(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0126389
(43) 공개일자 2024년08월20일

(51) 국제특허분류(Int. Cl.)
G06T 13/40 (2011.01) *G06F 40/279* (2020.01)
G06F 40/40 (2020.01) *G06T 15/04* (2011.01)
G06T 17/20 (2006.01)
(52) CPC특허분류
G06T 13/40 (2013.01)
G06F 40/279 (2020.01)
(21) 출원번호 10-2023-0139071
(22) 출원일자 2023년10월17일
심사청구일자 2023년10월17일
(30) 우선권주장
1020230018505 2023년02월13일 대한민국(KR)

(71) 출원인
포항공과대학교 산학협력단
경상북도 포항시 남구 청암로 77 (지곡동)
(72) 발명자
오태현
경상북도 포항시 남구 청암로 77
김유왕
경상북도 포항시 남구 청암로 77
김지연
경상북도 포항시 남구 청암로 77
(74) 대리인
특허법인이상

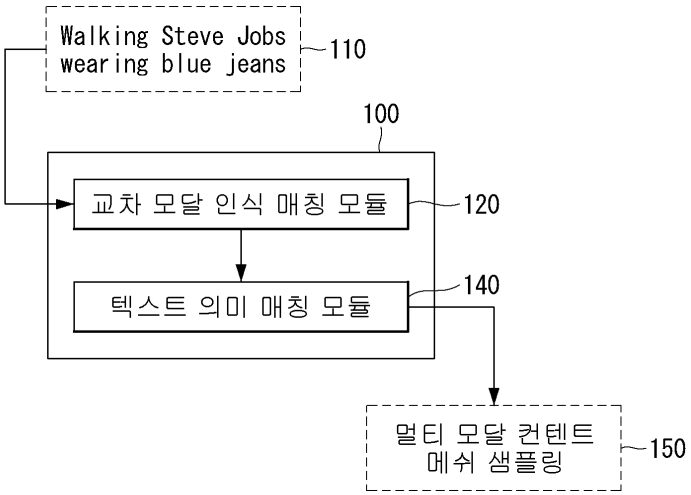
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **텍스트 기반 모션 추천을 이용하는 모션 애니메이팅 및 메쉬 스타일라이제이션 방법 및 장치**

(57) 요 약

텍스트 입력에 기초하여 계층적 다중 모델 모션 검색을 통해 모션을 추천하고, 추천된 멀티모달 콘텐츠 메쉬 샘플링에 스타일라이제이션을 더함으로써 텍스트의 의미에 가장 적합한 동작 시퀀스의 4차원 아바타를 생성하는 모션 애니메이팅 및 메쉬 스타일라이제이션 방법 및 장치가 개시된다. 이 방법은, 데이터베이스에 저장된 휴먼 모션 데이터세트에서 텍스트 프롬프트로 주어지는 쿼리에 일치하는 원시 액션 레이블을 찾는 단계, 원시 액션 레이블과 쿼리를 인코딩하여 벡터화하는 단계, 및 벡터화된 벡터들 간의 유사도를 계산하는 단계를 포함한다.

대 표 도 - 도3



(52) CPC특허분류
G06F 40/40 (2020.01)
G06T 15/04 (2013.01)
G06T 17/20 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711174048
과제번호 00164860
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 ICT융합산업혁신기술개발(R&D)
연구과제명 동적 객체 행동 모델링 및 휴먼-사물-공간 상호작용 기반 휴먼 디지털 트윈 기술 개발
기 여 율 1/3
과제수행기관명 한국전자기술연구원
연구기간 2022.06.01 ~ 2023.02.28
이 발명을 지원한 국가연구개발사업
과제고유번호 1711152863
과제번호 2021-0-02068-002
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성(R&D)
연구과제명 인공지능 혁신 허브 연구 개발
기 여 율 1/3
과제수행기관명 고려대학교 산학협력단
연구기간 2022.01.01 ~ 2022.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711195728
과제번호 00225630
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 실감콘텐츠핵심기술개발(R&D)
연구과제명 문장으로부터의 3차원 동영상 자동 생성 기술
기 여 율 1/3
과제수행기관명 한국과학기술연구원
연구기간 2023.04.01 ~ 2023.12.31

명 세 서

청구범위

청구항 1

컴퓨팅 장치에 의해 수행되는 텍스트 기반 모션 애니메이팅 및 메쉬 스타일라이제이션 방법으로서,

데이터베이스에 저장된 휴먼 모션 데이터세트에서 텍스트 프롬프트로 주어지는 쿼리에 일치하는 원시 액션 레이블을 찾는 단계;

상기 원시 액션 레이블과 상기 쿼리를 인코딩하여 벡터화하는 단계; 및

상기 벡터화된 벡터들 간의 유사도를 계산하는 단계;

를 포함하는 모션 애니메이팅 및 메쉬 스타일라이제이션 방법.

청구항 2

청구항 1에 있어서,

상기 계산하는 단계에서 계산된 유사도 값들 중 복수의 상위 값들을 색인하는 단계를 더 포함하는 모션 애니메이팅 및 메쉬 스타일라이제이션 방법.

청구항 3

청구항 2에 있어서,

상기 복수의 상위 값들에 기초하여 상기 휴먼 모션 데이터세트에서 찾은 액션 레이블과 상기 쿼리를 언어 모델에 입력하여 벡터화하는 단계를 더 포함하는 모션 애니메이팅 및 메쉬 스타일라이제이션 방법.

청구항 4

청구항 3에 있어서,

상기 복수의 상위 값들에 대한 벡터와 상기 쿼리에 대한 벡터 간의 유사도를 계산하여 상기 쿼리에 가장 잘 따르는 액션 레이블을 출력하는 단계를 더 포함하는, 모션 애니메이팅 및 메쉬 스타일라이제이션 방법.

청구항 5

청구항 4에 있어서,

상기 액션 레이블에 기초한 멀티모달 콘텐츠 메쉬 샘플링에 스타일 속성을 부여하여 움직이는 휴먼 메쉬 시퀀스를 생성하는 단계를 더 포함하는, 모션 애니메이팅 및 메쉬 스타일라이제이션 방법.

청구항 6

청구항 5에 있어서,

상기 스타일 속성을 예측하는 단계를 더 포함하며, 상기 스타일 속성은 색상과 변위를 포함하는, 모션 애니메이팅 및 메쉬 스타일라이제이션 방법.

청구항 7

청구항 6에 있어서,

상기 예측하는 단계 전에, 외부 데이터베이스로부터 획득한 템플릿 휴먼 메쉬로부터 상기 스타일 속성을 매핑하는 단계를 더 포함하는, 모션 애니메이팅 및 메쉬 스타일라이제이션 방법.

청구항 8

청구항 7에 있어서,

상기 매핑하는 단계는, 상기 템플릿 휴먼 메쉬 중 기본 뉴럴 스타일 필드와 동일한 메쉬 스타일을 찾는 동시에 콘텐츠 메쉬에서 스타일을 효과적으로 분리하는, 모션 애니메이팅 및 메쉬 스타일라이제이션 방법.

청구항 9

청구항 7에 있어서,

상기 매핑하는 단계는, 상기 템플릿 휴먼 메쉬의 꼭짓점 또는 메쉬 정점을 포즈에 상관없이 색상 스타일 속성 및 변위 스타일 속성에 매핑하는, 모션 애니메이팅 및 메쉬 스타일라이제이션 방법.

청구항 10

청구항 7에 있어서,

상기 멀티모달 콘텐츠 메쉬 샘플링에 상기 스타일 속성을 부여하는 단계를 더 포함하는, 모션 애니메이팅 및 메쉬 스타일라이제이션 방법.

청구항 11

청구항 10에 있어서,

상기 스타일 속성을 부여된 아바타에 대하여 시공간 뷰 증강을 수행하는 단계를 더 포함하는, 모션 애니메이팅 및 메쉬 스타일라이제이션 방법.

청구항 12

청구항 11에 있어서,

상기 시공간 뷰 증강을 수행하는 단계 후에 상기 아바타에 대응하는 멀티모달 콘텐츠 메쉬 샘플링의 각 이미지의 전경 픽셀 비율에 따라 복수의 서로 다른 카메라 포즈들에서 임베딩 벡터에 가중치를 부여하는 단계를 더 포함하는, 모션 애니메이팅 및 메쉬 스타일라이제이션 방법.

청구항 13

청구항 12에 있어서,

상기 가중치가 부여된 렌더링된 이미지를 사전 훈련된 인코더로 인코딩하는 단계를 더 포함하는, 모션 애니메이팅 및 메쉬 스타일라이제이션 방법.

청구항 14

텍스트 기반 휴먼 모션 추천을 이용하는 모션 애니메이팅 및 메쉬 스타일라이제이션 장치로서,

적어도 하나의 명령을 저장하는 메모리에 연결되는 프로세서를 포함하며,

상기 적어도 하나의 명령에 의해, 상기 프로세서는,

데이터베이스에 저장된 휴먼 모션 데이터세트에서 텍스트 프롬프트로 주어지는 쿼리에 일치하는 원시 액션 레이블을 찾는 단계;

상기 원시 액션 레이블과 상기 쿼리를 인코딩하여 벡터화하는 단계; 및

상기 벡터화된 벡터들 간의 유사도를 계산하는 단계;를 수행하는,

모션 애니메이팅 및 메쉬 스타일라이제이션 장치.

청구항 15

청구항 14에 있어서,

상기 프로세서는, 상기 계산하는 단계에서 계산된 유사도 값들 중 복수의 상위 값들을 색인하는 단계를 더 수행하는, 모션 애니메이팅 및 메쉬 스타일라이제이션 장치.

청구항 16

청구항 15에 있어서,

상기 프로세서는, 상기 복수의 상위 값들에 기초하여 상기 휴먼 모션 데이터세트에서 찾은 액션 레이블과 상기 쿼리를 언어 모델에 입력하여 벡터화하는 단계를 더 수행하는, 모션 애니메이팅 및 메쉬 스타일라이제이션 장치.

청구항 17

청구항 16에 있어서,

상기 프로세서는, 상기 복수의 상위 값들에 대한 벡터와 상기 쿼리에 대한 벡터 간의 유사도를 계산하여 상기 쿼리에 가장 잘 따르는 액션 레이블을 출력하는 단계를 더 수행하는, 모션 애니메이팅 및 메쉬 스타일라이제이션 장치.

청구항 18

청구항 17에 있어서,

상기 프로세서는, 상기 액션 레이블에 기초한 멀티모달 콘텐츠 메쉬 샘플링에 스타일 속성을 부여하여 움직이는 휴먼 메쉬 시퀀스를 생성하는 단계를 더 수행하는, 모션 애니메이팅 및 메쉬 스타일라이제이션 장치.

청구항 19

청구항 18에 있어서,

상기 프로세서는, 상기 스타일 속성을 예측하는 단계를 더 수행하며, 상기 스타일 속성은 색상과 변위를 포함하는, 모션 애니메이팅 및 메쉬 스타일라이제이션 장치.

청구항 20

청구항 19에 있어서,

상기 프로세서는, 상기 예측하는 단계 전에, 외부 데이터베이스로부터 획득한 템플릿 휴먼 메쉬로부터 상기 스타일 속성을 매핑하는 단계를 더 수행하는, 모션 애니메이팅 및 메쉬 스타일라이제이션 장치.

발명의 설명

기 술 분 야

[0001] 본 개시는 입력 텍스트로부터 움직이는 아바타를 출력하는 기술에 관한 것으로, 보다 상세하게는, 텍스트 입력에 기초하여 계층적 다중 모델 모션 검색을 통해 휴먼 모션을 추천하고, 추천된 멀티모달 콘텐츠 메쉬 샘플링에 스타일라이제이션을 더함으로써 텍스트의 의미에 가장 적합한 동작 시퀀스의 4차원 휴먼 아바타를 생성하는 모션 애니메이팅 및 메쉬 스타일라이제이션 기술에 관한 것이다.

배 경 기 술

[0002] 움직일 수 있고 디테일한 3차원 아바타를 수동으로 생산하는 것이나 휴먼 모델링 및 에디팅 파이프라인을 수동으로 생성하는 것은 노동 집약적이고 제작의 고통을 수반하는 번거롭고 지루하며 시간이 많이 걸리는 작업이다. 이러한 부담을 줄이기 위해 해당 프로세스를 자동화하려는 많은 시도가 도입되고 있다.

[0003] 게다가, 변형이 심한 인체는 시간적으로 일관된 세부 기하학적 구조와 질감을 디자인하는 것을 더욱 어렵게 만든다.

[0004] 이와 같이 자연어 프롬프트만으로 디테일한 기하학적 구조와 질감으로 그럴듯하고 인간이 인식할 수 있는 스타일의 3D 휴먼 메쉬를 생성하는 새로운 방안이 요구되고 있다.

발명의 내용

해결하려는 과제

[0005] 본 개시는 전술한 요구에 부응하기 위해 도출된 것으로, 본 개시의 목적은 입력 텍스트로부터 움직이는 아바타

를 생성하기 위해, 텍스트 프롬프트에 기초하여 계층적 멀티모달 모션 검색을 통해 휴먼 모션을 추천하고, 추천된 휴먼 모션의 멀티모달 콘텐츠 메쉬 샘플링에 스타일라이제이션을 더함으로써 입력 텍스트의 의미체계에 가장 적합한 모션 시퀀스를 가진 4차원 휴먼 아바타를 생성할 수 있는 액션 아바타 생성 장치 및 방법을 제공하는데 있다.

[0006] 본 개시의 다른 목적은 전술한 액션 아바타 생성 장치에 채용할 수 있고 계층적 멀티모달 모션 검색을 기반으로 하는 텍스트 기반 휴먼 모션 추천 방법 및 장치를 제공하는데 있다.

과제의 해결 수단

[0007] 상기 기술적 과제를 해결하기 위한 본 개시의 일 측면에 따른 모션 애니메이팅 및 메쉬 스타일라이제이션 방법은, 컴퓨팅 장치에 의해 수행되고 텍스트 기반 모션 추천을 이용하는 방법으로서, 데이터베이스에 저장된 휴먼 모션 데이터세트에서 텍스트 프롬프트로 주어지는 쿼리에 일치하는 원시 액션 레이블을 찾는 단계; 상기 원시 액션 레이블과 상기 쿼리를 인코딩하여 벡터화하는 단계; 및 상기 벡터화된 벡터들 간의 유사도를 계산하는 단계를 포함한다.

[0008] 상기 방법은, 상기 계산하는 단계에서 계산된 유사도 값들 중 복수의 상위 값들을 색인하는 단계를 더 포함할 수 있다.

[0009] 상기 방법은, 상기 복수의 상위 값들에 기초하여 상기 휴먼 모션 데이터세트에서 찾은 액션 레이블과 상기 쿼리를 언어 모델에 입력하여 벡터화하는 단계를 더 포함할 수 있다.

[0010] 상기 방법은, 상기 복수의 상위 값들에 대한 벡터와 상기 쿼리에 대한 벡터 간의 유사도를 계산하여 상기 쿼리에 가장 잘 따르는 액션 레이블을 출력하는 단계를 더 포함할 수 있다.

[0011] 상기 방법은, 상기 액션 레이블에 기초한 멀티모달 콘텐츠 메쉬 샘플링에 스타일 속성을 부여하여 움직이는 휴먼 메쉬 시퀀스를 생성하는 단계를 더 포함할 수 있다.

[0012] 상기 방법은, 상기 스타일 속성을 예측하는 단계를 더 포함할 수 있다. 여기서, 상기 스타일 속성은 색상과 변위를 포함한다.

[0013] 상기 방법은, 상기 예측하는 단계 전에, 외부 데이터베이스로부터 획득한 템플릿 휴먼 메쉬로부터 상기 스타일 속성을 매핑하는 단계를 더 포함할 수 있다.

[0014] 상기 매핑하는 단계는, 상기 템플릿 휴먼 메쉬 중 기본 뉴럴 스타일 필드와 동일한 메쉬 스타일을 찾는 동시에 콘텐츠 메쉬에서 스타일을 효과적으로 분리하도록 구성될 수 있다.

[0015] 상기 매핑하는 단계는, 상기 템플릿 휴먼 메쉬의 꼭짓점 또는 메쉬 정점을 포즈에 상관없이 색상 스타일 속성 및 변위 스타일 속성에 매핑하도록 구성될 수 있다.

[0016] 상기 방법은, 상기 멀티모달 콘텐츠 메쉬 샘플링에 상기 스타일 속성을 부여하는 단계를 더 포함할 수 있다.

[0017] 상기 방법은, 상기 스타일 속성을 부여된 아바타에 대하여 시공간 뷰 증강을 수행하는 단계를 더 포함할 수 있다.

[0018] 상기 방법은, 상기 시공간 뷰 증강을 수행하는 단계 후에 상기 아바타에 대응하는 멀티모달 콘텐츠 메쉬 샘플링의 각 이미지의 전경 픽셀 비율에 따라 복수의 서로 다른 카메라 포즈들에서 임베딩 벡터에 가중치를 부여하는 단계를 더 포함할 수 있다.

[0019] 상기 방법은, 상기 가중치가 부여된 렌더링된 이미지를 사전 훈련된 인코더로 인코딩하는 단계를 더 포함할 수 있다.

[0020] 상기 기술적 과제를 해결하기 위한 본 개시의 다른 측면에 따른 모션 애니메이팅 및 메쉬 스타일라이제이션 장치는, 텍스트 기반 휴먼 모션 추천을 이용하는 장치로서, 적어도 하나의 명령을 저장하는 메모리에 연결되는 프로세서를 포함한다. 여기서, 상기 적어도 하나의 명령에 의해, 상기 프로세서는, 데이터베이스에 저장된 휴먼 모션 데이터세트에서 텍스트 프롬프트로 주어지는 쿼리에 일치하는 원시 액션 레이블을 찾는 단계; 상기 원시 액션 레이블과 상기 쿼리를 인코딩하여 벡터화하는 단계; 및 상기 벡터화된 벡터들 간의 유사도를 계산하는 단계를 수행한다.

[0021] 상기 프로세서는, 상기 계산하는 단계에서 계산된 유사도 값들 중 복수의 상위 값들을 색인하는 단계를 더 수행

할 수 있다.

[0022] 상기 프로세서는, 상기 복수의 상위 값들에 기초하여 상기 휴먼 모션 데이터세트에서 찾은 액션 레이블과 상기 쿼리를 언어 모델에 입력하여 벡터화하는 단계를 더 수행할 수 있다.

[0023] 상기 프로세서는, 상기 복수의 상위 값들에 대한 벡터와 상기 쿼리에 대한 벡터 간의 유사도를 계산하여 상기 쿼리에 가장 잘 따르는 액션 레이블을 출력하는 단계를 더 수행할 수 있다.

[0024] 상기 프로세서는, 상기 액션 레이블에 기초한 멀티모달 콘텐츠 메쉬 샘플링에 스타일 속성을 부여하여 움직이는 휴먼 메쉬 시퀀스를 생성하는 단계를 더 수행할 수 있다.

[0025] 상기 프로세서는, 상기 스타일 속성을 예측하는 단계를 더 수행할 수 있다. 여기서, 상기 스타일 속성은 색상과 변위를 포함한다.

[0026] 상기 프로세서는, 상기 예측하는 단계 전에, 외부 데이터베이스로부터 획득한 템플릿 휴먼 메쉬로부터 상기 스타일 속성을 매핑하는 단계를 더 수행할 수 있다.

발명의 효과

[0027] 본 개시에 의하면, 텍스트 구동 3D 아바타 생성, 즉 기계가 휴먼의 텍스트 프롬프트를 이해하게 하여 텍스트 프롬프트에 따르는 움직이는 3D 아바타를 생동감있게 완전 자동으로 생성할 수 있다. 텍스트 구동 3D 아바타 생성은 가상 휴먼 애니메이션(virtual human animation), 언어 구동 로봇 작업 계획(language-driven robot task planning), 영화 스크립트 시각화(movie script visualization) 등과 같은 기계 제작 매체(machine-created media)에 널리 효과적으로 적용될 수 있다.

도면의 간단한 설명

[0028] 도 1은 본 개시의 일실시예에 따른 텍스트 기반 모션 애니메이팅 및 메쉬 스타일라이제이션 장치(이하 간략히 '액션 아바타 생성 장치'라고도 한다)에 대한 개략적인 블록도이다.

도 2는 도 1의 액션 아바타 생성 장치에서 주어진 입력 텍스트 프롬프트(input text prompt)에 대하여 생성되는 움직이는 휴먼 아바타를 나타낸 예시도이다.

도 3은 도 1의 액션 아바타 생성 장치의 휴먼 모션 추천(human motion recommendation, HMR) 유닛에 대한 개략적인 블록도이다.

도 4는 도 3의 휴먼 모션 추천 유닛에 채용할 수 있는 계층적 멀티모달 모션 검색 모듈을 설명하기 위한 구성도이다.

도 5는 도 1의 액션 아바타 생성 장치의 최적화 유닛을 설명하기 위한 구성도이다.

도 6은 본 개시의 다른 실시예에 따른 액션 아바타 생성 장치에 대한 개략적인 블록도이다.

도 7은 도 6의 액션 아바타 생성 장치의 휴먼 모션 추천 유닛에 의해 생성되는 멀티모달 콘텐트 메쉬 샘플링을 나타내는 예시도이다.

도 8은 도 6의 액션 아바타 생성 장치의 비교예에서 부주의한 무작위 자르기의 문제를 예시하여 설명하기 위한 도면이다.

도 9는 본 실시예의 액션 아바타 생성 장치에 의해 생성된 움직이는 휴먼 아바타의 추천 모션 시퀀스에서 입력 텍스트 프롬프트와 함께 디테일 표면 지오메트리와 텍스처를 가진 대표 프레임들을 보여주기 위한 도면이다.

도 10은 본 실시예의 액션 아바타 생성 장치의 두 양태들의 성능 실험 결과와 함께 두 비교예들의 성능 실험 결과를 대비하여 설명하기 위한 도면이다.

도 11은 본 실시예의 액션 아바타 생성 장치와 두 비교예들의 모션-텍스트 일관성, 스타일화 품질 및 전반적인 일관성에 대한 성능 평가를 위해 사전에 주어진 5개의 랜덤 텍스트-아바타 쌍의 결과들에 대한 46명의 비전문가 평가를 점수화하여 나타낸 그래프이다.

도 12는 본 실시예의 액션 아바타 생성 장치에 의해 생성된 특정 텍스트 기반의 움직이는 휴먼 아바타에서 가중치 배제, 샘플링 배제, 시간 증강 배제, 공간 증강 배제 및 시공간 증강 배제의 경우들을 비교하여 나타낸 도면

이다.

도 13은 본 개시의 또 다른 실시예에 따른 액션 아바타 생성 장치에 대한 개략적인 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0029] 본 개시는 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세하게 설명하고자 한다. 그러나, 이는 본 개시를 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 개시의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0030] 제1, 제2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 개시의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

[0031] 본 출원의 실시예들에서, "A 및 B 중에서 적어도 하나"는 "A 또는 B 중에서 적어도 하나" 또는 "A 및 B 중 하나 이상의 조합들 중에서 적어도 하나"를 의미할 수 있다. 또한, 본 출원의 실시예들에서, "A 및 B 중에서 하나 이상"은 "A 또는 B 중에서 하나 이상" 또는 "A 및 B 중 하나 이상의 조합들 중에서 하나 이상"을 의미할 수 있다.

[0032] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.

[0033] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 개시를 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0034] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 개시가 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가진 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0035] 이하, 첨부한 도면들을 참조하여, 본 개시의 바람직한 실시예를 보다 상세하게 설명하고자 한다. 본 개시를 설명함에 있어 전체적인 이해를 용이하게 하기 위하여 도면상의 동일한 구성요소에 대해서는 동일한 참조부호를 사용하고 동일한 구성요소에 대해서 중복된 설명은 생략한다.

[0036] 본 개시는 텍스트 구동 3D 객체 콘텐츠 및 스타일 조작과 밀접한 관련이 있다. 멀티모달 객체 스타일화는 주로 CLIP(contrastive language-image pre-training) 및 3D 콘텐츠/스타일 조작 방법과 같은 학습된 멀티모달 임베딩 공간을 사용하여 연구될 수 있다. 이러한 연구 방향을 간략하게 살핀다.

[0037] 여기서, CLIP은 더 나은 일반화 성능을 갖도록 이미지 캡처링(image captioning) 문제로 정의되고 사전 훈련된(pre-trained) 모델로서 이미지와 자연어(텍스트)를 서로 대조하면서 하나의 이미지와 그 이미지를 설명하는 텍스트를 매칭시키는 문제를 풀도록 학습된 모델을 지칭할 수 있다. CLIP-Actor는 CLIP 모델을 사용하여 구현한 시스템을 지칭할 수 있다.

[0038] 본 개시의 바람직한 실시예를 설명하기에 앞서 본 개시와 관련된 기존 작업을 먼저 간략히 언급하면 다음과 같다.

[0039] 먼저, 텍스트 구동 시각적 데이터 조작(text-driven visual data manipulation)과 관련된 기존 작업은 다음과 같다. 즉, 학습된 텍스트 및 이미지 조인트 임베딩 공간의 최근 발전은 이미지 및 3D 객체의 스타일 조작에 대한 연구에 불을 붙였다. CLIP 임베딩 공간은 풍부한 자연 이미지와 텍스트로 학습되며 원래 제로샷 이미지 및 언어 분석 작업을 위해 개발되었다. 흥미롭게도 그 표현은 직관적인 텍스트 가이드를 통해 시각적 데이터를 조작할 수 있을 만큼 강력하다는 것이 밝혀졌다. 이미지의 경우, 텍스트 조건부 이미지 생성이 CLIP에 의해 눈에

띄게 향상되었다. 대표적인 작업인 StyleCLIP은 자연어 텍스트 프롬프트가 주어지면 사전 훈련된 생성 모델의 잠재 코드를 최적화하여 입력 이미지를 조작한다. CLIPDraw는 경사하강법을 통해 곡선 세트의 매개변수를 최적화하여 텍스트 안내와 함께 이미지를 합성한다.

[0040] 이미지 영역과 유사하게, 여러 기존 작업들은 미분 가능 렌더링의 발전을 활용하여 조작 대상 영역을 3D 객체로 확장한다. 미분 가능 렌더링 기술을 사용하면 2D 렌더링 이미지에서 3D 개체로의 원활한 그라데이션 흐름이 가능하다. 따라서 CLIP은 2D 이미지를 통해 언어와 3D 양식을 연결한다. Dream Fields는 텍스트 프롬프트가 주어질 때 자유 공간의 암시적 표현을 사용하여 3D 구조를 생성한다. 다만, 3D 콘텐츠를 학습하거나 조작하기 위해 구조적 사전 지식을 활용하지는 않는다. 이를 통해 새로운 스타일로 유연한 콘텐츠 탐색이 가능하지만 추상적인 시각적 콘텐츠로 이어지는 경우가 많다.

[0041] 또 다른 기본 작업들로서 Michel et al.은 대상 텍스트 조건 프롬프트에 맞게 주어진 고정 소스 메쉬 스타일을 조작하는 CLIP 기반 최적화 방법인 Text2Mesh를 제안한다. Dream Fields와 달리, Text2Mesh는 고정된 T- 포즈 템플릿 휴먼 메쉬에 정의된 변위 및 텍스처 맵에 대해 3D 개체의 스타일을 지정하므로 강력한 구조적 사전 설정을 적용한다. 그리고 텍스트 프롬프트를 통해 그럴듯하고 흥미로운 메쉬 스타일과 질감을 보여준다. 그러나 주어진 템플릿 메쉬가 주어진 텍스트 프롬프트를 따르기 어려울 때 바람직하지 않은 스타일을 생성하는 문제가 있다. 예를 들어, 디테일한 인간 행동을 포함하는 텍스트는 주어진 인간 템플릿 메쉬의 포즈와 동작이 서로 일치하지 않을 때 스타일화에 실패한다.

[0042] 따라서, 후술하는 본 실시예에서는 입력 텍스트 프롬프트에 따라 디테일과 스타일로 휴먼 메쉬에 애니메이션을 적용하는 데 중점을 둔다. 또한, 파라메트릭 휴먼 메쉬 모델을 활용하여 기하학적 콘텐츠, 즉 포즈로부터 스타일을 분리한다. 이러한 솔루션을 통해, 본 실시예에서는 휴먼 메쉬의 포즈, 디테일 및 스타일이 입력 텍스트에 순차적으로 일치할 수 있다. 그리고 이를 통해 액션에서부터 스타일까지 입력 텍스트 프롬프트에 더 잘 따르는 3D 휴먼 객체를 안정적으로 조작할 수 있다.

[0043] 다음으로, 텍스트 구동 휴먼 모션 조작(text-driven human motion manipulation)과 관련된 기존 작업은 다음과 같다. 즉, 주어진 자연어 설명을 사용하여 인체 모션을 생성하기 위한 최근의 많은 접근들 중 어느 한 작업 라인은 자연어 설명을 순차적으로 번역하고 반복 신경 모델을 사용하여 인간의 골격 동작을 생성하도록 기계를 안내한다. 또 다른 작업 라인은 제한된 수의 폐쇄 세트 액션 범주에 따라 휴먼 모션을 생성한다. 한편, 본 실시예에서는 전체 문장에서의 텍스트 및 시각적 의미체계에 중점을 두고 다양한 자연어 설명을 다루도록 구성된다.

[0044] 최근 MotionCLIP와 TEMOS는 자연어 조건 메쉬 모션 생성 학습을 제안했다. MotionCLIP은 휴먼 모션 오토인코더를 학습하고 의미론적 유사도를 사용하여 잠재 공간을 CLIP 텍스트 및 이미지 공간과 호환되도록 만든다. 마찬가지로 TEMOS는 Transformer-VAE(variational autoencoder)를 사용하여 생성적 휴먼 메쉬 모션 잠재 공간을 학습하고 이를 DistilBERT를 통해 자연어 잠재 공간과 정렬하여 교차 모달 모션 잠재 공간을 구성한다. 두 가지 방법들 모두는 자연어 설명의 텍스트 및 시각적 의미를 캡처하기 위해 잠재 공간에 중점을 둔다. 한편, 후술할 본 실시예에서는 추천 시스템을 사용하여 설명을 실제 모션에 직접 매핑한다. 더욱이, 본 실시예의 디테일한 체적 메쉬는 앞서 언급한 기존 연구들보다 훨씬 더 표현력이 뛰어난 외관 속성으로 스타일화될 수 있다.

[0045] 다음으로, 움직이는 휴먼 메쉬의 질감과 기하학적 모양의 스타일화(texture and geometric stylization of human mesh in motion)과 관련된 기존 작업은 다음과 같다. 즉, 3D 메쉬 포즈 외에도 최근 작업에서는 휴먼 메쉬에 의상 모델링이나 텍스처 색상과 같은 다양한 수준의 디테일을 추가하고 있다. 휴먼 메쉬와 천 메쉬의 별도 모델링, 파라메트릭 휴먼 메쉬 모델의 뉴럴 확장(neural extension), 뉴럴 파라메트릭 접근법 및 뉴럴 암시적 접근법는 주어진 휴먼 스캔(human scans)에서 옷을 입은 휴먼 메쉬 결과를 보여주지만 표면 색상은 없다. 그러하니 관련 작업들은 질감과 기하학적 스타일을 별도로 다루고 있다. 최근, Saito et al.는 질감과 기하학적 스타일을 모두 복구하는 약지도 방법을 제안하고 있다. 하지만, 이러한 관련 작업들 중 어느 것도 제로샷 방식으로 예컨대 입력 텍스트만 사용하여 휴먼 모션의 다양한 색상 및 옷감 디테일을 생성하고 있지 않다.

[0046] 따라서 본 실시예에서는 제로샷(zero-shot)에서 휴먼 메쉬의 애니메이팅에 관련된 새로운 텍스트 구동 추천, 디테일화 및 질감화를 제시하고, 이를 통해 작업 데이터세트 없이 기계의 상상력을 통해 질감과 기하학적 디테일을 가진 움직이는 휴먼 메쉬를 생성한다.

[0047] 도 1은 본 개시의 일실시예에 따른 액션 아바타 생성 장치에 대한 개략적인 블록도이다. 그리고 도 2는 도 1의 액션 아바타 생성 장치에서 주어진 입력 텍스트 프롬프트(input text prompt)에 대하여 생성되는 움직이는 휴먼 아바타를 나타낸 예시도이다.

[0048] 도 1을 참조하면, 액션 아바타 생성 장치는 휴먼 모션 추천(human motion recommendation, HMR) 유닛(100)과 텍스트-시간 스타일화(text-time stylization) 유닛(300)을 포함한다.

[0049] 휴먼 모션 추천 유닛(100)은 텍스트를 입력받고, 텍스트 프롬프트로 주어지는 쿼리에 일치하는 액션 레이블을 데이터베이스에 저장된 휴먼 모션 데이터세트에서 찾아 휴먼 모션을 추천한다. 그리고, 텍스트-시간 스타일화 유닛(300)은 추천된 휴먼 모션에 대해 최적화 스타일 속성을 부여하여 움직이는 아바타를 생성한다.

[0050] 본 실시예의 액션 아바타 생성 장치는 인간에 대한 텍스트-시각 결합 이해에 기반한다. 예를 들어, 배우가 연극의 대본을 읽을 때, 통상 배우는 대본에 설명된 맥락에 따라 몸짓, 말투 및 옷의 이미지를 떠올린다. 이러한 텍스트-시각 결합 상상력은 기계-제작 미디어, 예를 들어 움직이는 스타일화 3D 휴먼들을 가속화하는 돌파구가 될 수 있다. CLIP의 텍스트-이미지 결합 임베딩 공간(text-image joint embedding space)을 활용하여 움직이는 스타일화 3D 휴먼을 기계에 구현할 수 있다. CLIP 임베딩 공간의 표현력을 통해, 텍스트와 이미지 간의 유사도 측정은 텍스트-3D 휴먼 메쉬(text-to-3D human mesh)를 모션으로 구축할 때 구체적인 지도를 제공한다. 이러한 지도를 발판으로 하여 본 실시예에서는 3D 휴먼 메쉬를 애니메이션화하는 텍스트 구동 추천 및 스타일화의 자동화된 프레임워크인 액션 아바타 생성 장치를 제공한다.

[0051] 액션 아바타 생성 장치는, 도 2에 나타낸 바와 같이 인간의 행동과 스타일을 설명하는 텍스트 프롬프트가 주어지면, 텍스트 프롬프트에 따라 움직이는 휴먼 메쉬의 짧은 클립을 만들 수 있다. 본 실시예에서 텍스트 프롬프트는 다음과 같다: "a baseball player swings a baseball bat."(야구 선수가 야구 방망이를 휘두른다.)

[0052] 본 실시예의 액션 아바타 생성 장치는, 주어진 쿼리 텍스트와 밀접한 상관 관계가 있는 움직이는 메쉬를 데이터베이스에서 검색하기 때문에 아티스트가 디자인한 추가 3D 메쉬 입력을 필요로 하지 않는다. 또한, 텍스트 프롬프트 내에서 시각적 및 텍스트 신호를 포착하기 위해 세분화된 텍스트 의미론적 매칭을 활용하는 계층적 텍스트 구동 휴먼 모션 추천 모듈을 활용할 수 있다.

[0053] 또한, 액션 아바타 생성 장치는, 포즈(pose)에 구애받지 않는 방식으로 분리형 뉴럴 스타일 필드(Decoupled Neural Style Fields, DNSF)를 최적화하여 메쉬 시퀀스를 상세하게 설명하고 텍스처화할 수 있다. DNSF는 시간적으로 일관되고 포즈에 구애받지 않는 방식으로 움직이는 휴먼 메쉬의 스타일 속성을 학습한다. 또한, 텍스트 기반 뉴럴 DNSF 최적화의 수렴을 개선하기 위해, 멀티모달 콘텐츠 메쉬 샘플링, 시공간 뷰 증강 및 마스크 가중 임베딩 어텐션을 추가로 적용할 수 있다. 마스크 가중 임베딩 어텐션은 안정적인 뉴럴 최적화를 위해 적용될 수 있다. 여기서, 최적화의 목적은 입력 텍스트 프롬프트와 스타일화된 3D 메쉬의 2D 렌더링 이미지 간의 상관 관계를 최대화하는 것이다. 이러한 뉴럴 최적화를 활용하면, 시공간적으로 증강된 렌더링된 이미지로 DNSF를 최적화하고 멀티모달 샘플링 전략으로 우수한 성능의 초기 콘텐츠 메쉬를 제공할 수 있다.

[0054] 전술한 구성에 의하면, 액션 아바타 생성 장치는, 제로샷에서 다양한 텍스트 설명과 함께 움직이는 시각적 및 물리적으로 그럴듯한 3D 휴먼 메쉬를 효과적으로 스타일링할 수 있다.

[0055] 도 3은 도 1의 액션 아바타 생성 장치의 휴먼 모션 추천(human motion recommendation, HMR) 유닛에 대한 개략적인 블록도이다.

[0056] 도 3을 참조하면, 휴먼 모션 추천 유닛(100)은 교차 모달 인식 매칭(cross-modal aware matching) 모듈(120) 및 텍스트 의미론적 매칭(textual semantic matching) 모듈(140)을 포함한다. 휴먼 모션 추천 유닛(100)은 프레임 단위로 액션 레이블이 정렬된 휴먼 모션 데이터세트를 사용할 수 있다.

[0057] 휴먼 모션 추천 유닛(100)은 입력 텍스트 프롬프트(110)을 받고 데이터베이스(도 6의 114 참조)에 저장된 휴먼 모션 데이터세트에서 텍스트 프롬프트로 주어지는 쿼리에 일치하는 액션 레이블을 찾고, 액션 레이블에 기초한 텍스트 의미론적(sementic) 매칭을 통해 멀티모달 콘텐츠 메쉬 샘플링(150)을 생성할 수 있다.

[0058] 교차 모달 인식 매칭 모듈(120)은 입력 텍스트 프롬프트(110)을 받고 데이터베이스에 저장된 휴먼 모션 데이터세트에서 텍스트 프롬프트로 주어지는 쿼리에 일치하는 액션 레이블을 찾을 수 있다. 또한, 교차 모달 인식 매칭 모듈(120)은 쿼리와 시각적 및 언어적으로 일치하는 원시 액션 레이블과 상기의 텍스트 프롬프트를 벡터화할 수 있다. 또한, 교차 모달 인식 매칭 모듈(120)은 벡터화를 통해 생성된 벡터들 간의 유사도를 계산할 수 있다. 이를 위해, 교차 모달 인식 매칭 모듈(120)은 벡터화 모듈과 유사도 계산 모듈을 구비할 수 있다.

[0059] 텍스트 의미론적 매칭 모듈(140)은 입력 텍스트 프롬프트(110) 내 문장에서 텍스트 의미체계를 캡쳐할 수 있다. 또한, 텍스트 의미론적 매칭 모듈(140)은, 캡쳐한 텍스트 의미체계를 통해 텍스트 프롬프트와 가장 관련성이 높은 액션 레이블을 찾을 수 있다.

[0060] 이와 같이, 휴먼 모션 추천 유닛(100)은 텍스트 프롬프트를 따르는 모션 시퀀스를 얻기 위해 데이터세트에서 시각적 및 텍스트적으로 관련된 액션 레이블을 검색하여 모션을 추천할 수 있다. 즉, 텍스트 프롬프트가 쿼리로 주어지면, 교차 모달 인식 매칭과 텍스트 의미론적 매칭의 2-단계 검색 기능을 가진 휴먼 모션 추천 유닛(100)을 통해, 쿼리와 시각적 및 언어적으로 연결된 원시 액션 레이블이 매칭될 수 있다. 이러한 계층적 매칭은 휴먼 모션 추천 유닛(100)이 교차 모달 인식 컨텍스트와 언어적 의미체계를 포착할 수 있도록 한다. 이러한 포괄적인 매칭 모듈은 우수한 초기 콘텐츠를 후속 신경망 스타일화 모듈(도 1 및 도 6의 300 참조)로 전달할 수 있다. 초기 콘텐츠는 멀티모달 콘텐츠 메쉬 샘플링(130)을 포함할 수 있다. 휴먼 모션 추천 유닛(100)은 계층적 멀티모달 모션 검색(hierarchical multi-modal motion retrieval) 유닛 또는 간략히 모션 추천 시스템으로 지칭될 수 있다.

[0061] 도 4는 도 3의 휴먼 모션 추천 유닛에 채용할 수 있는 계층적 멀티모달 모션 검색 모듈을 설명하기 위한 구성도이다.

[0062] 도 4를 참조하면, 계층적 멀티모달 모션 검색 모듈은 교차 모달 인식 매칭 모듈(120) 및 텍스트 의미론적 매칭 모듈(140)을 포함한다. 계층적 멀티모달 모션 검색 모듈은 교차 모달 인식 매칭 모듈(120)과 텍스트 의미론적 매칭 모듈(140)과의 사이에 위치하는 색인부(132) 및 필터(134)를 더 구비할 수 있다.

[0063] 교차 모달 인식 매칭 모듈(120)은 제1 인코더(122), 제1 벡터(123a), 제1 데이터베이스(123b), 및 제1 유사도 계산부(124)를 포함할 수 있다. 제1 인코더(122)는 텍스트 인코더로 지칭될 수 있다. 교차 모달 인식 매칭 모듈(120)은 텍스트 프롬프트(112)를 입력받는다. 텍스트 프롬프트(112)는 "Steve Jobs stretching arms"일 수 있다.

[0064] 교차 모달 인식 매칭 모듈(120)에서, 제1 인코더(122)는 입력되는 텍스트 프롬프트(112)와 텍스트 프롬프트(112)로 주어지는 쿼리에 일치하는 액션 레이블을 데이터베이스(160)에 저장된 휴먼 모션 데이터세트에서 찾아 입력받고, 쿼리와 액션 레이블을 인코딩하여 인코딩된 쿼리와 인코딩된 액션 레이블을 생성할 수 있다. 여기서 인코딩된 쿼리(123a)는 쿼리를 벡터화한 제1 벡터(123a)로 지칭될 수 있다. 그리고, 인코딩된 액션 레이블은 액션 레이블을 벡터화한 제2 벡터로 각각 지칭될 수 있고, 제1 데이터페이스(123b)에 저장될 수 있다. 또한, 교차 모달 인식 매칭 모듈(120)에서, 제1 유사도 계산부(124)는 벡터화를 통해 생성된 벡터들 간의 유사도를 계산할 수 있다.

[0065] 계산된 유사도 값들 중 높은 순서에서 가장 위에 있는 k개(Top-k)의 값들은 색인부(132)에 저장될 수 있다. 색인부(132)는 예를 들어 우선순위가 가장 높은 값에서부터 3개의 값(예컨대, 521,710, 1207)을 선택하여 인덱스할 수 있다. 색인된 값들과 데이터베이스(160)으로부터의 값들은 각각 필터(134)에 입력될 수 있다. 필터(134)는, 상위 k개(Top-k)의 인덱스가 선택된 상태에서, top-k 필터로서 해당 원시 액션 레이블을 검색할 수 있다.

[0066] 텍스트 의미론적 매칭 모듈(140)은 제2 인코더(142), 제2 벡터(143a), 제2 데이터베이스(143b) 및 제2 유사도 계산부(144)를 구비할 수 있다. 제2 인코더(142)는 언어 모델(language model)로 지칭될 수 있다.

[0067] 교차 모달 인식 매칭 모듈(120)에서, 제2 인코더(142)는 텍스트 프롬프트(112)를 받고 또한 필터(134)로부터 상위 k개의 유사도 값들에 대한 액션 레이블(이하 '제2 액션 레이블')을 데이터베이스(160)에서 찾아 입력받고, 텍스트 프롬프트(112)의 쿼리와 제2 액션 레이블을 인코딩하여 인코딩된 쿼리(143a)와 인코딩된 top-k 액션 레이블을 생성할 수 있다.

[0068] 여기서 인코딩된 쿼리(143a)는 쿼리를 벡터화한 제3 벡터로 지칭될 수 있다. 그리고, 인코딩된 top-k 액션 레이블은 액션 레이블을 벡터화한 제4 벡터로 각각 지칭될 수 있고, 제2 데이터페이스(143b)에 저장될 수 있다. 또한, 교차 모달 인식 매칭 모듈(120)에서, 제2 유사도 계산부(144)는 벡터화를 통해 생성된 벡터들 간의 유사도를 계산할 수 있다.

[0069] 텍스트 의미론적 매칭 모듈(140)은 입력 텍스트 프롬프트(110) 내 문장에서 텍스트 인코더로 캡쳐하고, 캡쳐한 텍스트 의미체계를 통해 텍스트 프롬프트와 가장 관련성이 높은 액션 레이블을 찾을 수 있다. 그리고, 텍스트 프롬프트와 가장 관련성이 높은 액션 레이블(152)에 기초하여 멀티모달 콘텐츠 메쉬 샘플링(MMCMS, 154)을 획득할 수 있다.

[0070] 전술한 계층적 멀티모달 모션 검색 모듈(이하 간략히 '검색 모듈'이라고도 한다)의 작동 원리를 좀더 상세히 설명하면 다음과 같다.

[0071] 텍스트 프롬프트가 쿼리로 주어지면, 검색 모듈은 데이터베이스(도 6의 160 참조)에서 가장 관련성이 높은 원시

액션 레이블을 찾는다. 다음, 검색 모듈은 쿼리와 모든 원시 액션 레이블을 텍스트 인코더 h(·)에 의해 인코딩하고 이들 간의 유사도를 측정한다. 원시 액션 레이블의 상위 k개 인덱스가 선택되고, 해당 원시 액션 레이블이 top-k 필터(134)에 의해 검색될 수 있다. 언어 모델 인코더 m(·)는 쿼리 및 top-k 액션 레이블을 벡터화할 수 있다. 가장 높은 유사도 점수를 받은 원시 액션 레이블은 입력 텍스트 프롬프트에 대해 최종적으로 일치하는 결과로 검색될 수 있다.

[0072] 좀더 구체적으로, 텍스트 프롬프트 y가 주어지면, 검색 모듈은 지속시간(duration) T 동안의 SMPL(skinned multi-persion linear) 포즈 매개변수의 시퀀스 $R_{1:T} = [R_1, \cdots, R_T]$를 검색할 수 있다. 단일 프레임 t에서, 메쉬 정점(mesh vertices) M_t는 다음과 같이 선형 매핑을 사용하여 획득될 수 있다.

수학식 1

[0073] $$M_t = M(R_t, \beta_t), \quad \forall t \in \{1, \cdots, T\}$$

[0074] 수학식 1에서, R_t는 포즈 매개변수(pose parameters)를 나타내고 β_t는 휴먼 메쉬의 모양 매개변수(shape parameters)를 나타낸다.

[0075] 프레임 t에서 단일 메쉬는 면들(faces, F) 및 3D 메쉬 정점(mesh vertices) $M_t \in R^{V\times 3}$(여기서 V는 정점 수)에 의해 표현될 수 있다. 모든 프레임에 대한 SMPL 메쉬면(mesh faces) F는 주어진 삼각 측량과 동일하므로, 메쉬 정점 M_t를 사용하여 단일 메쉬를 나타낼 수 있다. 따라서 모든 메쉬 정점들 $M_{1:T} = [M_1, \cdots, M_T]$는 휴먼 메쉬의 전체 시퀀스를 나타내며, "콘텐츠"(content)로서 분리형 뉴럴 스타일 필드(DNSF)로 옮겨질 수 있다. 그 다음에 DNSF는 각 메쉬 정점의 색상 및 변위와 같은 "스타일"(style)을 학습하고 텍스처된 메쉬의 시퀀스($M^*_{1:T}$)를 생성할 수 있다.

[0076] 다시 말해서, 검색 모듈은 교차 모달 인식 매칭을 통해 공동 이미지-텍스트 공간의 입력 텍스트 프롬프트와 유사한 액션 레이블을 찾을 수 있다. 여기서, 검색 모듈은 데이터베이스, 즉 특정 데이터베이스(예컨대, BABEL)에서 수집한 원시 액션 레이블 A 세트를 준비할 수 있. 따라서 텍스트 프롬프트 y가 주어지면 원시 액션 레이블 ai ∈ A의 집합 $A_k \subset A$를 다음과 같이 표현할 수 있다.

수학식 2

[0077] $$\mathcal{A}_k = \text{top-k}[\mathcal{S}(\mathbf{h}(a_i), \mathbf{h}(y))], \quad \text{where } \mathcal{S}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}$$

[0078] 수학식 2에서, h(·)는 사전 훈련된 텍스트 인코더이고, top-k[·]는 k개의 최적 일치를 반환하는 함수를 나타낸다. 유사도는 코사인 유사도에 의해 측정될 수 있다.

[0079] 구체적으로, 예를 들면, "거꾸로 걷는 남자"라는 입력 프롬프트를 생각해 볼 수 있다. 그 경우, 교차 모달 인식 매칭은 텍스트 인코더를 사용하여 입력 프롬프트와 액션 레이블을 벡터화하고 이들 간의 유사도를 계산한다. 이때, 일치하는 액션 레이블 집합 A_k는 {"walking in place", "walking backward", "walking laterally"}로 결정될 수 있고, 상위 1개 매칭 레이블은 "walking in place"일 수 있다. 이와 같이, 텍스트 인코더는 시각적으로 나타나는 단어에 초점을 맞추도록 학습되기 때문에 시각적 의미 예컨대 '걷기'를 포착할 수 있다. 여기서 A_k의 모든 요소는 시각적 공간의 입력 프롬프트와 밀접한 관련이 있다. 한편, 텍스트 인코더는 비디오 대신 스틸 이미지로 학습되므로 세밀한 액션 예컨대, "제자리에서 걷기"와 "뒤로 걷기"의 2개의 스틸 이미지에서 동일하게 나타나기 때문에 이들을 구분할 수 없다. 따라서 본 실시예의 검색 모듈에서는 단일 단계 검색을 보상하기 위해 다음과 같은 텍스트 의미론적 매칭(textual semantic matching)을 이용한다.

[0080] 텍스트 의미론적 매칭 모듈은 문장에서 텍스트 의미 체계를 캡처하여 입력 프롬프트와 가장 관련성이 높은 액션 레이블을 찾는다. 언어 전문가(예컨대, MPNet)를 활용하여 2단계 모듈이 텍스트 의미론과 문법 구조를 구별할

수 있도록 학습할 수 있다. 가장 일치하는 레이블 a는 다음의 [수학식 3]과 같이 검색될 수 있다.

수학식 3

[0081] $$a_* = \arg\max_{a_j \in \mathcal{A}_k} \mathcal{S}(\mathbf{m}(a_j), \mathbf{m}(y))$$

[0082] [수학식 3]에서 m(·)은 미리 훈련된 MPNet 인코더를 나타낸다.

[0083] 다시 한 번, 위의 "거꾸로 걷는 사람"의 예를 적용해 보면, top-k 액션 레이블의 순위가 다시 지정되고 가장 유사한 액션 레이블인 "뒤로 걷기"가 최종 결과로 검색된다. 검색된 액션 레이블과 연결된 메쉬 $M_{1:T}$ 시퀀스는 콘텐츠 메쉬 시퀀스로서 다음의 뉴럴 메쉬 스타일지정 파이프라인(neural mesh stylization pipeline)에 전달될 수 있다.

[0084] 이와 같이, 본 실시예의 검색 모듈은 움직이는 휴먼 메쉬의 분리된 스타일화 휴먼 메쉬(Decoupled Stylization of Human Meshes in Motion)를 이용할 수 있다. 즉, 검색 모듈은 컨텐츠 메쉬를 T 프레임들의 검색된 휴먼 모션 시퀀스 $M_{1:T}$로부터 샘플링된 $M_i \in R^{V\times 3}$으로서 표시할 수 있다. 이 경우, 메쉬의 표면 스타일 속성 $\{c,d\}\in\{R^{V\times 3}, R^{V}\}$ 은 지정된 삼각 측량을 통해 표면에 적용되는 꼭짓점별 RGB 색상 및 꼭짓점별 변위로 해석될 수 있다.

[0085] 한편, 기존의 뉴럴 스타일 필드는 고정된 스태틱 메쉬를 입력으로 받아 MLP(multi-layer perceptron)를 사용하여 스타일 속성을 학습한다. 그러나 뉴럴 스타일 필드는 한 번에 하나의 포즈를 취하는 메쉬를 사용하므로 휴먼 메쉬 시퀀스를 스타일화하려면 상당한 수의 MLP가 필요하다. 이에 본 실시예의 휴먼 메쉬 애니메이션 장치에서는 뉴럴 최적화 유닛에 분리형 뉴럴 스타일 필드(decoupled neural style field)를 새롭게 도입한다.

[0086] 도 5는 도 1의 액션 아바타 생성 장치의 뉴럴 최적화 유닛을 설명하기 위한 구성도이다.

[0087] 도 5를 참조하면, 뉴럴 최적화 유닛은 템플릿 휴먼 메쉬 정점들(template human mesh vertices, THMV, 210)을 받아 텍스트 인코더(220)에서 벡터화한 후 분리형 뉴럴 스타일 필드(decoupled neural style field, DNSF) 모듈(240)에 의해 콘텐츠 메쉬에서 스타일 필드를 분리할 수 있다. 분리되는 스타일 필드는 색상(color, C, 250a)과 변위(displacement, D, 250b)를 포함할 수 있다.

[0088] 콘텐츠 메쉬에서 스타일 필드를 분리하여 동작 중인 메쉬에 대한 스타일 속성을 학습하기 위해, 뉴럴 최적화 유닛은 하나의 신경망만을 필요로 할 수 있다. 특히, 먼저 템플릿 휴먼 메쉬 M_c의 스타일 속성을 매핑하고 렌더링 직전에 콘텐츠 메쉬 M_i, $\forall i\in\{1,\cdots, T\}$와 병합할 수 있다(도 6 참조). DNSF 모듈(240)은 기본 뉴럴 스타일 필드와 동일한 메쉬 스타일을 찾는 동시에 콘텐츠 메쉬에서 스타일을 효과적으로 분리할 수 있다. 실제로 MLP G_Θ로 매개변수화된 DNSF 모듈(240)은 템플릿 휴먼 메쉬의 꼭짓점 예컨대 T-포즈형(posed) SMPL M_c을 다음의 [수학식 4]와 같이 포즈에 구애받지 않는 방식으로 색상 스타일 속성(C, 250a) 및 변위 스타일 속성(D, 250b)에 매핑할 수 있다.

수학식 4

[0089] $$DNSF:\ G_\theta(M_c) \mapsto \{c, d\}$$

[0090] 또한, 뉴럴 최적화 유닛에서는 메쉬 꼭짓점에 푸리에 기능 기반 위치 인코딩(도 6의 220 참조)을 사용하여 스타일 필드가 더 높은 주파수 디테일을 처리하도록 구성될 수 있다. 구체적으로, MLP G_Θ는 꼭짓점별 RGB 값, $c\in[0, 1]^{V\times 3}$ 및 꼭짓점별 변위 값, $d\in[-0.1, 0.1]^{V}$를 입력값으로 위치 인코딩된 특징을 가져올 수 있다. 그 후, 예측된 스타일 속성들은 컨텐츠 포즈된 메쉬 M_i에 적용되어, 스타일화된 휴먼 메쉬 M_i를 생성하는데 이용될 수 있다.

[0091] 이와 같이 텍스트 기반 DNSF 최적화. 텍스트 기반 DNSF op[1] 타이밍의 핵심은 시각적 메쉬 관찰과 입력 텍스트 프롬프트 간의 의미론적 상관 관계를 최대화하는 것입니다.

[0092] 한편, 텍스트 인코더는 2D 이미지에 대해서만 설계 및 훈련되기 때문에 생성된 3D 메쉬 자체와의 의미론적 상관관계를 측정하기 위해 CLIP을 직접 활용할 수 없다. 이에 본 실시예의 뉴럴 최적화 유닛에서는 3D 물체의 관찰이 어떤 관점에서든 유사하게 설명될 수 있다는 직관적인 아이디어를 활용한다. 즉, CLIP의 표현 능력을 지도 신호로 활용하기 위해 먼저 호환성을 위해 3D 메쉬의 이미지를 렌더링한다. 랜덤하게 샘플링된 N개의 카메라 포즈들, $p=[p_1,\cdots, p_N]$를 사용하여 스타일화된 메쉬 M^*_i를 차등적으로 렌더링하여 N-뷰 렌더링된 이미지 $I^*_{ij},\ \forall j \in \{1, 2,\cdots, N\}$를 얻을 수 있다. 따라서 주요 최적화 목표인 의미론적 손실(semantic loss, SL)(도 6의 700 참조)은 사전 훈련된 CLIP 이미지 및 텍스트 인코더, g(·) 및 h(·)를 이용하여 다음의 [수학식 5]와 같이 정의될 수 있다.

수학식 5

[0093]
$$\mathcal{L}_s = 1 - \frac{\bar{\mathbf{g}}(\mathbf{I}_i^*)^\top \mathbf{h}(y)}{\|\bar{\mathbf{g}}(\mathbf{I}_i^*)\|_2 \|\mathbf{h}(y)\|_2}, \quad \bar{\mathbf{g}}(\mathbf{I}_i^*) = \frac{1}{N}\sum\nolimits_{j=1}^{N} \mathbf{g}(\mathbf{I}_{ij}^*)$$

[0094] [수학식 5]에서, y는 입력 텍스트 프롬프트를 나타내고, $g(I^*_i)$ 및 $h(y)\in R^{512}$는 이미지 및 텍스트 프롬프트에 대한 정규화되지 않은 CLIP 임베딩 벡터들을 나타낸다.

[0095] [수학식 5]에서 알 수 있듯이, 의미론적 손실은 기본적으로 스타일화된 메쉬 M^*_i에 대한 정규화된 평균 임베딩 벡터와 입력 텍스트 프롬프트 y에 대한 정규화된 임베딩 간의 코사인 유사도일 수 있다.

[0096] 도 6은 본 개시의 다른 실시예에 따른 액션 아바타 생성 장치에 대한 개략적인 블록도이다.

[0097] 도 6을 참조하면, 액션 아바타 생성 장치는 휴먼 모션 추천 유닛(100)과 텍스트-시간 스타일화 유닛(300)에 더하여 뉴럴 최적화(neural optimization) 유닛(200)을 더 포함할 수 있다. 뉴럴 최적화 유닛(200)은 시공간 뷰 증강(spatio-temporal view augmentation) 유닛(500)을 포함할 수 있으나, 이에 한정되지 않는다. 시공간 뷰 증강 유닛(500)은 액션 아바타 생성 장치의 독립적인 유닛으로 구성될 수 있다.

[0098] 휴먼 모션 추천 유닛(100)은 교차 모달 인식 매칭(cross-modal aware matching, CMAM) 모듈(120)과 텍스트 의미론적 매칭(textual semantic matching, TSM) 모듈(140)을 구비할 수 있다. 교차 모달 인식 매칭 모듈(120)은 입력 텍스트 프롬프트(110)로 주어지는 쿼리에 일치하는 메쉬 정점들(132)을 데이터베이스(160)에 저장된 휴먼 모션 데이터세트에서 찾을 수 있다. 그리고, 텍스트 의미론적 매칭 모듈(140)은 휴먼 모션 데이터세트에 대하여 메쉬 정점들(132)에 기초한 텍스트 의미론적(sementic) 매칭을 통해 액션 레이블(144)을 추출할 수 있다. 추출된 액션 레이블에 대응하는 휴먼 모션은 멀티모달 콘텐츠 메쉬 샘플링(multi-modal content mesh sampling, MMCMS, 154)으로 추출될 수 있다. 데이터베이스(160)는 넓은 의미에서 휴먼 모션 추천 유닛(100)에 포함될 수 있으나, 이에 한정되지 않고 별도의 유닛이나 시스템으로 구성될 수 있다.

[0099] 뉴럴 최적화 유닛(200)은 위치 인코딩(position encoding) 모듈(220)과 DNSF 모듈(240)을 구비할 수 있다. 위치 인코딩 모듈(220)은 데이터베이스에서 템플릿 휴먼 메쉬 벡터(template human mesh vertices, THMV, 210)을 받아 위치 기반 인코딩을 수행할 수 있다. 템플릿 휴먼 메쉬 벡터(210)는 텍스트 프롬프트에 의해 주어지는 휴먼 템플릿 메쉬의 자세(pose)를 포함할 수 있다. 자세는 휴먼 템플릿 메쉬 상의 적어도 일부의 3차원 좌표들로 표현될 수 있다. 그리고, DNSF 모듈(240)은 특정 자세를 가지는 휴먼 템플릿 메쉬의 전체 시퀀스를 "콘텐츠"로 받고 메쉬 꼭지점의 색상(color, 250a)과 변위(displacement, 250b) 등의 "스타일"을 학습하고, 텍스쳐 메쉬의 시퀀스를 생성할 수 있다.

[0100] 텍스트-시간 스타일화 유닛(300)은 휴먼 모션 추천 유닛(100)으로부터 멀티모달 콘텐츠 메쉬 샘플링(154)을 받아 최적화된 스타일 속성(optimized style attributes, OSA)를 부여하여 움직이는 3D 휴먼 모션들에 의한 모션 시퀀스(900)를 생성할 수 있다. 또한, 텍스트-시간 스타일화 유닛(300)은 뉴럴 최적화 유닛(200)으로부터 색상(250a)과 변위(250b)가 학습된 텍스쳐 메쉬의 시퀀스를 이용하여 멀티모달 콘텐츠 메쉬 샘플링(154)에 대한 최적화된 스타일 속성을 부여할 수 있다.

[0101] 시공간 뷰 증강 유닛(500)은 시간 뷰 증강(temporal view augmention) 모듈(510), 공간 뷰 증강(spation view

augmention) 모듈(530), 마스크 가중 임베딩 어텐션(mask-weighted emboding attention) 모듈(550), 및 CLIP 인코더(570)를 구비할 수 있다. 시간 뷰 증강 모듈(510)과 공간 뷰 증강 모듈(530)은 단일 시공간 뷰 증강 모듈로 구성될 수 있다.

[0102] 시공간 뷰 증강(spatio-temporal view augmentation) 모듈과 관련하여, 기존 연구에서는 3D 시점 또는 2D 이미지 증강과 같은 공간적 증강이 콘텐츠 생성의 품질을 향상시킨다는 것을 보여주었다. 이에 본 실시예에서는 시간적 움직임(temporal movement)에서 비롯된 멀티-뷰 속성(multi-view property)을 가진 휴먼 메쉬를 활용한다. 즉, 휴먼 모션의 시공간적 맥락과 상기의 분리형 뉴럴 스타일 필드 표현을 활용하기 위해, 시공간 뷰 증강을 사용할 수 있다.

[0103] 시공간적 뷰 증강에 의하면, DNSF의 강점이 증폭될 수 있다. 위에서 DNSF G_Θ가 포즈에 구애받지 않는 템플릿 SMPL 메쉬를 입력으로 사용할 수 있음을 언급한 바 있다. 따라서, 의미론적 손실(L_s)는 DNSF 학습을 위한 모션 시퀀스 내의 임의의 콘텐츠 메쉬 $M_i \in \{1,...,T\}$로 측정될 수 있다. 중앙 프레임 또는 텍스트 프롬프트에 가장 잘 맞는 프레임을 샘플링하여 사용할 수 있다. 이것은 DNSF 학습에 유리한 관점으로 손실을 측정할 기회를 증가시킨다. 콘텐츠 메쉬의 고유한 선택이 텍스트 프롬프트를 따르지 않을 때 그럴듯한 색상과 기하학적 디테일을 생성하지 못하는 점을 고려할 때 콘텐츠 메쉬 샘플링 전략은 중요한 디자인 선택일 수 있다.

[0104] 마스크 가중 임베딩 어텐션 모듈(550)은 _____(발명자님의 추가 설명 부탁 드립니다).

[0105] CLIP 인코더(570)는 시공간 뷰 증강 유닛(500)에 의해 증강된 벡터를 출력할 수 있다.

[0106] 전술한 시공간 뷰 증강 유닛(500)은 휴먼 모션 추천 모듈(100)에서 생성되는 멀티모달 콘텐츠 메쉬 샘플링(154)에 DNSF 모듈(240)에서 생성되는 스타일 속성을 덧셈 연산기(270)로 더하여 얻어지는 3D 아바타(280)를 입력 데이터로 사용할 수 있다.

[0107] 시공간 뷰 증강 유닛(500)에서 출력되는 휴먼 모션 시퀀스와 입력 텍스트 프롬프트(110)의 출력 값에 기초하여 의미론적 손실(SL, 700)을 계산하고, 이를 토대로 액션 아바타 생성 장치나 그 구성요소들 중의 적어도 일부를 학습시킬 수 있다.

[0108] 전술한 바와 같이, 액션 아바타 생성 장치의 텍스트 구동 휴먼 모션 추천 모듈은, 휴먼 액션에 대한 텍스트 설명이 주어질 때, 모션 데이터베이스로부터 의미상 가장 잘 일치하는 모션 시퀀스를 찾는다. 콘텐츠 메쉬는 멀티모달 컨텍스트(context)에서 샘플링된다. 분리형 뉴럴 스타일 필드는 T-포즈를 취한 휴먼 메쉬를 가져와서 텍스트 기반 스타일 특성을 학습한 다음, 콘텐츠 메쉬에 적용된다. 시공간 뷰 증강 및 가중치 렌더링된 이미지를 적용하여 렌더링된 이미지와 텍스트 간의 유사도로 신경망 최적화를 안내할 수 있다.

[0109] 또한, 전술한 바와 같이, 액션 아바타 생성 장치는 메쉬의 정점 색상과 변위를 사용하여 스타일을 지정하고 입력 텍스트의 설명에 맞는 3D 모션을 시각화할 수 있다. 예를 들어, "청바지를 입고 스티브 잡스를 걷고 있다(walking Steve Jobs wearing blue jeans)"라는 자연어 프롬프트를 생각해 볼 수 있다. 이때, 추가로 고정된 3D 메쉬 입력을 준비하는 대신, 본 실시예에서는 사전에 준비된 데이터세트에서 모션 시퀀스를 검색하여 입력 프롬프트(예컨대, 걷기)에 따르는 3D 메쉬 시퀀스를 얻을 수 있다. 검색된 메쉬 시퀀스는 메쉬 스타일화의 "콘텐츠"가 된다. 그리고, 메쉬 정점의 색상과 변위를 학습하기 위해 신경망 모델을 최적화하여 메쉬에 특성 예컨대, 옷, 머리카락 등의 스타일 속성을 부여할 수 있다. 마지막으로, 본 실시예의 휴먼 메쉬 애니메이션 장치는 청바지를 입은 스티브 잡스가 걷고 있는 짧은 클립(도 6의 900 참조)을 생성할 수 있다.

[0110] 도 7은 도 6의 액션 아바타 생성 장치의 휴먼 모션 추천 유닛에 의해 생성되는 멀티모달 콘텐트 메쉬 샘플링을 나타내는 예시도이다.

[0111] 도 7을 참조하면, 멀티모달 콘텐트 메쉬 샘플링(multi-modal content mesh sampling)을 위해 스타일을 지정할 콘텐츠 메쉬를 선택하는 방법들 중 하나는, 메쉬 시퀀스 내에서 단일 메쉬를 임의로 선택하는 것이다. 하지만 신중한 텍스트 프롬프트와 메쉬의 렌더링된 이미지와의 의미론적 정렬은 최적화에 중요하기 때문에, 모션 내에서 최상의 텍스트 준수 메쉬를 찾는 멀티모달 콘텐츠 메쉬 샘플링을 이용할 수 있다. 특히, 콘텐츠 메쉬 시퀀스 $I(M_{1:T})$의 이미지를 렌더링하고 입력 텍스트 프롬프트(y)를 사용하여 각 이미지의 유사도 점수를 계산할 수 있다.

[0112] 예를 들어, 위에서 언급한 멀티모달 콘텐츠 메쉬 샘플링과 관련하여, 텍스트 프롬프트(text prompt)가 "a man jumping kick"(점프 킥하는 남자)로 주어지면 검색된 모션의 각 메쉬를 이미지로 렌더링하고 점프 킥 동작과 의

미상 일치하는 메쉬 프레임을 찾을 수 있다. 즉, 도 7에서는 주어진 텍스트 프롬프트를 따르는 5개의 메쉬 프레임들을 찾고, 그 중에 가장 일치하는 최상위 k(top-k) 메쉬 프레임으로서 3개의 메쉬 프레임을 선택한 것을 보여준다. 각 프레임의 유사도 점수가 각 아바타의 하부에 표시되어 있다.

[0113] 도 8은 도 6의 액션 아바타 생성 장치의 비교예에서 부주의한 무작위 자르기의 문제를 예시하기 위한 도면이다.

[0114] 본 실시예에서 액션 아바타 생성 장치는 사전 훈련된 CLIP 인코더 g(·)가 렌더링된 이미지를 인코딩하기 전에 무작위 자르기 및 원근 변환을 포함하여 미분 가능한 2D 이미지 증강을 적용할 수 있다. 이러한 2D 증강은 DNSF가 다양한 관점의 이미지들에서 스타일 속성을 학습하는 데 도움이 되므로 3D 콘텐츠에서 더 나은 일반화를 달성할 수 있다.

[0115] 한편, 부주의 한 무작위 자르기를 적용할 때 문제가 발생할 수 있다. 즉, 도 8에 예시한 바와 같이, 기존의 작업들에서는 렌더링된 이미지를 자르기 위해 극단적인 클로즈업을 적용하여 빈 렌더링에 의해 심하게 샘플링할 수 있다. 이러한 불필요한(redundant) 이미지는 적절하게 스타일화된 메쉬에 대해서도 텍스트 프롬프트를 준수하지 않으며 안정적인 DNSF 최적화를 방해한다.

[0116] 따라서, 본 실시예에서는 각 이미지의 전경 픽셀 비율에 따라 N개의 서로 다른 카메라 포즈들 $\{P_j\}_{j=1}^{N}$에서 CLIP 임베디드 벡터 $g(I_{ij}^{*})$에 가중치를 부여하여 상기의 문제를 완화할 수 있다. 즉, 렌더링된 이미지 I_{ij}^{*}가 메쉬 전경 픽셀의 극히 작은 부분을 가지고 있다면 임베딩 벡터 $g(I_{ij}^{*})$를 거부하도록 구성할 수 있다. 이것을 마스크 가중 임베딩 어텐션(mask-weighted embedding attention)이라고 지칭하고, [수학식 5]에 가중치(w_{ij})를 다음과 같이 추가하여 구현합니다.

수학식 6

[0117]
$$\bar{\mathbf{g}}(\mathbf{I}_i^*) = \sum_{j=1}^{N} \frac{w_{ij}}{\sum_{k=1}^{N} w_{ik}} \mathbf{g}(\mathbf{I}_{ij}^*), \quad w_{ij} = \frac{1}{HW} \sum_{H,W} \mathbb{1}[\mathbf{m}_{ij}(h,w) = 1]$$

[0118] [수학식 6]에서, H와 W는 렌더링된 이미지의 높이와 너비를 나타내고, 주어진 소정 콘텐츠 메쉬 M_i, 카메라 포즈 P_j에 대한 스타일화된 메쉬 M_i^{*}일 때, m_{ij}는 랜더링된 이미지 I_{ij}^{*}의 전경 마스크를 가리킨다.

[0119] 전술한 액션 아바타 생성 장치를 여러 측면에서 평가한 결과를 예시하면 다음과 같다.

[0120] 도 9는 본 실시예의 액션 아바타 생성 장치에 의해 생성된 움직이는 휴먼 아바타의 추천 모션 시퀀스에서 입력 텍스트 프롬프트와 함께 디테일 표면 지오메트리와 텍스쳐를 가진 대표 프레임들을 보여주기 위한 도면이다.

[0121] 본 실시예에서 액션 아바타 생성 장치는, 텍스트 프롬프트에 부합하는 가장 잘 어울리는 상위 3개의 메쉬 프레임들을 사용하는 모델로서, 마스크 가중 임베딩 어텐션(mask-weighted embedding attention)과 함께 멀티모달 콘텐츠 메쉬 샘플링 및 시공간 뷰 증강을 사용하는 모델로 정의할 수 있다. 또한 액션 아바타 생성 장치의 베이스는 기본적으로 검색된 모션 시퀀스의 중앙 프레임만을 활용하고, DNSF를 사용하지 않는 즉, 스타일 필드를 학습하기 위해 포즈형 메쉬(posed mesh)를 이용하는 모델로 준비될 수 있다. 액션 아바타 생성 장치의 베이스는 뉴럴 최적화에 초기 메쉬를 제안함으로써 Text2Mesh의 한계를 최소한 완화하기 때문에 여전히 강력한 베이스라인 모델이 될 수 있다. 여기서, Text2Mesh의 한계는 주어진 템플릿 메쉬가 주어진 텍스트 프롬프트를 따르기 어려울 때 바람직하지 않은 스타일을 생성하는 문제를 포함한다

[0122] 도 9에서와 같이, 액션 아바타 생성 장치의 정성적 결과를 보면, 각 아바타에서는 추천 모션 시퀀스의 대표 프레임을, 입력 텍스트 프롬프트와 함께 디테일 표면 형상과 텍스처를 함께 볼 수 있다. 즉, 액션 아바타 생성 장치는 우수한 동작과 우수한 스타일 일관성, 생생하고 매력적인 텍스처 결과를 보여줍니다.

[0123] 도 9에서, 아바타들 각각에 대한 입력 텍스트 프롬프트들은, "jumping spiderman"(점핑 스파이더맨), "Messing jumping over object"(물체 위로 점프하는 메쉬), "Freddle Mercury dancing"(춤추는 프레디 머큐리), "walking Gandhi"(걷고 있는 간디), "Alan Turing walking forwards"(앞으로 걸어가는 앨런 튜링), "bruno

mars dance stepping"(브루노 마스의 댄스 스텝핑), "Daft Punk turning music on"(음악을 켜는 다프트 펑크) 및 "Steve Jobs stretching arms"(팔을 쭉 뻗는 스티브 잡스)를 포함할 수 있다.

[0124] 도 9에서는 주어진 텍스트 프롬프트에 대한 액션 아바타 생성 장치의 추천과 메쉬 스타일화 결과를 시각화하고 있다. 단 하나의 텍스트 프롬프트만으로, 액션 아바타 생성 장치는 대표적인 포즈가 포함된 시각적으로 일치하는 모션 시퀀스를 검색할 수 있다. 또한 액션 아바타 생성 장치는 피사체의 대표 아이덴티티를 포착할 수 있다. 예를 들어, 스파이더맨의 물갈퀴가 있는 의상, 리오넬 메쉬 유니폼의 상징적인 색상, 프레디 머큐리의 헤어스타일, 간디가 입는 가운(robe)과 같은 기하학적 질감(geometric and texture)의 디테일이 각 피사체 또는 각 아바타에 잘 표현되어 있다.

[0125] 도 10은 본 실시예의 액션 아바타 생성 장치의 두 양태들의 성능 실험 결과와 함께 두 비교예들의 성능 실험 결과를 대비하여 설명하기 위한 도면이다.

[0126] 두 비교예는 Dream Fields로 표시된 제1 비교예(a)와 Text2Mesh로 표시된 제2 비교예(b)이고, 본 실시예의 두 양태는 CLIP-Actor(base)로 표시된 액션 아바타 생성 장치의 베이스(c)와 CLIP-Actor(full)로 표시된 액션 아바타 생성 장치(d)이다. 그리고 본 성능 실험에서, 입력 텍스트 프롬프트는 "a baseball player throwing a ball"(공을 던지고 있는 야구 선수)와 "Tony Stark wearing blue suit is walking forwards"(파란색 슈트를 입은 토니가 앞으로 걸어가고 있다)를 사용한다.

[0127] 전술한 두 비교예들과 본 실시예의 강력한 베이스라인 모델인 액션 아바타 생성 장치의 베이스(c)를 사용하여 본 실시예의 액션 아바타 생성 장치(d)를 평가할 수 있다.

[0128] 도 10에 나타낸 바와 같이 동일한 텍스트 프롬프트가 주어질 때의 시각적 질적 비교 결과를 확인할 수 있다. 즉, 제1 비교예(a)의 Dream Fields는 생성된 3D 컨텐츠의 흐릿하고 인식하기 어려운 렌더링을 보여준다.

[0129] 이러한 성능 저하는 Dream Field를 훈련할 때 구조적 사전 지식이 부족하기 때문에 발생한다고 가정할 수 있다. Dream Fields는 어떠한 구조적 안내 없이 가상 공간에서 3D 포인트의 점유도와 색상을 학습할 수 있다. 예를 들어 특정 동작을 수행할 때, Dream Field에 휴먼 바디의 물리적 제약을 가할 수 없다(도 10의 (a) 참조). 이것은 고도로 제한되지 않은 콘텐츠 생성 프로세스에 의미론적 감독만을 적용하면 물리적으로 제한된 휴먼 모션 및 질감을 처리하는 데 실패한다는 사실을 나타낸다.

[0130] 제2 비교예(b)의 Text2Mesh는 Dream Fields보다 향상된 텍스처 생성을 보여준다. 그러나 아티스트가 디자인한 휴먼 메쉬는 타겟 휴먼 액션과 전혀 상관관계가 없기 때문에 여전히 실패한다. 즉, 제2 비교예(b)의 Text2Mesh는 Dream Fields보다 더 나은 성능을 보여주지만 표면에 상당한 결함이 있다. 이러한 제2 비교예(b)의 한계는 "posed"(특정 자세를 가진) 콘텐츠 메쉬에서 스타일 필드를 학습하는 Text2Mesh의 고도로 결합된 스타일 필드에서 비롯될 수 있다.

[0131] 한편, Text2Mesh는 제한된 범위에 있도록 정점별 변위를 고정하여 스타일 속성이 콘텐츠를 크게 변경하는 것을 방지할 수 있다. 따라서, Text2Mesh 앞단에 본 실시예의 텍스트 구동 휴먼 모션 추천 모듈을 추가하고 텍스트에 맞는 콘텐츠 메쉬를 초기 지점으로서 제공하면, 즉 액션 아바타 생성 장치의 베이스와 같이 사용하면, Text2Mesh의 질적 성능을 크게 향상시킬 수 있다.

[0132] 제1 실시예(c)의 CLIP-Actor(base)는 사람이 인식할 수 있는 스타일 속성을 가진 더 많은 텍스트 준수 메쉬를 보여주지만 여전히 표면 결함을 가지고 있다.

[0133] 제2 실시예(d)의 CLIP-Actor(full)는 헤어스타일, 얼굴 아이덴티티 등 디테일한 색상과 기하학적 구조를 보여주면서 휴먼이 인식할 수 있고 의미에 부합하는 동작을 보여줍니다.

[0134] 또한, 제2 실시예(d)의 CLIP-Actor(full)는 야구 선수의 모자, 토니 스타크(Tony Stark)의 헤어스타일과 같이 의미론적으로 의미있는 디테일을 캡처하는 동시에 지저분한 스파이크(messy spikes)를 줄여 질적 결과를 더욱 향상시키고 있다.

[0135] 이와 같이, 새로운 DNSF, 멀티모달 콘텐츠 메쉬 샘플링 및 시공간 뷰 증강을 통해 본 실시예의 CLIP-Actor(full)는 다중 프레임 휴먼 모션에서 발생하는 멀티-뷰 렌더링을 활용할 수 있다. 따라서 그 결과는 훨씬 더 매끄럽고 텍스트와 일치할 수 있다. 한편, 본 실시예의 CLIP-Actor를 제외한 다른 비교예들의 방법은 휴먼 모션을 제대로 처리하지 못하고 있다. 이처럼, 본 실시예의 CLIP-Actor는, 시간적으로 일관되고 포즈에 구애받지 않는 메쉬 스타일 속성을 가지는, 텍스트 프롬프트를 따르는 휴먼 모션을 추천할 수 있다.

[0136] 본 실시예의 액션 아바타 생성 장치를 정량적으로 평가하기 위한 방법들 중 하나로써 이하의 사용자 직접 평가 결과를 참조할 수 있다.

[0137] 도 11은 본 실시예의 액션 아바타 생성 장치와 두 비교예들의 모션-텍스트 일관성, 스타일화 품질 및 전반적인 일관성에 대한 성능 평가를 위해 사전에 주어진 5개의 랜덤 텍스트-아바타 쌍의 결과들에 대한 46명의 비전문가 평가를 점수화하여 나타낸 그래프이다.

[0138] 도 11을 참조하면, 본 실시예의 CLIP-Actor은 모션-텍스트 일관성(motion-text consistency), 스타일화 품질(stylization quality) 및 전반적인 일관성(overall consistency)의 모든 측면들에서 두 비교예들(Text2Mesh 및 Dream Fields)보다 성능이 뛰어난 것을 알 수 있다.

[0139] 특히, 두 비교예들 모두의 성능은 그 중립점(neural point) 예컨대 3보다 낮은 점수를 받은 것을 보여주고, 본 실시예의 성능은 모션-텍스트 일관성, 스타일화 품질 및 전반적인 일관성 각각에서 4.33, 3.92 및 3.94와 같이 중립점을 크게 상회하는 것을 알 수 있다. 이러한 차이는 모션-텍스트 일관성에서 더욱 두드러지게 나타나며, 이는 본 실시예의 액션 일관성이 우수한 것을 검증한다.

[0140] 도 12는 본 실시예의 액션 아바타 생성 장치에 의해 생성된 특정 텍스트 기반의 움직이는 휴먼 아바타의 제작 과정에서 가중치 배제(-weight), 샘플링 배제(-sample), 시간 뷰 증강 배제(-aug_t), 공간 뷰 증강 배제(-aug_s) 및 시공간 뷰 증강 배제(-aug_st)의 경우들을 비교하여 나타낸 도면이다. 가중치 배제 버전은 단순히 가중치를 배제하기 보다 마스크 가중치 임베딩 어텐션(mask-weighted embedding attention)을 배제한 버전일 수 있다.

[0141] 도 12를 참조하면, 텍스트 구동 유먼 메쉬 애니메이팅 장치는 움직이는 휴먼 아바타의 제작 과정에서 분리형 뉴럴 스타일 필드(decoupled neural style field, DNSF)의 주요 구성요소를 선택적으로 제거하여 그 성능을 검증할 수 있다.

[0142] 주요 구성요소는 가중치(weight), 샘플링(sampling), 시간 뷰 증강(temporal view augmentation), 공간 뷰 증강(spatio view augmentation) 및 시공간 뷰 증강(spatio-temporal view augmentation)을 포함할 수 있다.

[0143] 텍스트 구동 휴먼 아바타의 제작 과정에서 NDSF의 주요 구성요소에 대한 선택적 배제 결과(ablation results), 본 실시예의 휴먼 아바타의 전체(full) 모델은 가중치 배제(-weight), 샘플링 배제(-sample), 시간 뷰 증강 배제(-aug_t), 공간 뷰 증강 배제(-aug_s) 및 시공간 뷰 증강 배제(-aug_st) 각각과 비교할 때 가장 부드러운 기하학적 형태와 생생한 색상을 보여준다.

[0144] 또한, 비교를 위해, 텍스트 구동 휴먼 모션 추천 모듈에 의해 생성한 멀티모달 콘텐츠(content)의 대응 부분을 각각의 케이스의 상단에 함께 보여주고 있다.

[0145] 실험 과정과 그 결과를 좀더 구체적으로 설명하면, 먼저 시간적 증강 효과(effects of temporal augmentation)를 확인하기 위해, DNSF가 단일 메쉬 프레임(top-1)만 활용하도록 시간적 뷰 증강을 제거할 수 있다. 즉, 다중 프레임 렌더링을 제거하면 시각적 품질이 크게 저하되어 샘플 표면에 눈에 띄는 스파이크와 비현실적인 색상이 나타날 수 있다(-aug_t 참조). 한편, 전체(full) 모델은 상위 3개 관련 프레임과 2D 증강, 3D 증강을 활용하므로 모델이 과적합되지 않도록 정규화하는 양식화된 메쉬의 다중 뷰에 영향을 미칠 수 있다.

[0146] 다음으로, 멀티모달 콘텐츠 메쉬 샘플링(multi-modal content mesh sampling)을 통해, DNSF는 텍스트 프롬프트에 맞는 더 나은 초기화로 최적화를 시작할 수 있다. 따라서 콘텐츠 메쉬의 나이브 샘플링의 경우, 더 나은 초기화로 최적화를 할 수 없으므로, 인식할 수 없는 얼굴 신원(unrecognizable face identifies), 저하된 질감(degraded texture) 및 저하된 기하학적 디테일(degraded geometric details)이 생성될 수 있다(-sample 참조). 이와 같이, 전체(full) 모델은 멀티모달 콘텐츠 메쉬 샘플링의 효과(effects of multi-modal content mesh sampling)를 가지는 것을 알 수 있다.

[0147] 다음으로, 마스크 가중 임베딩 어텐션(mask-weighted embedding attention)은 스타일화된 메쉬에 세밀한 터치(detailed touches)를 더한다. 최적화를 유도할 때 빈 렌더링(empty renderings)을 방지함으로써, 역전파에서 집중형 경사 흐름(focused gradient flow)을 통해 거친 것에서부터 미세한 것까지(coarse-to-fine) 기하학적 디테일과 텍스쳐 디테일을 학습할 수 있다.

[0148] 증강된 렌더링 이미지에 발끝이나 손가락 끝과 같은 신체 말단 영역의 극단적인 클로즈업이 포함된 경우, 임베딩 어텐션 방법은 집중형 경사 흐름을 통해 빈 공간이 아닌 메쉬 전경 픽셀(mesh foreground pixels)에 DNSF의

어텐션을 끌어당길 수 있다.

[0149] 도 12에서, 가중치 배제(-weight) 모델을 어텐션 메커니즘 없이 DNSF를 훈련할 때의 결과를 보여준다. 전체(full) 모델은 가중치 배제 모델보다 훨씬 부드럽고 세밀한 기하학적 디테일을 보인다.

[0150] 이와 같이, 본 실시예의 새로운 어텐션 메커니즘은 텍스트 구동 3D 객체 조작 파이프라인(text-driven 3D object manipulation pipelines) 뿐만 아니라 차별화가능한 렌더링 애플리케이션(differentiable rendering applications)에도 효과적으로 적용될 수 있다.

[0151] 도 13은 본 개시의 또 다른 실시예에 따른 액션 아바타 생성 장치에 대한 개략적인 블록도이다.

[0152] 도 13을 참조하면, 액션 아바타 생성 장치(1300)는 적어도 하나의 프로세서(processor, 1210)를 포함할 수 있다. 또한, 액션 아바타 생성 장치(1300)는 외부 네트워크와 연결되어 신호를 주고받거나 통신을 수행하는 송수신 장치(transceiver, 1330)를 포함할 수 있다. 또한, 액션 아바타 생성 장치(1300)는, 메모리(memory, 1320), 입력 인터페이스 장치(input interface device, 1340), 출력 인터페이스 장치(output interface device, 1350), 저장 장치(storage device, 1360) 등을 선택적으로 더 포함할 수 있다. 액션 아바타 생성 장치(1300)에 포함되는 구성 요소들은 버스(bus, 1370)에 의해 연결되어 서로 통신을 수행할 수 있다.

[0153] 프로세서(1310)는 중앙 처리 장치(central processing unit, CPU), 그래픽 처리 장치(graphics processing unit, GPU), 또는 본 개시의 실시예들에 따른 방법들이 수행되는 전용의 프로세서를 의미할 수 있다. 프로세서(1310)는 메모리(1320) 및 저장 장치(1360) 중 적어도 어느 하나에 저장된 프로그램 명령(program command)을 실행할 수 있다. 프로그램 명령은 텍스트 기반 휴먼 모션 추천을 이용하는 휴먼 메쉬 애니메이팅 방법을 구현하기 위한 적어도 하나의 명령을 포함할 수 있다. 이러한 적어도 하나의 명령은 소프트웨어 모듈이나 프로그램에 포함될 수 있다.

[0154] 메모리(1320) 및 저장 장치(1360) 각각은 휘발성 저장 매체 및 비휘발성 저장 매체 중에서 적어도 하나로 구성될 수 있다. 예를 들어, 메모리(1320)는 읽기 전용 메모리(read only memory, ROM) 및 랜덤 액세스 메모리(random access memory, RAM) 중에서 적어도 하나로 구성될 수 있다.

[0155] 송수신 장치(1330)는 근거리 무선 네트워크나 케이블 연결, 위성과의 통신, 범용 기지국과의 유선 또는 무선 통신, 모바일 에지 코아 네트워크나 코아 네트워크(core network)와의 아이디얼 백홀 링크(ideal backhaul link) 또는 넌(non)-아이디얼 백홀 링크의 연결 등을 위한 통신인터페이스나 서브통신시스템을 포함할 수 있다.

[0156] 입력 인터페이스 장치(1340)는 키보드, 마이크, 터치패드, 터치스크린 등의 입력 수단들에서 선택되는 적어도 하나와 적어도 하나의 입력 수단을 통해 입력되는 신호를 기저장된 명령과 매핑하거나 처리하는 입력 신호 처리부를 포함할 수 있다.

[0157] 출력 인터페이스 장치(1350)는 프로세서(1310)의 제어에 따라 출력되는 신호를 기저장된 신호 형태나 레벨로 매핑하거나 처리하는 출력 신호 처리부와, 출력 신호 처리부의 신호에 따라 진동, 빛 등의 형태로 신호나 정보를 출력하는 적어도 하나의 출력 수단을 포함할 수 있다. 적어도 하나의 출력 수단은 스피커, 디스플레이 장치, 프린터, 광 출력 장치, 진동 출력 장치 등의 출력 수단들에서 선택되는 적어도 하나를 포함할 수 있다.

[0158] 전술한 실시예들에 의하면, 텍스트 구동 애니메이팅 휴먼 메쉬를 위한 자동화 시스템이 제공될 수 있다. 이 자동화 시스템은, 멀티모달 인식(multi-modal aware) 및 텍스트 의미론적 매칭(textual semantic matching)을 활용하는 계층적 방법을 통해 입력 텍스트 프롬프트와 가장 적합하게 의미체계적으로 일치하는 휴먼 모션 시퀀스를 추천할 수 있다. 그리고 자동화 시스템은, 분리형 뉴런 스타일 필드(decoupled neural style field, DNSF)를 통해 포즈에 구애받지 않는 방식으로 합성을 통한 최적화를 통해 권장 모션의 메쉬들을 스타일화할 수 있다. 또한, 자동화 시스템은 멀티모달 샘플링 및 임베딩 가중치를 활용하는 새로운 뉴런 최적화 기술을 추가로 적용할 수 있고, 이에 의해 상세화(detailization) 품질 및 텍스처화(texturization) 품질을 안정화하고 향상시킬 수 있다.

[0159] 또한, 전술한 실시예들에 의하면, 텍스트 구동 애니메이팅 휴먼 메쉬를 위한 자동화 시스템은, 손 및 동물과 같은 다른 파라메트릭 메쉬 모델로 확장될 수 있으며, 이러한 3D 객체의 다양한 애니메이션을 가능하게 한다. 즉, 본 실시예의 자동화 시스템은, 자연어 설명(natural language description)과 결합된(paired), 모션으로 스타일화된 메쉬들의 데이터세트를 생성하는 다양한 응용에 활용될 수 있다.

[0160] 또한, 본 개시에 따른 방법들은 다양한 컴퓨터 수단을 통해 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등

을 단독으로 또는 조합하여 포함할 수 있다. 컴퓨터 판독 가능 매체에 기록되는 프로그램 명령은 본 개시를 위해 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다.

[0161] 컴퓨터 판독 가능 매체의 예에는 롬(rom), 램(ram), 플래시 메모리(flash memory) 등과 같이 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러(compiler)에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터(interpreter) 등을 사용해서 컴퓨터에 의해 실행될 수 있는 고급 언어 코드를 포함한다. 상술한 하드웨어 장치는 본 개시의 동작을 수행하기 위해 적어도 하나의 소프트웨어 모듈로 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

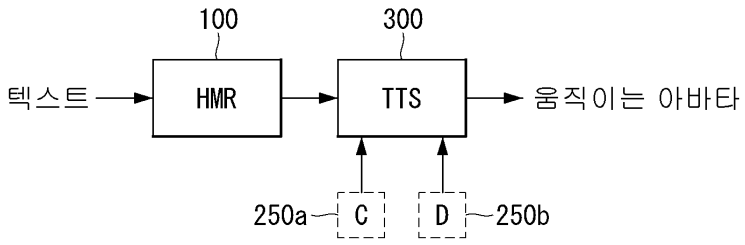
[0162] 본 개시의 일부 측면들은 장치의 문맥에서 설명되었으나, 그것은 상응하는 방법에 따른 설명 또한 나타낼 수 있고, 여기서 블록 또는 장치는 방법 단계 또는 방법 단계의 특징에 상응한다. 유사하게, 방법의 문맥에서 설명된 측면들은 또한 상응하는 블록 또는 아이템 또는 상응하는 장치의 특징으로 나타낼 수 있다. 방법 단계들의 몇몇 또는 전부는 예를 들어, 마이크로프로세서, 프로그램 가능한 컴퓨터 또는 전자 회로와 같은 하드웨어 장치에 의해(또는 이용하여) 수행될 수 있다. 몇몇의 실시 예에서, 가장 중요한 방법 단계들의 적어도 하나 이상은 이와 같은 장치에 의해 수행될 수 있다.

[0163] 실시 예들에서, 프로그램 가능한 로직 장치 예를 들어, 필드 프로그래머블 게이트 어레이가 여기서 설명된 방법들의 기능의 일부 또는 전부를 수행하기 위해 사용될 수 있다. 실시 예들에서, 필드 프로그래머블 게이트 어레이(field-programmable gate array)는 여기서 설명된 방법들 중 하나를 수행하기 위한 마이크로프로세서(microprocessor)와 함께 작동할 수 있다. 일반적으로, 방법들은 어떤 하드웨어 장치에 의해 수행되는 것이 바람직하다.

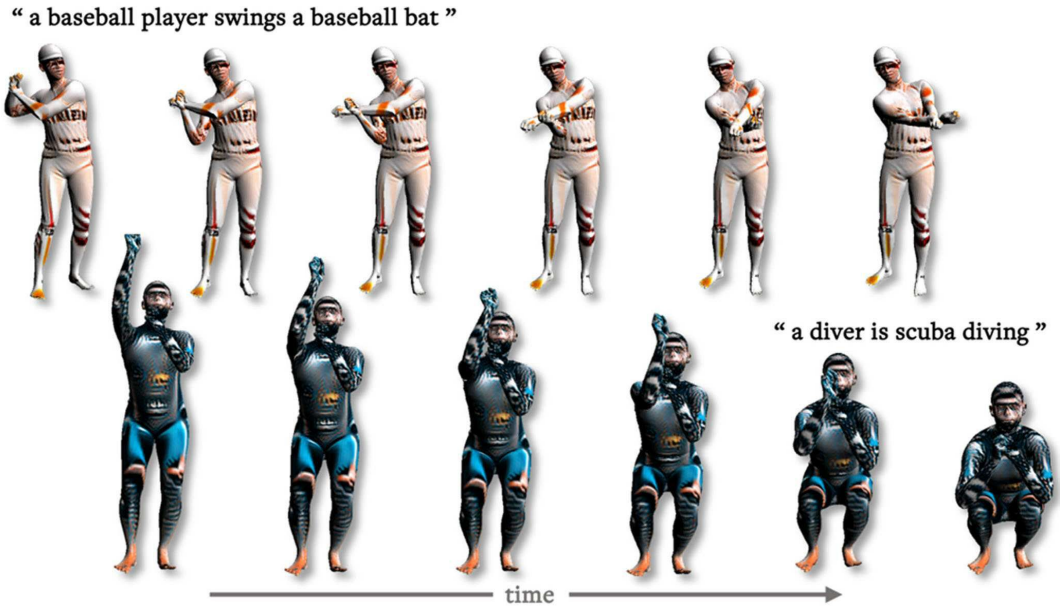
[0164] 이상 실시예를 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 특허 청구의 범위에 기재된 본 개시의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 개시를 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

도면

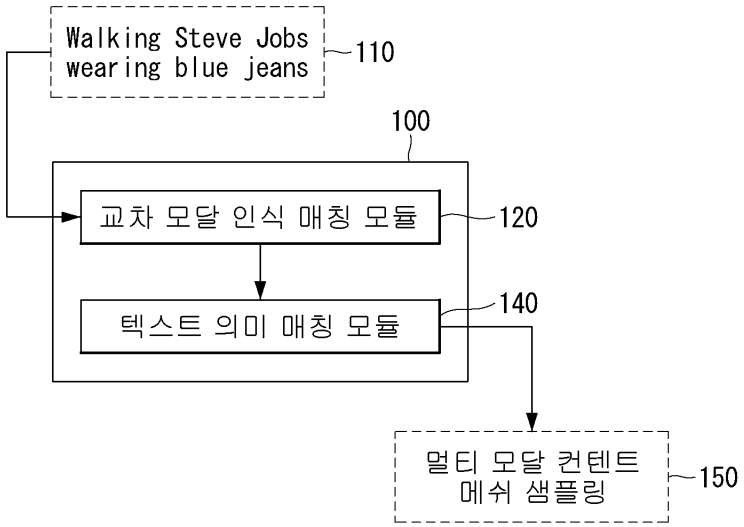
도면1



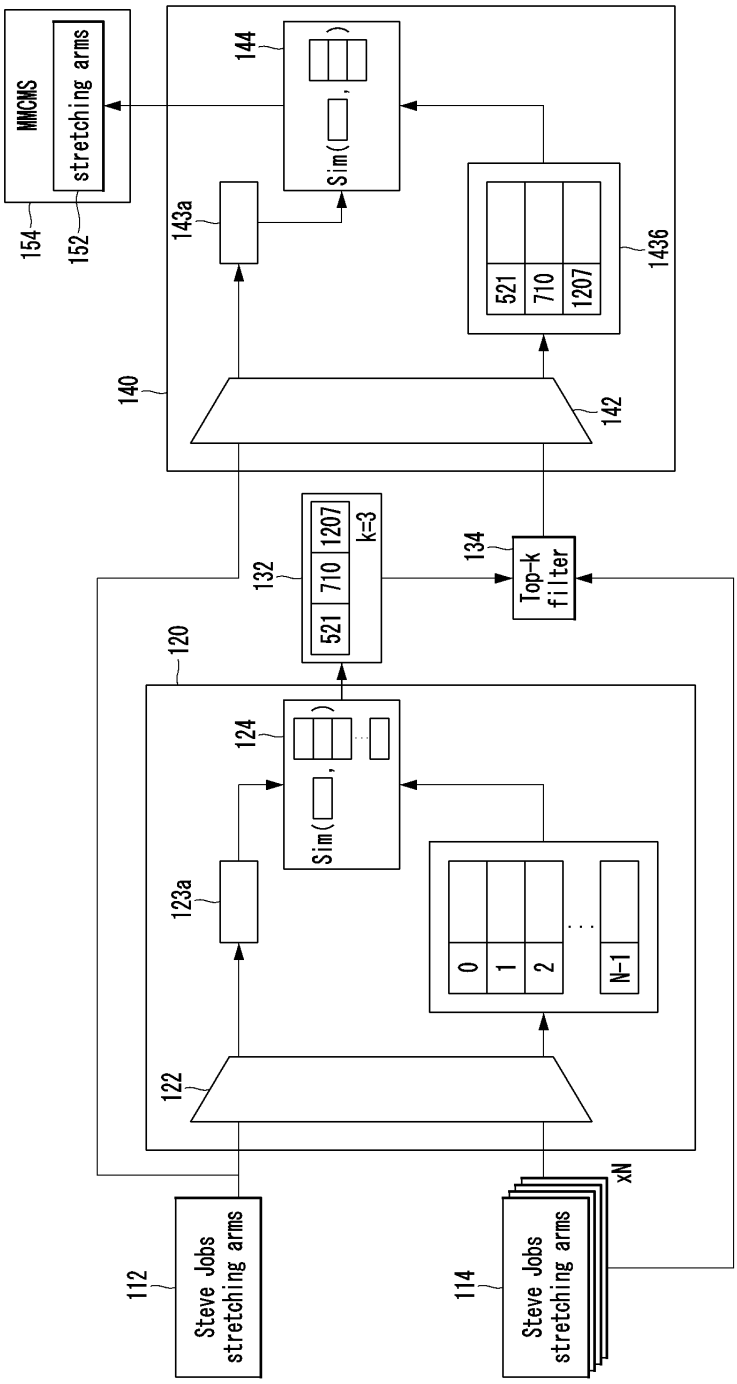
도면2


“ a baseball player swings a baseball bat ”
“ a diver is scuba diving ”
time

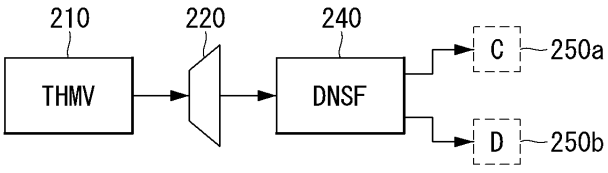
도면3


Walking Steve Jobs wearing blue jeans
110
100
교차 모달 인식 매칭 모듈
120
텍스트 의미 매칭 모듈
140
멀티 모달 컨텐트 메쉬 샘플링
150

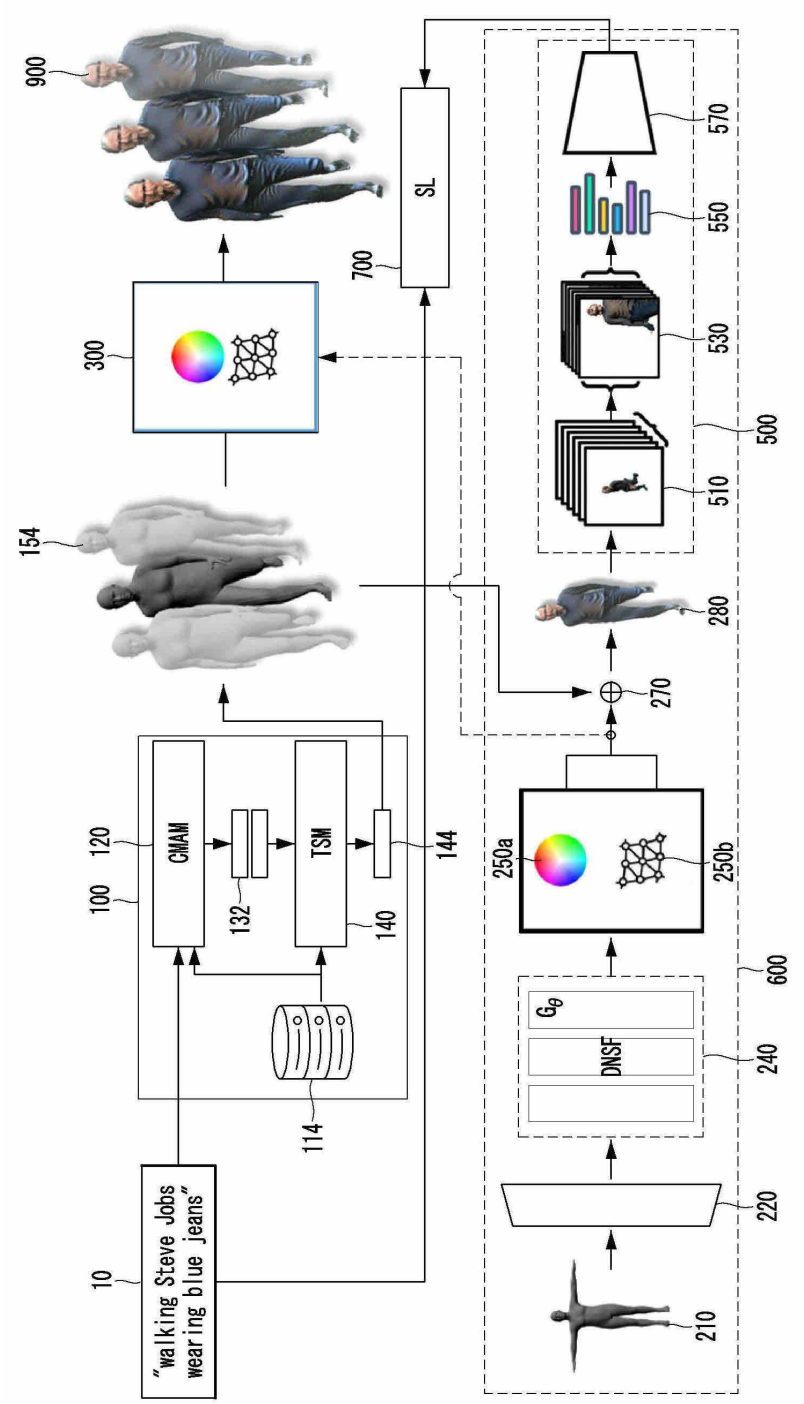
도면4


MMCMS
stretching arms
154
152
144
Sim(,)
143a
521
710
1207
1436
140
142
132
521
710
1207
k=3
134
Top-k filter
120
124
Sim(,)
123a
0
1
2
N-1
122
112
Steve Jobs stretching arms
114
Steve Jobs stretching arms
xN

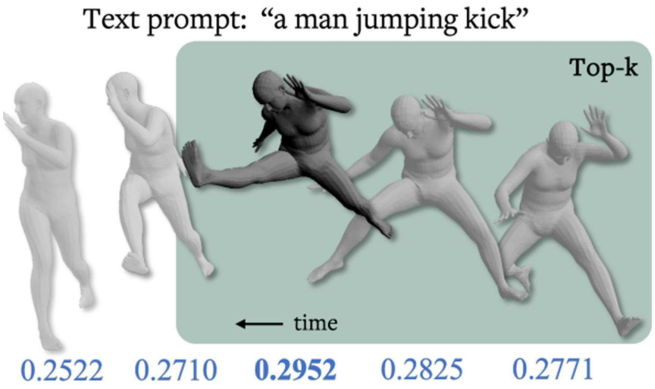
도면5


210
220
240
THMV
DNSF
C
250a
D
250b

도면6


900
SL
700
570
550
530
500
510
300
154
280
270
CMAM
120
100
132
TSM
140
144
250a
250b
600
G_θ
DNSF
240
220
210
114
"walking Steve Jobs wearing blue jeans"
10

도면7

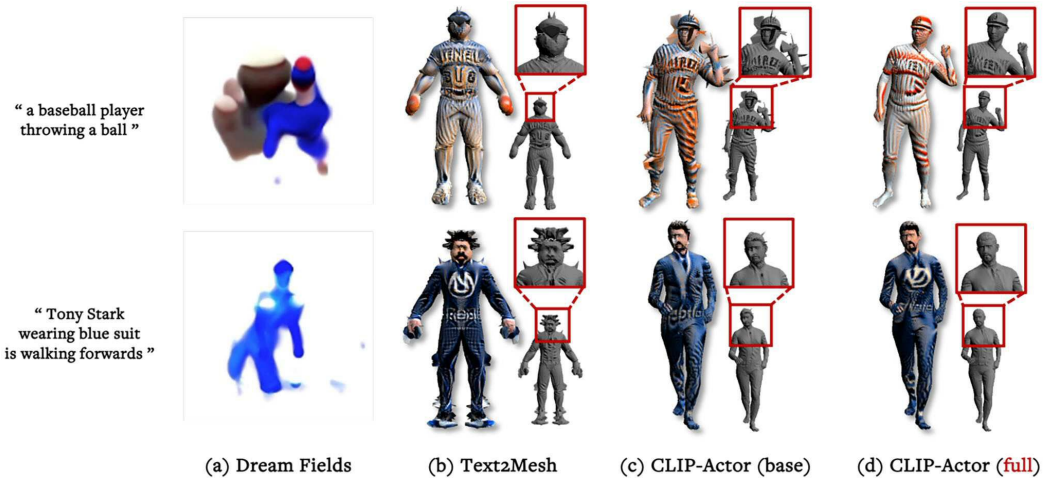

Text prompt: “a man jumping kick”
Top-k
time
0.2522
0.2710
0.2952
0.2825
0.2771

도면8

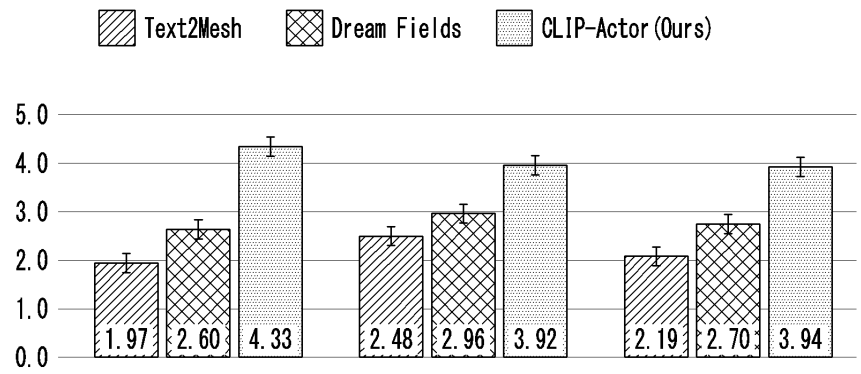
도면9


“ jumping Spiderman ”
“ Messi jumping over object ”
“ Freddie Mercury dancing ”
“ walking Gandhi ”
“ Alan Turing walking forwards ”
“ Bruno Mars dance stepping ”
“ Daft Punk turning music on ”
“ Steve Jobs stretching arms ”

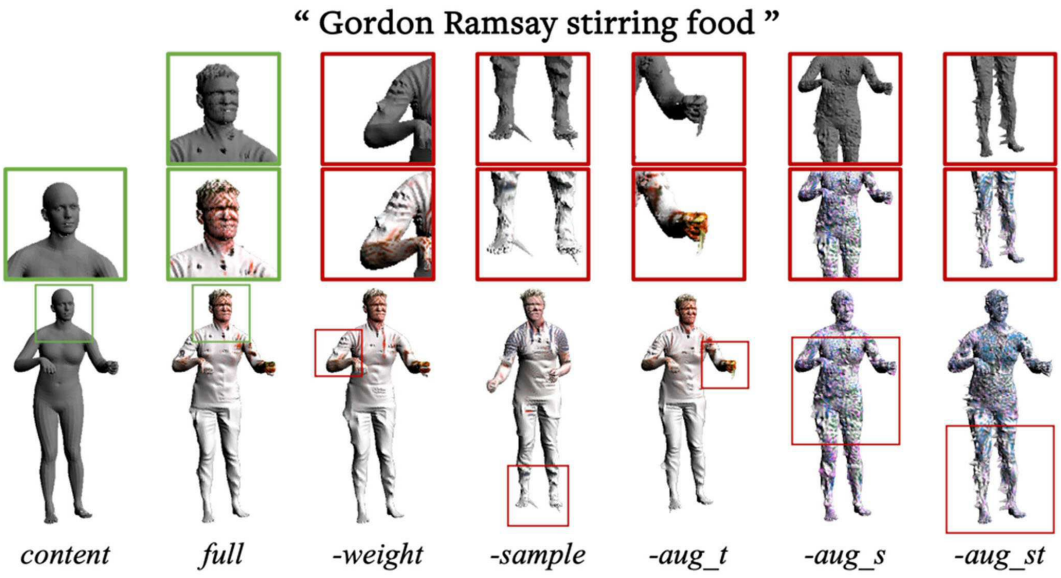
도면10



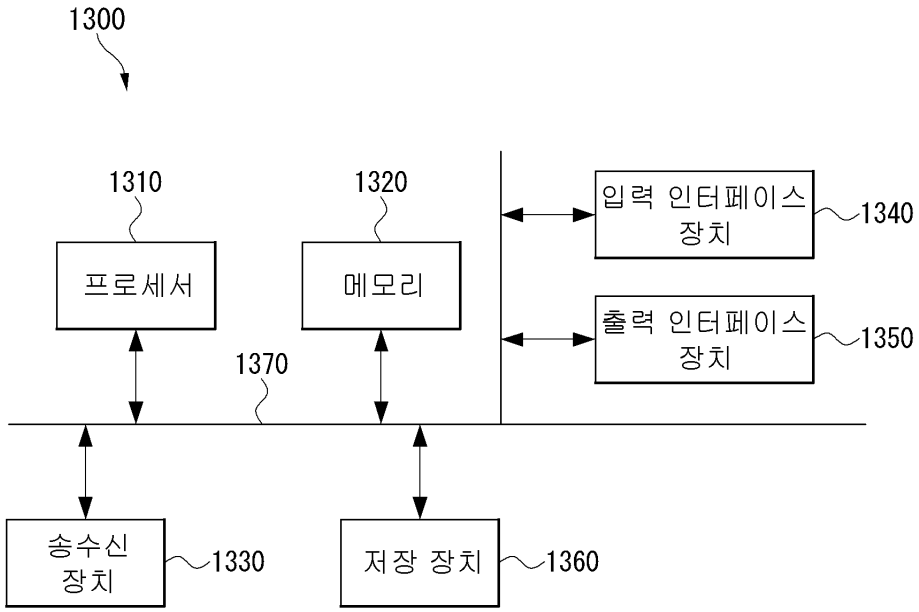
도면11



도면12



도면13


1300
1310
프로세서
1320
메모리
입력 인터페이스 장치
1340
출력 인터페이스 장치
1350
1370
송수신 장치
1330
저장 장치
1360