Human and Interaction

2024년도 동계 세미나

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

Presented By
ChanHee Kang

Outline

- Background
  - Camera model
    - Weak perspective camera model
    - Perspective camera model
- Paper
  - CHORUS: Learning Canonicalized 3D Human-Object Spatial Relations from Unbounded Synthesized Images

Background

- **Perspective Camera Model (원근 투영 카메라 모델)**

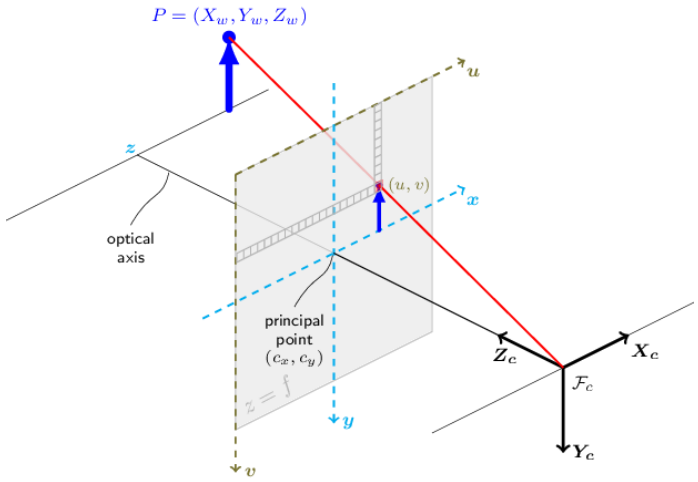
  ▪ 개요

  - 3D 공간의 점들을 2D 이미지 평면으로 투영할 때, 물체와 카메라 사이의 거리에 따라 투영되는 물체의 크기가 달라지는 현상을 반영한 카메라 모델

  - $P' = K[R|t]P$로 표현할 수 있음

    ⁂ P: 3D 공간의 점, K: 카메라 내부 파라미터 행렬, R: rotation matrix, t: translation vector

    ⁂ P': 2D 이미지 평면에 투영된 점

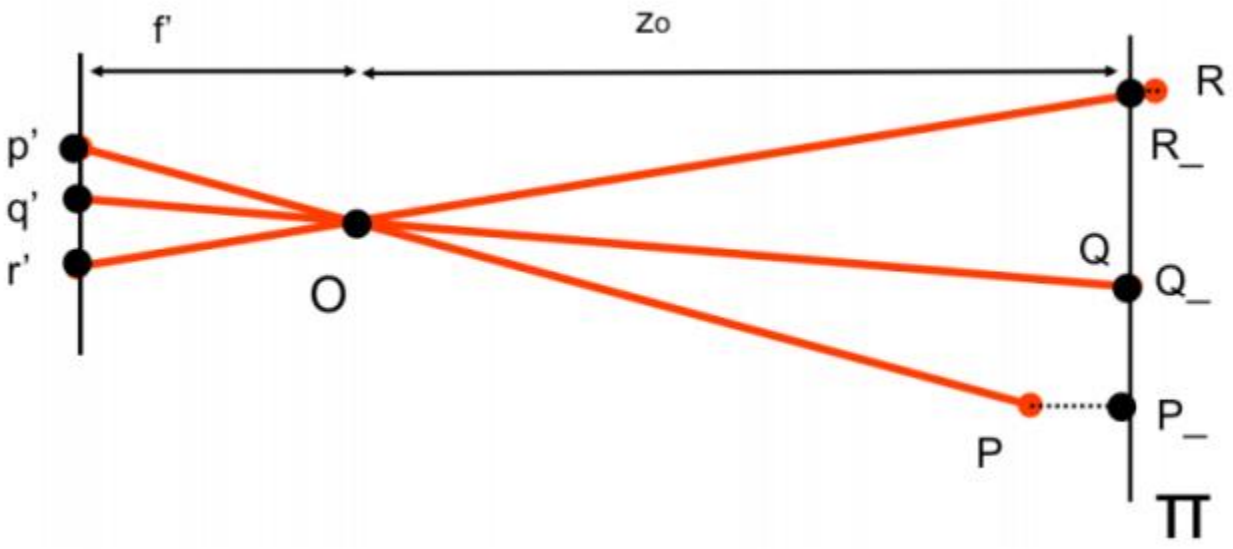


Background

- **Weak Perspective Camera Model (약한 원근 투영 카메라 모델)**

  ▪ 개요

  - Weak perspective camera model은 perspective model을 단순화 하여, 3D 공간의 객체가 카메라에 충분히 멀리 떨어져 있어 모든 점들이 거의 동일한 거리에 있다고 가정하는 카메라 모델
  - 이 모델에서는 객체의 깊이 변화가 이미지 상의 크기 변화에 미치는 영향을 무시할 수 있음
  - 이로 인해, 모든 투영(projection)은 단일 스케일링 요소로 근사할 수 있음



Background

[경고] 다음 내용을 모르고 세미나 내용을 들으면 바보로 느껴질 수 있습니다.

- **Skinning**
  - **Skinning**
    - 뼈대만 있는 물체에 표면을 덮는 것
  - **Rigid Skinning**
    - 각 skin vertex 가 정확히 하나의 뼈(bone)에 연관되어 있다고 가정하는 방법
      - 그렇기 때문에 skin의 각 vertex를 위치시키기 위해 연관된 bone의 transform을 이용
      - $v'_j = T_i v_j$로 deformation이 진행
    - 단점
      - Skin vertices들의 경계에서 원하지 않는 불연속성 발생
  - **Linear Blend Skinning (LBS)**
    - 관절(joint) 근처의 vertices (mesh를 구성하는 점) 을 선형적으로 결합하여 rigid skinning의 단점을 극복
      - $v'_j = \sum_i w_{ij} T_i v_j^i$로 deformation 진행
        - ✓ i : bone의 index, j: vertex의 index

Background

[경고] 다음 내용을 모르고 세미나 내용을 들으면 바보로 느껴질 수 있습니다.

- **Precision, Recall, AP (Average Precision)**
  - **Precision**
    - **Positive라고 한 예측 중, 정답을 positive라고 예측한 것의 비율**
  - **Recall**
    - **실제로 정답인 것 중 모델이 정답이라고 예측한 것의 비율**
  - **Average precision**
    - **Confidence score의 적절한 threshold를 조절함에 따라 precision과 recall 값이 변함**
    - **이렇게 조절하여 recall값의 변화에 따른 precision 값을 나타낸 것을 PR곡선이라고 함**
    - **이렇게 얻어진 precision 값의 평균을 AP라고 하며 PR 곡선의 면적 (AUC, Area Under Curve)로 계산**

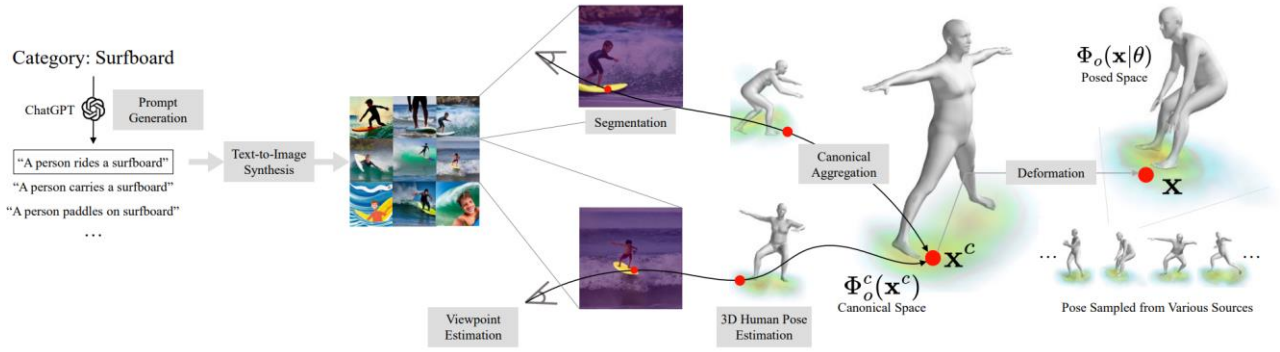
Paper

- **Overview**

  - **목표**

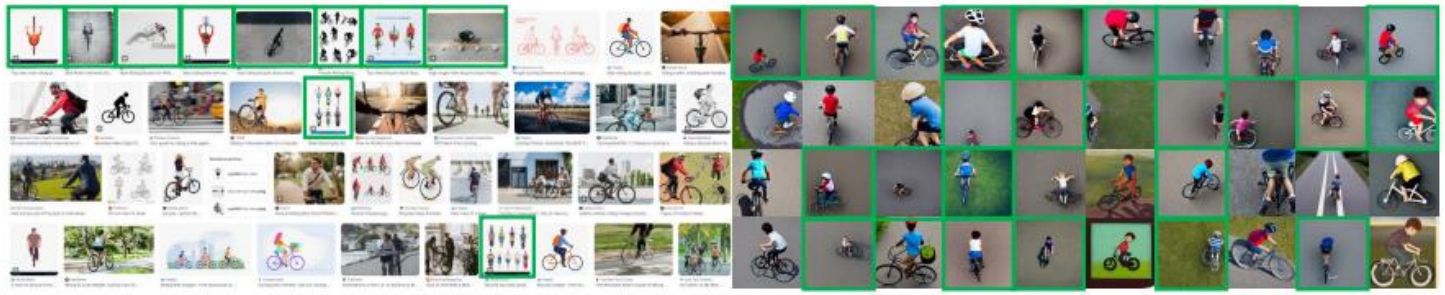
    - **Human Object Interaction (HOI)에 대한 3D spatial knowledge를 사람의 표면 위치에 대한 occupancy probability distribution로 모델링 하는 것**

      - **Occupancy probability distribution : 사람 표면의 좌표가 주어졌을 때, 그 좌표에 object 가 접촉할 확률 분포**
      - $\Phi_o(x|\theta, s) \in [0, 1]$
        - ✓ o: interaction 하는 object, x: HOI가 일어나는 위치, θ: SMPL pose 파라미터, s: HOI type



Paper

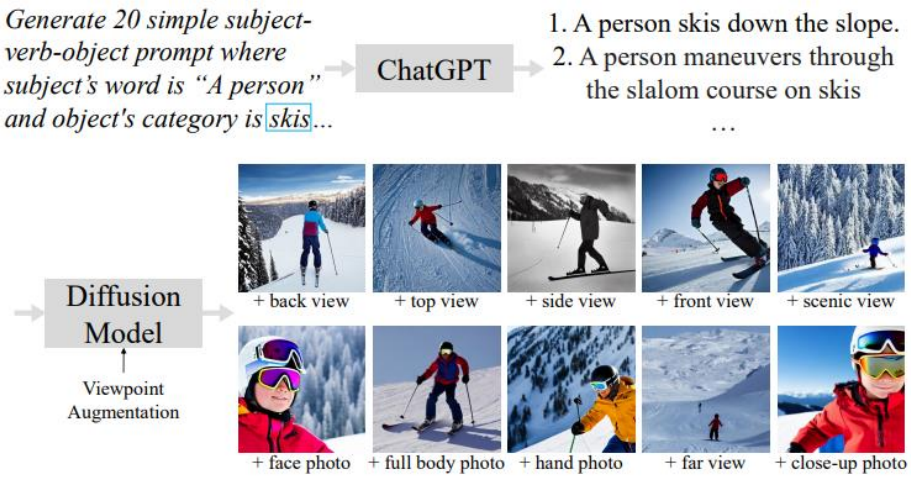
- **ChatGPT를 이용한 prompt 생성과, LDM을 이용한 이미지 합성**
  - **Prompt : “A person is riding a bicycle, top view”**
    - **Internet Image Search V.S. Image Synthesis**
      - 비슷한 대상에 대한 multi-view 이미지를 얻기 원함
      - 합성된 이미지가 ‘top view’라는 viewpoint keyword에 더 부합하는 이미지 생성
      - 인터넷 크롤링을 해서 얻은 이미지보다 실험적으로 합성적 이미지가 더 적합할 것이라 판단
    - **Image filtering**
      - 생성된 이미지를 사람의 존재 유무, 생성된 사람의 수, 생성된 사람의 상체 존재 유무 등의 기준을 통해 필터링



Internet Image Search Results　　Synthesized Images

Paper

- **ChatGPT를 이용한 prompt 생성과, LDM을 이용한 이미지 합성**
  - **Viewpoint augmentation**
    - **ChatGPT로 생성된 prompt 에 viewpoint를 의미하는 keyword를 추가하여 augmentation 진행**
      - **Weakness : viewpoint가 제대로 반영되었는지를 filtering 하는 protocol이 존재 하지 않았음**

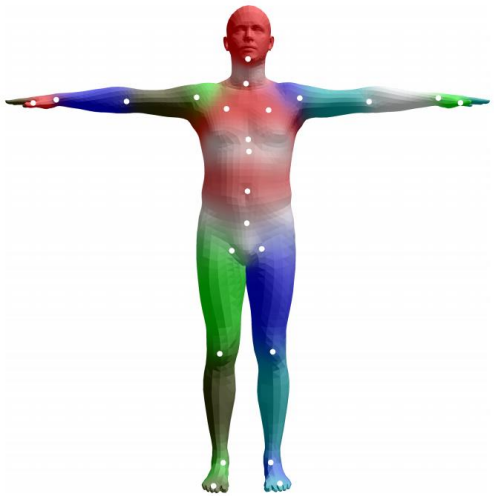


Paper

- 생성된 이미지의 문제점
  - 생성 모델을 통해 생성된 2D 이미지의 inconsistency
    - 생성 모델을 통해서 생성된 이미지의 경우 정확히 동일한 대상에 대한 multi-view 이미지 생성 불가
      - 이는 입력하는 prompt로 control 할 수 있는 것이 아님
      - Prompt로 control할 수 있는 것은 interaction하는 object와 interaction type 뿐임
        - ✓Object to interact with : surfboard, backpack, …
        - ✓Interaction type : hold, ride, …
    - 생성 모델을 통해서 생성된 이미지는 pose를 제어할 수 없음
      - 그렇기 때문에 canonical space에서의 occupancy distribution을 구해야 함
      - 이를 논문에서는 HOI cues들을 canonical space에서 aggregate 한다고 이야기함

Paper

- **Canonical Space**
  - **Human pose 에서의 canonical space**
    - **이 논문에서 이야기하는 canonical space는 SMPL의 rest pose로 정의됨 (48^3 voxel space)**
      - 대부분의 관절에 대해 회전을 0으로 설정
      - 왼쪽과 오른쪽 골반(pelvis)에 각각 z축 회전을 $\pm\frac{\pi}{6}$를 적용
        - ✓실험적으로 다리 사이에 충분한 거리를 유지하는 것이 유리하다는 것을 경험적으로 발견
      - $\Phi_o^c(x^c|s) \in [0,1]$를 우선 찾고 이를 pose deformed space로 warping
        - ✓Φ_o^c : canonical space에서 특정 object와 interaction할 확률 분포
        - ✓x^c : canonical space 상에서의 vertex 좌표
        - ✓s : HOI type (e.g. hold, ride, …)



Paper

- **Pose Canonicalization**

  - **방법론**

    - 3D human pose estimation model (top-down)을 이용해 camera viewpoint를 추정
    - 생성된 2D object를 semantic segmentation 모델을 이용해 object와 human에 대한 mask를 구함
    - 이를 이용하여 3D occupancy distribution을 추정

  - **Viewpoint estimation via 3D Human Pose Estimation**

    - Off-the-shelf 3D human pose estimator 사용 (top-down approach)
      - Input : I (image), B (person bounding box)
      - Output : ϕ (global orientation), θ (pose parameter=angles), β (shape parameter), π (weak perspective camera parameter), j (image space에서의 projected 2D joint location)
    - Camera parameter conversion
      - ϕ, π의 경우 perspective camera parameter로 변환됨
        - ✓WPC parameter를 이용한 projection과 PC parameter를 통해 구한 projection이 일치하도록 최적화

Paper

- **3D Occupancy Estimation via Human Pose Canonicalization**
  - **Canonical space의 Linear Blend Skinning (LBS) weight 계산**

관절 index는 고정된 상태
or
Vectorized weight

$$\omega(\mathbf{x}^c) = \frac{\sum_{i \in \mathbf{N_k}(\mathbf{x}^c)} w_i \,/\, \|\mathbf{x}^c - \mathbf{v}_i\|}{\sum_{i \in \mathbf{N_k}(\mathbf{x}^c)} 1 \,/\, \|\mathbf{x}^c - \mathbf{v}_i\|}$$

- **SMPL 메시의 k 개의 가장 가까운 이웃 꼭짓점에서의 가중치를 이용**
  - 직관적으로 3D 공간은 가장 가까운 SMPL 꼭짓점들의 움직임을 따라 변형됨
  - Canonical space의 특정 위치의 메시가, 인접한 mesh의 꼭짓점의 영향을 더 받게끔 설계
  - 기존의 SMPL에서 사전 정의된 LBS weight (w_i)를 조금 수정함

Paper

- **3D Occupancy Estimation via Human Pose Canonicalization**
  - **Canonical space to pose deformed space warping**

$$\mathbf{x} = \mathcal{W}(\mathbf{x}^c) = \sum_{j=1}^{n_b} \omega_j(\mathbf{x}^c) \cdot \mathbf{B}_j(\theta_j) \cdot \mathbf{x}^c$$

- j : joint의 index (관절)
- $B_j(\theta_j) \in SE(3)$, w_j: j-번째 관절에 해당하는 LBS weight

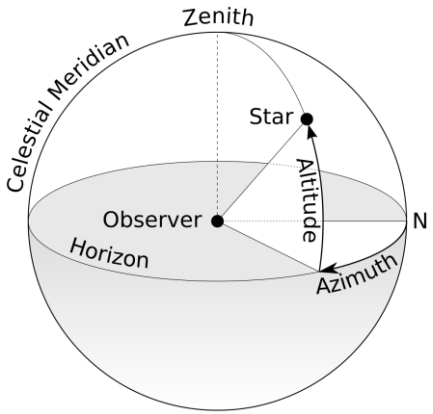
  - **Occupancy distribution computation**

$$\Phi_o^c(\mathbf{x^c}) = \frac{\sum_{k=1}^{|\mathbf{G}|} r_k \mathcal{M}_k(\Pi_k(\mathcal{W}(\mathbf{x^c})))}{\sum_{k=1}^{|\mathbf{G}|} r_k \mathcal{I}_k(\Pi_k(\mathcal{W}(\mathbf{x^c})))}$$

- G: a set of generated images, r_k : accumulation score
- $\mathcal{M}_k$: mask operator, $\mathcal{I}_k$: image operator, Π_k: perspective camera parameter

Paper

- **3D Occupancy Estimation via Human Pose Canonicalization**
  - **Uniform View Sampling**
    - **Viewpoint augmentation을 했음에도 불구하고 특정 viewpoint의 편향이 존재**
      - **Azimuth (방위각)을 고정된 숫자의 bin으로 나눔**
      - **각 bin에 해당하는 카메라 숫자의 역수를 accumulation score r_k로 지정**
  - **Inference for posed space**
    - **Inference 시에는 canonical space에서의 occupancy distribution을 posed deformed space로 변형**
      - **Backward skinning 이용 (train 시에는 forward skinning 이용)**



Paper

- **Experiments**
  - **Dataset**
    - **Image search dataset**
      - **AutoCrawler를 이용하여 이미지 크롤링된 데이터 셋 사용**
    - **COCO-EFT dataset (for testing)**
      - **GT-2D object mask 존재**
      - **Pseudo 3D GT (SMPL) 존재**

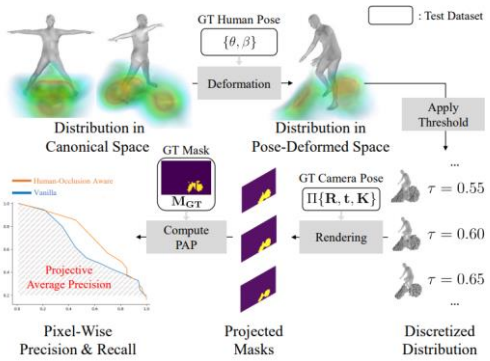
Paper

- **Evaluation Protocol**

  - **Projective Average Precision (PAP)**

    - **3D annotation 없이 occupancy distribution의 타당성을 평가하기 위한 지표**

      - 분포 변형 : 분류 키워드에 따라 canonical space에서 pose deformed space로 분포 변형
      - 분포 이산화 : pose deformed space에서 분포를 다양한 threshold에 대해 이산화 후, perspective camera를 이용하여 projection 시킴
      - Precision and Recall 계산 : Threshold에 대해 렌더링된 마스크와 annotation mask의 pixelwise precision 과 recall 계산
      - AP 계산 : Precision과 Recall 값을 계산하여 AP 값을 계산하고, 이를 모든 카테고리 내 , 모든 테스트 이미지에 대해 평균하여 PAP 값 계산



감사합니다.