15 하이브리드 지능형 가상 캐릭터 동작 생성 개발

소속 정보컴퓨터공학부 분과 A 팀명 HCB 참여학생 박현성, 최우창, 황원식 지도교수 이명호

서론

과제 개요

✓ 하이브리드 지능형 가상 캐릭터의 한글 텍스트 기반 동작 생성을 개발하는 것을 목표로 둔다.

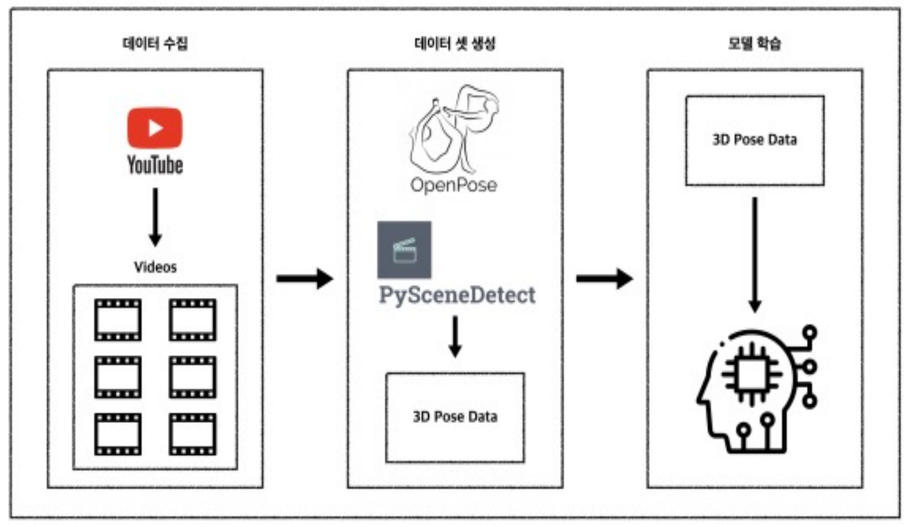
✓ 기존의 생성 모델 중 한글 텍스트 기반 동작 생성 모델이 존재하지 않았기에 한글 텍스트 데이터 셋 생성과 모델 구성을 수행한다.

과제 목표

✓ 한글 텍스트 데이터 셋 구축을 위해 유튜브 채널을 통해 영상과 자막을 얻는다.

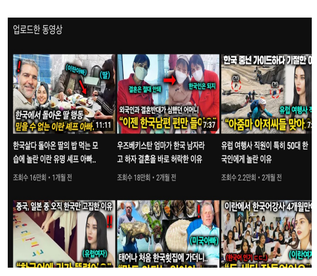
✓ OpenPose 라이브러리를 이용하여 2차원 Pose 추출 후 3차원 Pose 추정 작업을 진행한다.

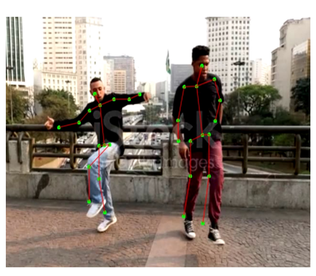
✓ 해당 데이터 셋을 모델에 학습시켜 한글 텍스트에 대한 제스처를 생성한다.

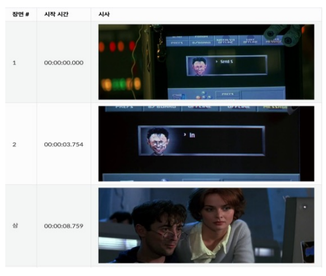


본론

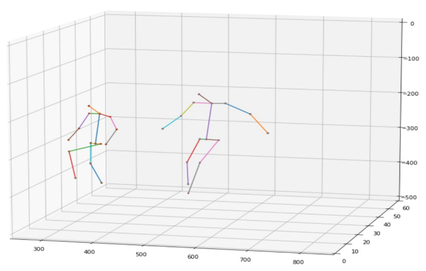
데이터 셋 생성





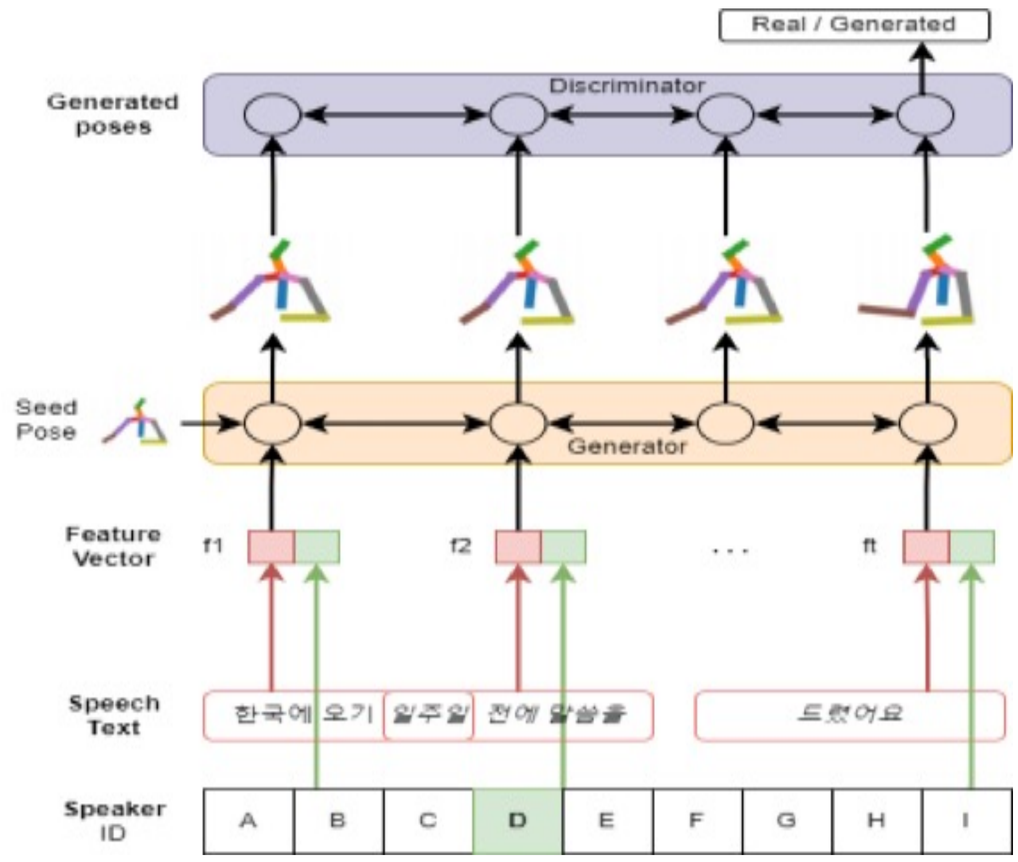


유튜브 영상 및 자막 웹크롤링 → OpenPose를 이용해 2차원 Human Pose 추출 → PySceneDetect를 이용해 영상을 클립으로 나눈 뒤 유효한 클립 선별 → 유효한 클립의 2차원 Pose를 3차원 Pose로 추정 → 추출된 데이터를 통해 LMDB 데이터 셋 생성



사용한 비디오 수	328개
비디오 평균 길이	6.4분
전체 비디오 프레임 수	3857343 frames
사용한 프레임 비율	28% (1094727/3857343)
사용 비디오 길이	10.03시간

모델 구조



✓ Text는 FastText로 워드 임베딩되어 TCN Network를 거쳐 인코딩된다.

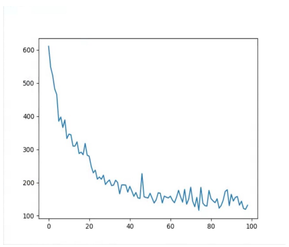
✓ Speaker ID는 Style Embedding Space를 학습시킬 때 사용된다.

✓ Style Embedding Space는 같은 문장에서 다른 동작을 생성한다.

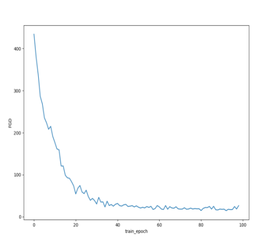
✓ Gesture Generator는 GRU Network로 구성되고 GAN로 학습된다.

결론

FGD 비교



Sequence2Sequence
(Min FGD : 116.269)

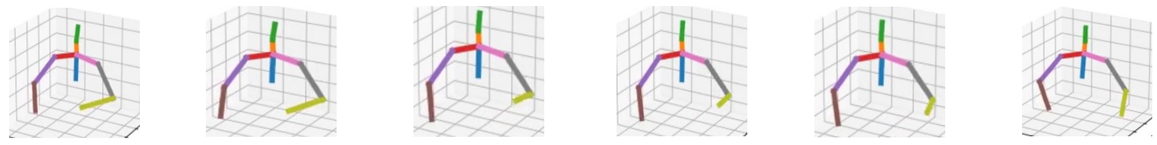


Multimodal Context
(Min FGD: 14.637)

❑ **FGD**

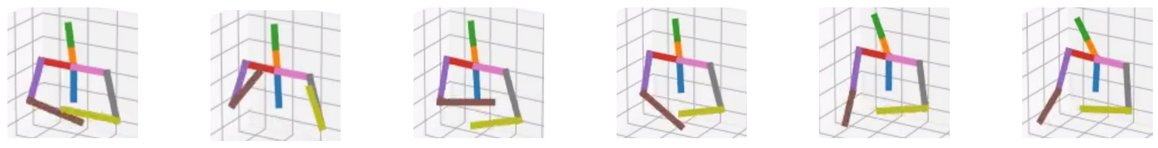
GAN 모델을 평가할 때 자주 사용되는 FID(Frechet Inception Distance)라는 생성된 이미지 분포와 기존 이미지 분포가 얼마나 유사한지 측정하는 지표를 Gesture Generation Problem에 적용한 값

기존 모델과 결과 비교



H e l l o, N i c e t o m e e t y o u

[기존 모델]



안 녕 하 세 요 만 나 서 반 갑 습 니 다.

[개발 모델]

참고 문헌

[1] Y. Yoon, B. Cha, J. Lee, M. Jang, J. Lee, J. Kim, G. Lee, “Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity”, Journal of ACM: Transaction on Graphics, Vol. 39, No. 6, pp 1-16, Dec. 2020.(in Korean)

[2] R. Zhao, Y. Wang, A. Martinez, “A Simple, Fast and Highly-Accurate Algorithm to Recover 3D Shape from 2D Landmarks on a Single Image”, Journal of IEEE: Transactions on Pattern Analysis and Machine Intelligence, Vol. 40, No. 12, Dec. 2018. (3D Estimate)

[3] Github, openpose: https://github.com/CMU-Perceptual-Computing-Lab/openpose

[4] Github, PySceneDetect: https://github.com/Breakthrough/PySceneDetect