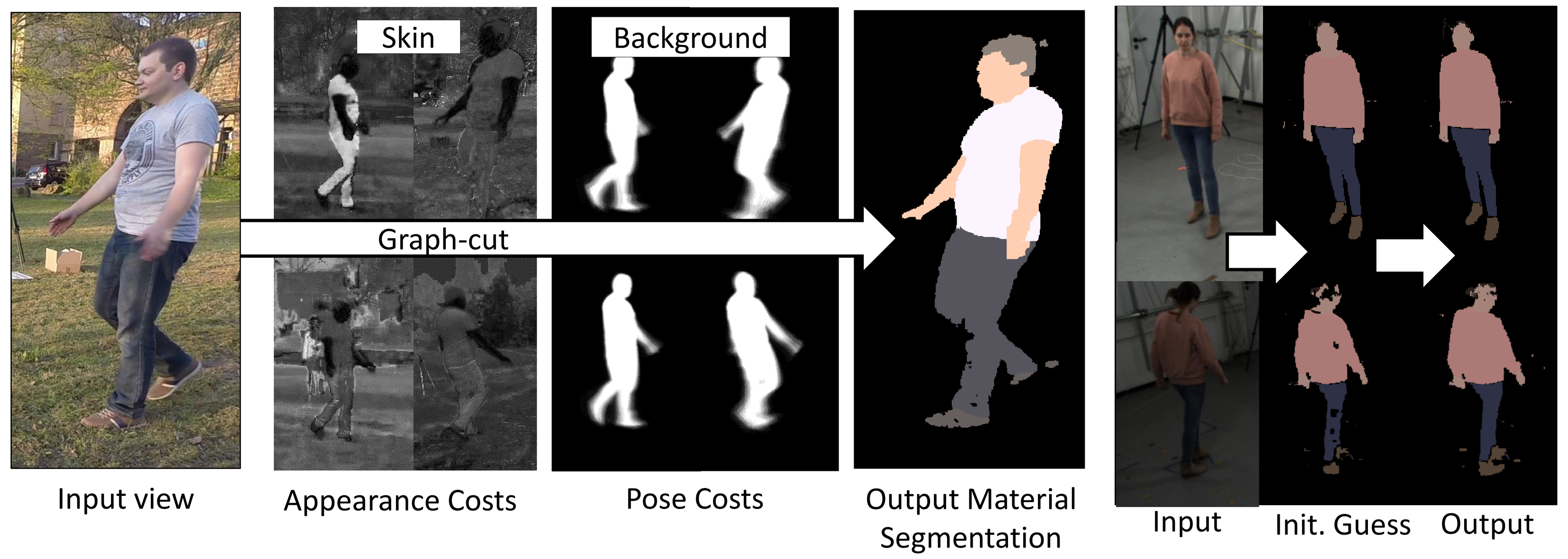


- **Key idea**: design an **iterative** approach to alternatively estimate materials and body pose using **temporal cues**.
- **Adaptive weighting (w_s, w_d)**: temporally measure the quality of material segmentations (*e.g.* abrupt changes) and scale down/up relevance for tracking accordingly.

[*BlobTracker*]: 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, 2011

Lighting-Invariant Segmentation:

- **Goal**: obtain temporally and spatially consistent material segmentations, which are invariant from background complexity and appearance changes due to light, to feed to *[Stoll et al.]*.



- **Graph-cut Energy**: cost of assigning material label ℓ_i to pixel $i, \forall i \in I$ (each frame/view is solved independently):

$$E(\boldsymbol{\ell}) = \sum_{i=1}^{|I|} [\boldsymbol{E}_i^p(\ell_i) * \boldsymbol{E}_i^a(\ell_i)] + \sum_{i \sim j} E_{ij}(\ell_i, \ell_j)$$

- **Pose Costs**: sample *50* random poses from a Gaussian distribution around the current pose prediction P^t based on previous P^{t-1}, P^{t-2}:

$$E_i^p(\ell_i) = 1 - H_{\ell_i}(x_i)$$

- **Appearance Costs**: *Mahalanobis* distance between pixels and labels:

$$E_i^a(\ell_i) = (\Phi(\mathrm{x_i}) - \mu_\ell)^T C_\ell^{-1} (\Phi(\mathrm{x_i}) - \mu_\ell)$$

  - Feature image $\Phi(x_i) = [sin(h_{x_i}), cos(h_{x_i}), s_{x_i}]$
  - Background feature $\Phi_{BG}(x_i) = [\Phi(x)^T, E_i^a(\ell_1), \dots, E_i^a(\ell_{L-1})]$
  - Material **geometric median** μ_ℓ and **covariance** C_ℓ on the *pose predicted locations* $\mathrm{X}_\ell = \{x_i | H_{\ell_i}(x_i) > t\}$:

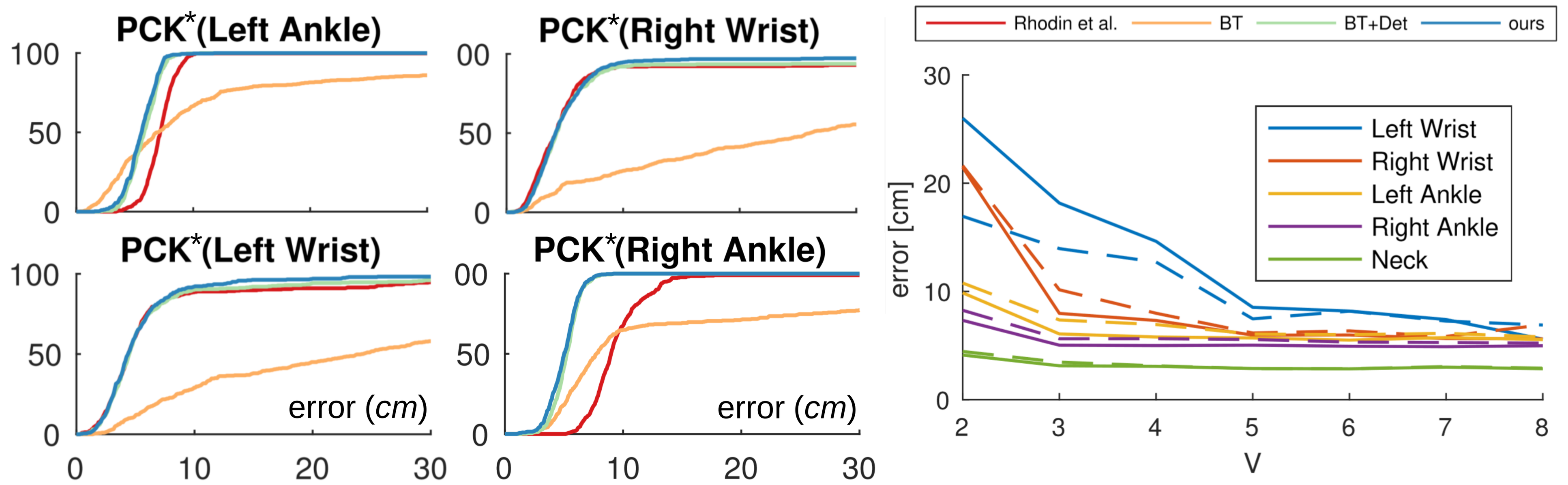
$$\mu_\ell = \underset{y}{\mathrm{argmin}} \sum_{x \in \mathrm{X}_\ell} ||x - y||_2, \ \boldsymbol{C_\ell} = \frac{1}{|\mathrm{X}_\ell| - 1} \sum_{x \in \mathrm{X}_\ell} (x - \mu_\ell)(x - \mu_\ell)^T$$

- **Smoothness**: neighboring pixels with similar color have similar materials:

$$E_{ij}(\ell_i, \ell_j) = \exp\left(\frac{\left\|I(x_i) - I(x_j)\right\|_2^2}{2}\right) \min(1, |\ell_i - \ell_j|)$$

Results:

- Our **quantitative** and **qualitative** results evidence that our approach accurately tracks the human pose and outperforms the existing methods.



AUC values of PCK curves above:

	Rhodin et al.	BT	BT+Det	ours
LW	0.9249	0.6858	0.9295	**0.9428**
RW	0.9298	0.6023	0.9326	**0.9451**
LA	0.8839	0.7979	0.9075	**0.9114**
RA	0.8279	0.7003	0.9061	**0.9105**

* Percentage of Correct Key Points

Average Ground Truth Errors (*cm*):

	Rhodin et al.	BT	BT+Det	ours
LW	7.35±9.68	30.92±26.18	6.89±8.77	**5.59**±4.91
RW	7.20±11.39	41.02±35.09	6.91±9.73	**5.61**±6.32
LA	7.28±3.05	12.69±14.35	5.79±1.28	**5.55**±1.33
RA	9.60±3.34	16.73±16.82	5.22±1.14	**4.98**±1.25
N	3.23±1.45	8.34±5.71	2.91±1.18	**2.86**±1.26