Interactive Human Generation

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By
강찬희

Table of Contents

- Three-dimensional Reconstruction of Human Interactions (CVPR 2020)
  - Contribution
  - CHI3D, FlickrCI3D

- Generative Proxemics: A prior for 3D social Interactions from Images (arxiv)
  - Contribution
  - BUDDI
  - Proposed Method
  - Evaluation metrics

Three-dimensional Reconstruction of Human Interactions

- Contribution
  - Proposed models for interaction signature estimation (ISP)
  - Investigate correlations between contact detection, segmentation and 3d contact signature prediction in 3d reconstruction
  - Construct several large datasets for learning and evaluate 3d contact prediction and reconstruction methods. (CHI3D, FlickrCI3D)
    - First datasets with ground-truth labels for the body regions in contact between humans

Three-dimensional Reconstruction of Human Interactions

- Closed Human Interactions 3D (CHI3D)
  - Environment setting
    - 10 motion cameras synchronized with 4 additional RGB cameras
    - In each recordings, one subject is motion tracked with a marker-based motion capture system
    - The second person is tracked using only RGB cameras
  - Contents
    - TRAIN set : 3 pairs of subjects
    - TEST set : 2 pairs of subjects
    - 4 different views
    - 900 x 900 resolution
    - Camera parameters : extrinsics, intrinsics (one assuming image distortion, one ignoring it)

Three-dimensional Reconstruction of Human Interactions

- FlickrCI3D Classification
  - Dataset
    - Images from the YFCC100M dataset (Images from Flickr by amateur photographers)
    - 55,095 images of 90,167 pairs of people in interaction scenarios
    - Classified by annotators in 3 contact classes (trainset)
      - No Contact: 49,372 pairs
      - Uncertain contact: 17,197 pairs
      - Contact: 14,733 pairs
- FlickrCI3D Signature
  - Dataset
    - 11,770 images of 14,866 pairs of people in contact with annotated contact signatures
    - Each contact signature represents a set of annotated correspondences between contact surfaces of two human bodies
    - Correspondences are provided on both GHUM and SMPLX templates, between
      - Vertex IDS, Facet IDs, Body region IDs (75 regions)

Generative Proxemics

- Generative Proxemics: A prior for 3D Social Interaction from Images
  - Contribution
    - Reconstruct pseudo-ground truth 3D meshes of interacting people with an optimization approach using existing ground-truth contact map (FlickrCI3D signatures dataset)
    - Model proxemics using a diffusion model that learns the joint distribution of two people in close social interaction directly in the SMPL-X parameter space
    - Introduce a new optimization method that uses the diffusion prior to reconstruct two peopl in close proximity from a single image without any contact annotation

Generative Proxemics

- Flow
  - Optimization process to create 3D pseudo-ground truth data
    - Use ground truth contact maps form the FlickrCI3D Signatures dataset
    - 2D image to 3D human interaction
  - Train diffusion model that learns the 3D proxemics prior between two people
    - The output from optimization process is used as training data
    - Random noise to 3D human interaction
  - Reconstruction of two people in close proximity from images
    - Do not require any ground-truth contact maps
    - Image to 3D human interaction

Generative Proxemics

- Optimization process
  - Body model representation
    - Use SMPL-X in baseline
    - Additionally use SMIL to support producing meshes for infants and children (interpolation)
  - Input
    - Discrete human-human contact annotations
    - Detected 2D keypoints
    - Initial estimates for pose, orientation, shape and translation $(\theta_0, \phi_0, \beta_0, \gamma_0)$
    - Initial estimates are provided from the output of BEV
    - Use least-square method to convert SMPL output from BEV to SMPL-X

Generative Proxemics

- Optimization process
  - Binary contact map
    - FlickrCI3D Signatures dataset divide the body into 75 regions and annotate their pairwise contact between both people.
    - Each region r, roughly covers a similar surface of the body and is associated with SMPL-X faces $\boldsymbol{f}_r$ and vertices $\boldsymbol{v}_r$
    - 3D contacts between meshes M^a, M^b are represented as a binary contact map $C \in \{0,1\}^{75 \times 75}$

$$\mathcal{C}_{ij}^{B} = \begin{cases} 1, & \text{if } r_i \text{ of } M^a \text{ is in contact with } r_j \text{ of } M^b \\ 0, & \text{otherwise.} \end{cases}$$

  - Two-stage approach
    - First stage: optimize pose, shape and translation
      - Encourages contact between discretely annotated body regions
      - Allow the bodies to intersect

Generative Proxemics

- Optimization process
  - Two stage approach
    - Second stage: use a loss term to resolve human-human intersection
      - The output of the first stage is usually close to the final pose with only slight intersections
      - Optimize only pose and translation
      - Fix the body shape
    - The objective function is:

$$
\begin{aligned}
L_{\text{fitting}} = & \lambda_{J2D}\mathbf{E}_{J2D} + \lambda_{\bar{\theta}}\mathbf{E}_{\bar{\theta}} + \lambda_{\theta}\mathbf{E}_{\theta} + {}_{\theta} + \\
& \lambda_{\boldsymbol{\beta}}\mathbf{E}_{\boldsymbol{\beta}} + \lambda_{\mathcal{C}^B}\mathbf{E}_{\mathcal{C}^B} + \lambda_P\mathbf{E}_P, \quad ,
\end{aligned}
$$

      - $\boldsymbol{E}_{J2D}$(re-projection error), $\boldsymbol{E}_{\theta}$(prior on the initial pose), $\boldsymbol{E}_{\bar{\theta}}$(GMM pose prior), $\boldsymbol{E}_{\beta}$ (L2-prior)
      - In the first and second stage this objective function is used but with different active terms

Generative Proxemics

- Optimization process
  - Objective function

$$L_{\text{fitting}} = \lambda_{J2D}\mathbf{E}_{J2D} + \lambda_{\bar{\theta}}\mathbf{E}_{\bar{\theta}} + \lambda_{\theta}\mathbf{E}_{\theta} + \lambda_{\boldsymbol{\beta}}\mathbf{E}_{\boldsymbol{\beta}} + \lambda_{\mathcal{C}^B}\mathbf{E}_{\mathcal{C}^B} + \lambda_P\mathbf{E}_P,$$

- $\boldsymbol{E}_{J2D}$(re-projection error), $\boldsymbol{E}_{\theta}$(prior on the initial pose), $\boldsymbol{E}_{\bar{\theta}}$(GMM pose prior), $\boldsymbol{E}_{\beta}$ (L2-prior)

– Discrete human-human contact loss

$$\mathbf{E}_{\mathcal{C}^B} = \sum_{i,j} \mathcal{C}_{ij}^B \min_{v \in \mathbf{v}_{r_i}, u \in \mathbf{v}_{r_j}} \|v - u\|^2$$

- Binary contact map works like adjacency matrix

Generative Proxemics

- Optimization process
  - Objective function

$$L_{\text{fitting}} = \lambda_{J2D}\mathbf{E}_{J2D} + \lambda_{\bar{\theta}}\mathbf{E}_{\bar{\theta}} + \lambda_{\theta}\mathbf{E}_{\theta} + \lambda_{\boldsymbol{\beta}}\mathbf{E}_{\boldsymbol{\beta}} + \lambda_{\mathcal{C}^B}\mathbf{E}_{\mathcal{C}^B} + \lambda_P\mathbf{E}_P,$$

- $\boldsymbol{E}_{J2D}$(re-projection error), $\boldsymbol{E}_{\theta}$(prior on the initial pose), $\boldsymbol{E}_{\bar{\theta}}$(GMM pose prior), $\boldsymbol{E}_{\beta}$ (L2-prior)

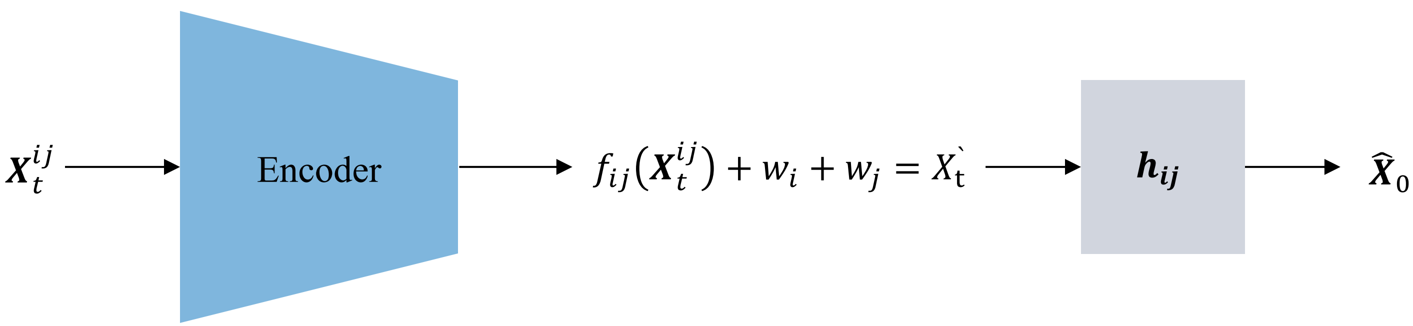
– Interpenetration loss

$$\mathbf{E}_P = \sum_{v \in \mathbf{v}_I^a} \min_{u \in \mathbf{v}^b} \|v - u\|^2 + \sum_{v \in \mathbf{v}_I^b} \min_{u \in \mathbf{v}^a} \|v - u\|^2$$

- Active in the second stage only
- Pushes inside vertices to the surface
- $\boldsymbol{V}_I^a$ denotes vertices of M^a intersecting the low-resolution mesh of M^b
- V_I^b follows the same notation

Generative Proxemics

- Unconditional Generative model train
  - Use Gaussian Diffusion to jointly learn parameter space of two human
    - First diffusion ground truth $\boldsymbol{X}_0 = [X^a, X^b]$, by uniformly sample a noise level t with noise $\epsilon_t \sim N(0, \sigma_t \boldsymbol{I})$
    - Noise ground truth by $\boldsymbol{X}_t = \alpha_t \boldsymbol{X}_0 + \epsilon_t$
    - Denoise with denoiser **D**, with transformer encoder
    - Before passing it to transformer encoder embed model parameters and human identity
      - $i \in \{\phi, \theta, \beta, \gamma\}$: i indicates each parameters where $j \in \{a, b\}$ indicates each person
      - f_{ij} : linear layers that generate learnable embeddings



$$L_D = L_\theta + L_\beta + L_\gamma + L_{v2v}$$

Generative Proxemics

- Optimization with the Proxemics Prior
  - Generating human-human interaction from an image
    - Use the diffusion model as a prior in optimization
      - Congruent to score distillation sampling in DreamFusion and Score Jacobian Chaining

$$L_{\text{Optimization w. BUDDI}} = L_{\text{fitting}} + L_{\text{diffusion}}$$

    - Flows
      - Input image to SMPL-X parameters using BEV
      - Optimize considering contact using $L_{fitting}$ (set λ_{C^B}=0)
        - ✓Initial SMPL-X parameters($\boldsymbol{X}_{no\_grad}$)predicted by BEV does not have gradients
        - ✓Add noise to initial parameters : $\boldsymbol{X}_t = \alpha_t \boldsymbol{X}_{no\_grad} + \epsilon_t$
        - ✓Denoise with denoiser using loss $L_{diffusion}$
          - Only use terms that incorporate parameters we optimize in both stages
          - The shape and orientation loss weights are set to zero

Generative Proxemics

- Evaluation Metrics
  - Use standard evaluation metrics from the human pose estimation literature
    - MPJPE, PA-MPJPE
  - Propose a new metric
    - PCC (Percentage of Correct Contact points)
      - Given two meshes and a contact map, compute pairwise vertex-to-vertex distance
      - Computed on annotated contact regions and consider pair to be correct when the distance is less than threshold

References

- Three-dimensional Reconstruction of Human Interactions (CVPR 2020)
- Generative Proxemics: A prior for 3D social Interactions from Images (arxiv)

감사합니다