



(19) 대한민국특허청(KR)

(12) 등록특허공보(B1)

(45) 공고일자 2021년03월23일

(11) 등록번호 10-2231391

(24) 등록일자 2021년03월18일

(51) 국제특허분류(Int. Cl.)

H04N 7/01 (2006.01) G06T 7/246 (2017.01)

H04N 21/845 (2011.01)

(52) CPC특허분류

H04N 7/0137 (2013.01)

G06T 7/251 (2017.01)

(21) 출원번호 10-2019-0172877

(22) 출원일자 2019년12월23일

심사청구일자 2019년12월23일

(56) 선행기술조사문헌

JP2004086592 A

KR101558202 B1

(73) 특허권자

연세대학교 산학협력단

서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)

(72) 발명자

김선주

서울특별시 서대문구 연세로 50, 제4공학관 723호(신촌동, 연세대학교)

김윤지

서울특별시 서대문구 연세로 50, 제4공학관 707호(신촌동, 연세대학교)

(뒷면에 계속)

(74) 대리인

특허법인우인

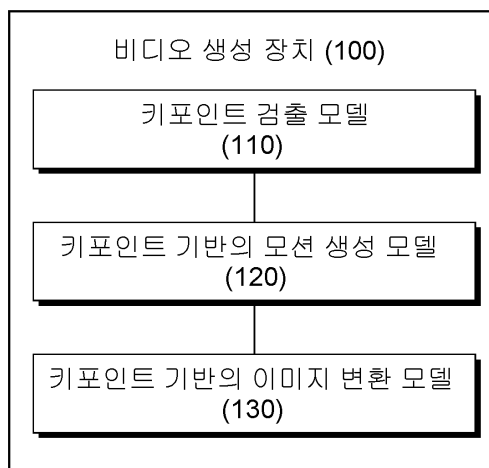
전체 청구항 수 : 총 18 항

심사관 : 박재학

(54) 발명의 명칭 키포인트 기반의 비디오 생성 방법 및 장치

(57) 요약

본 실시예들은 하나의 이미지에 존재하는 객체가 움직이도록 키포인트 검출 모델, 모션 생성 모델, 및 이미지 변환 모델을 학습하여 비디오를 생성하며, 이미지 내의 객체의 자연스러운 움직임을 생성할 수 있는 비디오 생성 방법 및 장치를 제공한다.

대표도 - 도1

(52) CPC특허분류

H04N 21/8455 (2013.01)

(72) 발명자

남성현

서울특별시 서대문구 연세로 50, 제4공학관 707호
(신촌동, 연세대학교)

조인

서울특별시 서대문구 연세로 50, 제4공학관 724호
(신촌동, 연세대학교)

이 발명을 지원한 국가연구개발사업

과제고유번호 2018-0-01858

부처명 과학기술정보통신부

과제관리(전문)기관명 정보통신기획평가원 (IITP)

연구사업명 인공지능SW 선도기술 및 유망 신기술 개발

연구과제명 가짜영상 판별성능 향상을 위한 동영상 조작 및 문장 기반 이미지 편집 알고리즘 개

발

기 여 율 1/1

과제수행기관명 연세대학교 산학협력단

연구기간 2019.02.01 ~ 2019.12.31

공지예외적용 : 있음

명세서

청구범위

청구항 1

컴퓨팅 디바이스에 의한 비디오 생성 방법에 있어서,

하나의 이미지로부터 키포인트 검출 모델을 통해 객체에 대한 키포인트를 추출하는 단계;

키포인트 기반의 모션 생성 모델을 통해 상기 키포인트가 변화된 키포인트 시퀀스를 생성하는 단계; 및

상기 하나의 이미지와 상기 키포인트 시퀀스를 이용하여 키포인트 기반의 이미지 변환 모델을 통해 비디오를 생성하는 단계를 포함하며,

상기 키포인트 검출 모델을 통해 기준 이미지(Reference Image)로부터 기준 키포인트를 추출하고, 상기 키포인트 검출 모델을 통해 대상 이미지(Target Image)로부터 대상 키포인트를 추출하고,

상기 키포인트 기반의 이미지 변환 모델은 상기 기준 키포인트 및 상기 대상 키포인트 간의 차이를 기반으로 상기 기준 이미지로부터 합성 이미지(Synthesized Image)를 생성하는 것을 특징으로 하는 비디오 생성 방법.

청구항 2

제1항에 있어서,

상기 키포인트 검출 모델은 상기 키포인트 기반의 이미지 변환 모델과 함께 학습되는 것을 특징으로 하는 비디오 생성 방법.

청구항 3

삭제

청구항 4

제1항에 있어서,

상기 키포인트 기반의 이미지 변환 모델은 상기 기준 이미지의 객체의 새로운 외형(Appearance)과 배경 마스크를 이용하여 상기 생성된 합성 이미지의 동적 영역을 처리한 변형 이미지(Translated Image)를 생성하는 것을 특징으로 하는 비디오 생성 방법.

청구항 5

제4항에 있어서,

상기 키포인트 기반의 이미지 변환 모델은 상기 기준 이미지 및 상기 합성 이미지를 혼합(Blend)하여 상기 변형 이미지(Translated Image)를 생성하는 것을 특징으로 하는 비디오 생성 방법.

청구항 6

제4항에 있어서,

상기 키포인트 기반의 이미지 변환 모델은

(i) 상기 기준 이미지에 상기 합성 이미지에 대한 배경 마스크를 적용한 제1 마스크 적용 이미지 및 (ii) 상기 합성 이미지에 상기 배경 마스크를 전환한 마스크를 적용한 제2 마스크 적용 이미지를 혼합(Blend)하여 상기 변형 이미지(Translated Image)를 생성하는 것을 특징으로 하는 비디오 생성 방법.

청구항 7

제4항에 있어서,

상기 키포인트 기반의 이미지 변환 모델과 상호 작용하는 이미지 판별기를 통해 상기 대상 이미지 및 상기 변형

이미지를 비교하여 구별하는 것을 특징으로 하는 비디오 생성 방법.

청구항 8

제1항에 있어서,

상기 키포인트 기반의 모션 생성 모델은 슈도 레이블과 함께 학습되며,

상기 키포인트 기반의 모션 생성 모델은 실제 이미지 시퀀스로부터 상기 키포인트 검출 모델을 통해 추출한 실제 키포인트 시퀀스를 상기 슈도 레이블로 설정하고,

상기 키포인트 기반의 모션 생성 모델과 상호 작용하는 키포인트 시퀀스 판별기를 통해 초기 키포인트로부터 생성한 키포인트 시퀀스 및 상기 실제 키포인트 시퀀스를 비교하여 구별하는 것을 특징으로 하는 비디오 생성 방법.

청구항 9

제8항에 있어서,

상기 키포인트 기반의 모션 생성 모델은 인코더와 디코더로 구성된 오토 인코더이며, 상기 인코더의 입력 레이어가 은닉 레이어에 매핑되고, 상기 은닉 레이어가 상기 디코더의 출력 레이어에 매핑되며,

상기 은닉 레이어에 노이즈가 추가되며,

상기 인코더에 상기 슈도 레이블, 상기 초기 키포인트 및 액션 클래스가 입력되고, 상기 디코더에 상기 초기 키포인트 및 액션 클래스가 입력되는 것을 특징으로 하는 비디오 생성 방법.

청구항 10

제9항에 있어서,

상기 인코더 및 상기 디코더에 은닉 노드가 방향을 가진 엣지로 연결된 순환 구조를 갖고 시계열 데이터를 처리하는 모델이 적용되는 것을 특징으로 하는 비디오 생성 방법.

청구항 11

하나의 이미지로부터 객체에 대한 키포인트를 추출하는 키포인트 검출 모델;

상기 키포인트가 변화된 키포인트 시퀀스를 생성하는 키포인트 기반의 모션 생성 모델; 및

상기 하나의 이미지와 상기 키포인트 시퀀스를 이용하여 비디오를 생성하는 키포인트 기반의 이미지 변환 모델을 포함하며,

상기 키포인트 검출 모델을 통해 기준 이미지(Reference Image)로부터 기준 키포인트를 추출하고, 상기 키포인트 검출 모델을 통해 대상 이미지(Target Image)로부터 대상 키포인트를 추출하고,

상기 키포인트 기반의 이미지 변환 모델은 상기 기준 키포인트 및 상기 대상 키포인트 간의 차이를 기반으로 상기 기준 이미지로부터 합성 이미지(Synthesized Image)를 생성하는 것을 특징으로 하는 비디오 생성 장치.

청구항 12

제11항에 있어서,

상기 키포인트 검출 모델은 상기 키포인트 기반의 이미지 변환 모델과 함께 학습되는 것을 특징으로 하는 비디오 생성 장치.

청구항 13

삭제

청구항 14

제11항에 있어서,

상기 키포인트 기반의 이미지 변환 모델은 상기 기준 이미지의 객체의 새로운 외형(Appearance)과 배경 마스크를 이용하여 상기 생성된 합성 이미지의 동적 영역을 처리한 변형 이미지(Translated Image)를 생성하는 것을 특징으로 하는 비디오 생성 장치.

청구항 15

제14항에 있어서,

상기 키포인트 기반의 이미지 변환 모델은 상기 기준 이미지 및 상기 합성 이미지를 혼합(Blend)하여 상기 변형 이미지(Translated Image)를 생성하는 것을 특징으로 하는 비디오 생성 장치.

청구항 16

제14항에 있어서,

상기 키포인트 기반의 이미지 변환 모델은

(i) 상기 기준 이미지에 상기 합성 이미지에 대한 배경 마스크를 적용한 제1 마스크 적용 이미지 및 (ii) 상기 합성 이미지에 상기 배경 마스크를 전환한 마스크를 적용한 제2 마스크 적용 이미지를 혼합(Blend)하여 상기 변형 이미지(Translated Image)를 생성하는 것을 특징으로 하는 비디오 생성 장치.

청구항 17

제14항에 있어서,

상기 키포인트 기반의 이미지 변환 모델과 상호 작용하는 이미지 판별기를 통해 상기 대상 이미지 및 상기 변형 이미지를 비교하여 구별하는 것을 특징으로 하는 비디오 생성 장치.

청구항 18

제11항에 있어서,

상기 키포인트 기반의 모션 생성 모델은 슈도 레이블과 함께 학습되며,

상기 키포인트 기반의 모션 생성 모델은 실제 이미지 시퀀스로부터 상기 키포인트 검출 모델을 통해 추출한 실제 키포인트 시퀀스를 상기 슈도 레이블로 설정하고,

상기 키포인트 기반의 모션 생성 모델과 상호 작용하는 키포인트 시퀀스 판별기를 통해 초기 키포인트로부터 생성한 키포인트 시퀀스 및 상기 실제 키포인트 시퀀스를 비교하여 구별하는 것을 특징으로 하는 비디오 생성 장치.

청구항 19

제18항에 있어서,

상기 키포인트 기반의 모션 생성 모델은 인코더와 디코더로 구성된 오토 인코더이며, 상기 인코더의 입력 레이어가 은닉 레이어에 매핑되고, 상기 은닉 레이어가 상기 디코더의 출력 레이어에 매핑되며,

상기 은닉 레이어에 노이즈가 추가되며,

상기 인코더에 상기 슈도 레이블, 상기 초기 키포인트 및 액션 클래스가 입력되고, 상기 디코더에 상기 초기 키포인트 및 액션 클래스가 입력되는 것을 특징으로 하는 비디오 생성 장치.

청구항 20

제19항에 있어서,

상기 인코더 및 상기 디코더에 은닉 노드가 방향을 가진 엣지로 연결된 순환 구조를 갖고 시계열 데이터를 처리하는 모델이 적용되는 것을 특징으로 하는 비디오 생성 장치.

발명의 설명

기술 분야

[0001] 본 발명이 속하는 기술 분야는 비디오를 생성하는 방법 및 장치에 관한 것이다.

배경 기술

[0002] 이 부분에 기술된 내용은 단순히 본 실시예에 대한 배경 정보를 제공할 뿐 종래기술을 구성하는 것은 아니다.

[0003] 네트워크 기술의 발달 및 서버 확장에 따른 미디어 플랫폼 산업 규모가 증가하는 추세이다. 스마트 디바이스의 보급으로 사용자들은 모바일 인터넷을 통해 다양한 콘텐츠에 언제든지 접근 가능하다. 사용자들은 인공지능을 이용하여 기존 비디오에서 일부 영역을 변경하거나 사진 이미지를 합성하여 가짜 비디오를 생성할 수 있다.

[0004] 비디오 예측은 단일 또는 적은 수의 이미지로부터 미래의 비디오 프레임을 합성하는 기술이다. 장면에서의 동적 움직임의 불확실성으로 인해 자연스러운 비디오를 생성하는 것은 쉽지 않다. 알려지지 않은 미래를 예측하는 것은 비디오 데이터와 실제 세계를 이해하는 데 필수적이고, 기계 학습 분야에서 중요한 기술이다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) US 10430642 (2019.10.01)

(특허문헌 0002) US 10410060 (2019.09.10)

(특허문헌 0003) US 2019-0289321 (2019.09.19)

발명의 내용

해결하려는 과제

[0006] 본 발명의 실시예들은 하나의 이미지에 존재하는 객체가 움직이도록 키포인트 검출 모델, 모션 생성 모델, 및 이미지 변환 모델을 학습하여 비디오를 생성하며, 이미지 내의 객체의 자연스러운 움직임을 구현하는 데 발명의 주된 목적이 있다.

[0007] 본 발명의 명시되지 않은 또 다른 목적들은 하기의 상세한 설명 및 그 효과로부터 용이하게 추론할 수 있는 범위 내에서 추가적으로 고려될 수 있다.

과제의 해결 수단

[0008] 본 실시예의 일 측면에 의하면 컴퓨팅 디바이스에 의한 비디오 생성 방법에 있어서, 하나의 이미지로부터 키포인트 검출 모델을 통해 객체에 대한 키포인트를 추출하는 단계, 키포인트 기반의 모션 생성 모델을 통해 상기 키포인트가 변화된 키포인트 시퀀스를 생성하는 단계, 및 상기 하나의 이미지와 상기 키포인트 시퀀스를 이용하여 키포인트 기반의 이미지 변환 모델을 통해 비디오를 생성하는 단계를 포함하는 비디오 생성 방법을 제공한다.

[0009] 본 실시예의 다른 측면에 의하면, 하나의 이미지로부터 객체에 대한 키포인트를 추출하는 키포인트 검출 모델, 상기 키포인트가 변화된 키포인트 시퀀스를 생성하는 키포인트 기반의 모션 생성 모델, 및 상기 하나의 이미지와 상기 키포인트 시퀀스를 이용하여 비디오를 생성하는 키포인트 기반의 이미지 변환 모델을 포함하는 비디오 생성 장치를 제공한다.

[0010] 본 실시예의 또 다른 측면에 의하면, 프로세서에 의해 실행 가능한 컴퓨터 프로그램 명령어들을 포함하는 비일시적(Non-Transitory) 컴퓨터 판독 가능한 매체에 기록된 컴퓨터 프로그램으로서, 상기 컴퓨터 프로그램 명령어들이 컴퓨팅 디바이스의 프로세서에 의해 실행되는 경우에, 하나의 이미지로부터 키포인트 검출 모델을 통해 객체에 대한 키포인트를 추출하는 단계, 키포인트 기반의 모션 생성 모델을 통해 상기 키포인트가 변화된 키포인트 시퀀스를 생성하는 단계, 및 상기 하나의 이미지와 상기 키포인트 시퀀스를 이용하여 키포인트 기반의 이미지 변환 모델을 통해 비디오를 생성하는 단계를 포함한 동작들을 수행하는 것을 특징으로 하는 컴퓨터 프로그램을 제공한다.

발명의 효과

- [0011] 이상에서 설명한 바와 같이 본 발명의 실시예들에 의하면, 하나의 이미지에 존재하는 객체가 움직이도록 키포인트 검출 모델, 모션 생성 모델, 및 이미지 변환 모델을 학습하여 비디오를 생성하며, 이미지 내의 객체의 자연스러운 움직임을 생성할 수 있는 효과가 있다.
- [0012] 여기에서 명시적으로 언급되지 않은 효과라 하더라도, 본 발명의 기술적 특징에 의해 기대되는 이하의 명세서에서 기재된 효과 및 그 잠정적인 효과는 본 발명의 명세서에 기재된 것과 같이 취급된다.

도면의 간단한 설명

- [0013] 도 1은 본 발명의 일 실시예에 따른 비디오 생성 장치를 예시한 블록도이다.
- 도 2는 본 발명의 일 실시예에 따른 비디오 생성 장치의 키포인트 검출 모델, 키포인트 기반의 모션 생성 모델, 및 키포인트 기반의 이미지 변환 모델을 예시한 도면이다.
- 도 3은 본 발명의 일 실시예에 따른 비디오 생성 장치가 키포인트 검출 모델 및 키포인트 기반의 이미지 변환 모델을 학습하는 동작을 예시한 도면이다.
- 도 4는 본 발명의 일 실시예에 따른 비디오 생성 장치가 배경 마스크를 적용하는 동작을 예시한 도면이다.
- 도 5는 본 발명의 일 실시예에 따른 비디오 생성 장치가 키포인트 기반의 모션 생성 모델을 학습하는 동작을 예시한 도면이다.
- 도 6은 본 발명의 다른 실시예에 따른 비디오 생성 방법을 예시한 흐름도이다.
- 도 7 내지 도 10은 본 발명의 실시예들에 따른 시뮬레이션 결과를 예시한 도면이다.
- 도 11은 본 발명의 실시예들을 실시하는 컴퓨팅 디바이스를 예시한 블록도이다.

발명을 실시하기 위한 구체적인 내용

- [0014] 이하, 본 발명을 설명함에 있어서 관련된 공지기능에 대하여 이 분야의 기술자에게 자명한 사항으로서 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략하고, 본 발명의 일부 실시예들을 예시적인 도면을 통해 상세하게 설명한다.
- [0015] 도 1은 본 발명의 일 실시예에 따른 비디오 생성 장치를 예시한 블록도이다.
- [0016] 도 1에 도시한 바와 같이, 비디오 생성 장치(100)는 키포인트 검출 모델(110), 키포인트 기반의 모션 생성 모델(120), 및 키포인트 기반의 이미지 변환 모델(130)을 포함한다. 도 1에서 예시적으로 도시한 다양한 구성요소들 중에서 일부 구성요소를 생략하거나 다른 구성요소를 추가로 포함할 수 있다. 비디오 생성 장치(100)는 이미지 판별기, 키포인트 시퀀스 판별기, 또는 이들의 조합을 포함할 수 있다.
- [0017] 비디오 생성 장치(100)는 하나의 이미지에 존재하는 객체가 움직이도록 키포인트 검출 모델, 모션 생성 모델, 및 이미지 변환 모델을 학습하여 비디오를 생성하며, 이미지 내의 객체의 자연스러운 움직임을 생성할 수 있다.
- [0018] 비디오 생성 장치(100)는 심층 신경망 학습 방법 중 하나인 적대적 생성 신경망을 구조로 하여 한 장의 객체 사진을 입력으로 받아 해당 물체가 움직이도록 보이는 동영상상을 합성한다.
- [0019] 비디오 생성 장치(100)는 한 장의 객체 사진을 입력받아 해당 객체의 키포인트를 추정한 후, 이 키포인트를 시작 프레임으로 하는 여러 장의 동영상 프레임을 합성한다. 이후 입력 사진의 객체 및 배경 외형을 상기 합성된 여러 장의 키포인트 프레임들과 결합하여 최종 동영상상을 합성한다.
- [0020] 비디오 생성 장치(100)는 레이블된 키포인트 정보 없이 비지도학습으로 다음 프레임을 예측하기 위해 한 개의 비디오에서 임의의 서로 다른 두 장의 프레임을 샘플링하여 그 중 한 프레임을 변환하여 다른 프레임을 만들 수 있도록 학습한다.
- [0021] 비디오 생성 장치(100)는 키포인트 학습을 통해 한 장의 프레임에서 키프레임을 추출할 수 있으며, 키포인트 시퀀스를 생성하기 위해서 비디오에서 학습된 방법으로 모든 프레임의 키프레임을 추출하여 다시 학습 데이터로 사용한다.
- [0022] 도 2는 본 발명의 일 실시예에 따른 비디오 생성 장치의 키포인트 검출 모델, 키포인트 기반의 모션 생성 모델,

및 키포인트 기반의 이미지 변환 모델을 예시한 도면이다.

- [0023] 키포인트 검출 모델은 키포인트 검출기(Keypoint Detector)라고도 한다. 키포인트 검출 모델은 복수의 레이어가 연결된 네트워크로 구현될 수 있다. 키포인트 검출 모델은 네트워크의 파라미터를 학습하여 가중치를 설정할 수 있다.
- [0024] 키포인트 검출 모델은 기준 이미지(Reference Image)로부터 기준 키포인트를 추출하고, 키포인트 검출 모델은 대상 이미지(Target Image)로부터 대상 키포인트를 추출한다. 키포인트 검출 모델은 이미지의 객체로부터 동적 영역 및 동적 특징을 추출하고 동적 특징의 대표값을 키포인트로 추출할 수 있다.
- [0025] 키포인트 기반의 이미지 변환 모델은 변형기(Translator)라고도 한다. 키포인트 기반의 이미지 변환 모델은 복수의 레이어가 연결된 네트워크로 구현될 수 있다. 키포인트 기반의 이미지 변환 모델은 네트워크의 파라미터를 학습하여 가중치를 설정할 수 있다.
- [0026] 키포인트 기반의 이미지 변환 모델은 기준 키포인트 및 대상 키포인트 간의 차이를 기반으로 기준 이미지로부터 합성 이미지(Synthesized Image)를 생성한다.
- [0027] 키포인트 기반의 이미지 변환 모델은 기준 이미지의 객체의 새로운 외형(Appearance)과 배경 마스크를 이용하여 상기 생성된 합성 이미지의 동적 영역을 처리한다. 배경 마스크는 이미지에 대해 설정된 기준치를 기준으로 바이너리 값으로 표현될 수 있다.
- [0028] 키포인트 기반의 이미지 변환 모델은 기준 이미지 및 합성 이미지를 혼합(Blend)하여 변형 이미지(Translated Image)를 생성할 수 있다.
- [0029] 키포인트 기반의 이미지 변환 모델은 (i) 기준 이미지에 상기 합성 이미지에 대한 배경 마스크를 적용한 제1 마스크 적용 이미지 및 (ii) 합성 이미지에 상기 배경 마스크를 전환한 마스크를 적용한 제2 마스크 적용 이미지를 혼합(Blend)하여 변형 이미지(Translated Image)를 생성할 수 있다.
- [0030] 키포인트 기반의 이미지 변환 모델과 상호 작용하는 이미지 판별기를 통해 대상 이미지 및 변형 이미지를 비교하여 구별할 수 있다.
- [0031] 키포인트 기반의 모션 생성 모델은 모션 생성기(Motion Generator)라고도 한다. 키포인트 기반의 모션 생성 모델은 복수의 레이어가 연결된 네트워크로 구현될 수 있다. 키포인트 기반의 모션 생성 모델은 네트워크의 파라미터를 학습하여 가중치를 설정할 수 있다.
- [0032] 키포인트 기반의 모션 생성 모델은 실제 이미지 시퀀스로부터 키포인트 검출 모델을 통해 추출한 실제 키포인트 시퀀스를 슈도 레이블로 설정한다.
- [0033] 키포인트 기반의 모션 생성 모델과 상호 작용하는 키포인트 시퀀스 판별기를 통해 초기 키포인트로부터 생성한 키포인트 시퀀스 및 실제 키포인트 시퀀스를 비교하여 구별한다.
- [0034] 키포인트 기반의 모션 생성 모델은 인코더와 디코더로 구성된 오토 인코더이며, 인코더의 입력 레이어가 은닉 레이어에 매핑되고, 은닉 레이어가 디코더의 출력 레이어에 매핑된다. 은닉 레이어에 노이즈가 추가될 수 있다.
- [0035] 인코더에 슈도 레이블, 초기 키포인트 및 액션 클래스가 입력되고, 디코더에 초기 키포인트 및 액션 클래스가 입력된다. 인코더 및 디코더에 은닉 노드가 방향을 가진 엣지로 연결된 순환 구조를 갖고 시계열 데이터를 처리하는 모델이 적용될 수 있다.
- [0036] 비디오 생성 장치(100)는 크게 두 단계로 학습을 수행한다. 비디오 생성 장치(100)는 키포인트 검출 모델을 키포인트 기반의 이미지 변환 모델과 함께 학습시킨다. 비디오 생성 장치(100)는 키포인트 기반의 모션 생성 모델을 슈도 레이블과 함께 학습시킨다.
- [0037] 도 3은 본 발명의 일 실시예에 따른 비디오 생성 장치가 키포인트 검출 모델 및 키포인트 기반의 이미지 변환 모델을 학습하는 동작을 예시한 도면이다.
- [0038] 키포인트 검출 모델은 동일한 비디오에서 두 프레임 (v , v') 사이의 이미지 변형을 학습하는 방식으로 전경 객체의 키포인트를 검출한다. v 를 v' 에 가깝게 변형하는 학습 방식은 네트워크가 이미지에서 가장 동적인 부분을 자동으로 찾게 하고, 기준 이미지에서 객체를 이동시키는 지침으로 사용할 수 있다.
- [0039] 키포인트 검출 모델은 키포인트와 이미지 간의 유사 관계를 추론하여 대상 이미지를 합성한다. 기준 이미지와

대상 이미지 간의 차이는 검출된 키포인트 세트의 차이에 대응한다.

[0040] 키포인트 검출 모델 Q은 입력 이미지의 키포인트 K를 검출한다. 키포인트 좌표 $\hat{\mathbf{k}} \in \mathbb{R}^{K \times 2}$ 는 특징 맵의 하나에 해당하는 K-채널 바이너리 맵 $\mathbf{l} \in \mathbb{R}^{H \times W \times K}$ 의 좌표를 산출하는 방식으로 획득될 수 있다. 소프트맥스 활성화 함수에 따른 마지막 특징 맵은 수식식 1과 같이 표현될 수 있다.

수식식 1

$$\mathbf{l}^n = \frac{e^{Q(\mathbf{v})^n}}{\sum_{\mathbf{u}} e^{Q(\mathbf{v})^n_{\mathbf{u}}}}$$

$$\hat{\mathbf{k}}^n = \sum_{\mathbf{u}} \mathbf{u} \cdot \mathbf{l}_{\mathbf{u}}^n$$

[0041]

[0042] $\hat{\mathbf{k}}^n$ 는 n 번째 키포인트의 좌표이고, u는 픽셀 좌표이다. 검출된 키포인트 $\hat{\mathbf{k}}$ 는 -1 내지 1의 범위에서 정규화되고, K 가우시안 분포 맵(Gaussian Distribution Map) $\mathbf{d} \in \mathbb{R}^{h \times w \times K}$ 으로 전환된다.

수식식 2

$$\hat{\mathbf{d}}_{\mathbf{u}'}^n = \frac{1}{\sigma \sqrt{2\pi}} e^{-\left(\mathbf{u}' - \hat{\mathbf{k}}^n\right)^2 / 2\sigma^2}$$

[0043]

[0044] σ 는 가우시안 분포의 표준편차이다.

[0045] 도 4는 본 발명의 일 실시예에 따른 비디오 생성 장치가 배경 마스크를 적용하는 동작을 예시한 도면이다.

[0046] 키포인트 기반의 이미지 변환 모델은 새로운 외형과 부드러운 배경 마스크 $\mathbf{m} \in \mathbb{R}^{H \times W \times 1}$ 를 갖는 합성 이미지 $\mathbf{s} \in \mathbb{R}^{H \times W \times 3}$ 를 생성하여 동적 영역을 처리한다. 키포인트 기반의 이미지 변환 모델은 합성 이미지의 동적 영역을 처리한 변형 이미지(Translated Image)를 생성한다. 동적 영역 처리를 통한 변형(Translation)은 객체 변형, 객체 복원(Inpainting), 및 객체 제거(Removal)를 수행한다.

[0047] 키포인트 기반의 이미지 변환 모델 T은 배경 마스크 m을 이용하여 입력 이미지 v 및 합성 이미지 s를 부드럽게 혼합한다.

수식식 3

$$\mathbf{m}, \mathbf{s} = T(\mathbf{v}; \hat{\mathbf{k}}; \hat{\mathbf{k}}')$$

$$\hat{\mathbf{v}} = \mathbf{m} \odot \mathbf{v} + (1 - \mathbf{m}) \odot \mathbf{s}$$

[0048]

[0049] \odot 는 두 텐서의 아다마르 곱(Hadamard Product)을 나타낸다. 키포인트 기반의 이미지 변환 모델은 기준 이미지에 합성 이미지에 대한 배경 마스크를 적용한 제1 마스크 적용 이미지를 생성한다. 합성 이미지에 배경 마스크를 전환한 마스크를 적용한 제2 마스크 적용 이미지를 생성한다. 제1 마스크 적용 이미지와 제2 마스크 적용 이미지를 혼합(Blend)하여 변형 이미지(Translated Image)를 생성한다.

[0050] 키포인트 검출 모델 Q와 키포인트 기반의 이미지 변환 모델 T의 학습 목표 함수는 출력과 대상 이미지 간의 거리로 정의된 복원 손실 함수를 포함하고, 진짜 이미지를 생성하도록 하는 적대적 손실 함수를 포함한다. 본원 손실 함수는 생성된 이미지와 대상 이미지 간의 인지적 유사성을 향상시킨다.

[0051] 비디오 생성 장치는 첫 번째 학습 과정에서 수학적 4의 두 가지 손실 함수를 최적화한다.

수학적 4

$$L_{D_{im}} = -\log D_{im}(\mathbf{v}') - \log(1 - D_{im}(\hat{\mathbf{v}}))$$

$$L_{Q,T} = -\log D_{im}(\hat{\mathbf{v}}) + \lambda_1 \mathbb{E}_l \|\Phi_l(\hat{\mathbf{v}}) - \Phi_l(\mathbf{v}')\|$$

[0052]

[0053] D_{im} 은 이미지 판별기이고, Φ_l 는 이미지 특징 컨볼루션 네트워크의 l 번째 레이어이고, λ_1 는 인지적 손실의 가중치이다.

[0054] 도 5는 본 발명의 일 실시예에 따른 비디오 생성 장치가 키포인트 기반의 모션 생성 모델을 학습하는 동작을 예시한 도면이다.

[0055] 비디오 생성 장치는 키포인트 검출 모델과 이미지 변환 모델의 학습을 마치면, 이미지로부터 키포인트를 검출하고 키포인트 세트로부터 이미지를 생성할 수 있다.

[0056] 비디오 생성 장치는 키포인트 검출 모델을 통해 진짜 비디오에서 검출된 키포인트를 슈도 레이블로 설정한다. 미래의 키포인트 시퀀스를 생성하기 위한 키포인트 기반의 모션 생성 모델을 슈도 레이블과 함께 학습시킨다.

[0057] 키포인트 기반의 모션 생성 모델은 주어진 조건에 따라 미래 이벤트의 분포를 학습하는 오토 인코더를 적용할 수 있다. 키포인트 기반의 모션 생성 모델은 슈도 레이블을 정규화된 분산 잠재 벡터로 인코딩하고

$q_\phi(\mathbf{z}|\hat{\mathbf{k}}_{1:T}, \hat{\mathbf{k}}_0, \mathbf{a})$, 잠재 벡터 \mathbf{z} 를 대응하는 키포인트 시퀀스로 디코딩한다
 $p_\theta(\hat{\mathbf{k}}_{1:T}|\mathbf{z}, \hat{\mathbf{k}}_0, \mathbf{a})$. 시계열 데이터를 처리하기 위해 장단기 메모리(Long Short Term Memory, LSTM) 구조를 인코더와 디코더 모두에 적용할 수 있다. 노이즈 $\mathcal{N}(0, I)$ 로부터 잠재 벡터 \mathbf{z} 를 랜덤 샘플링하여 미래 모션을 예측한다.

[0058] 키포인트 기반의 모션 생성 모델은 쿨백-라이블러 발산(Kullback-Leibler Divergence, KLD)과 복원 손실 함수를 최적화한다.

[0059] 쿨백-라이블러 발산은 두 확률 분포의 차이를 계산하는 데에 사용하며, 이상적인 분포에 대해 그 분포를 근사하는 다른 분포를 사용해 샘플링을 한다면 발생할 수 있는 정보 엔트로피 차이를 계산한다.

[0060] 비디오 생성 장치는 두 번째 학습 과정에서 수학적 5의 두 가지 손실 함수를 최적화한다.

수학식 5

$$L_{D_{seq}} = -\log D_{seq}(\hat{\mathbf{k}}_{1:T}) - \log(1 - D_{seq}(\tilde{\mathbf{k}}_{1:T}))$$

$$L_M = D_{KL}(q_\phi(\mathbf{z}|\hat{\mathbf{k}}_{1:T}; \hat{\mathbf{k}}_0; \mathbf{a})||p_z(\mathbf{z}))$$

$$+ \lambda_2 \|\tilde{\mathbf{k}}_{1:T} - \hat{\mathbf{k}}_{1:T}\|_1 - \lambda_3 \log D_{seq}(\tilde{\mathbf{k}}_{1:T})$$

[0061]

[0062] D_{seq} 는 적대적 손실 함수를 사용하는 키포인트 시퀀스 판별기이다. $\tilde{\mathbf{k}}$ 는 복원된 키포인트이고, λ_2 및 λ_3 는 모델의 하이퍼파라미터이다. 잠재 변수의 분포 $p_z(\mathbf{z})$ 는 노이즈 $\mathcal{N}(0, I)$ 로 설정될 수 있다.

[0063] 도 6은 본 발명의 다른 실시예에 따른 비디오 생성 방법을 예시한 흐름도이다. 비디오 생성 방법은 컴퓨팅 디바이스에 의하여 수행될 수 있으며, 비디오 생성 장치에 의해 수행될 수 있다.

[0064] 단계 S210에서 프로세서는 하나의 이미지로부터 키포인트 검출 모델을 통해 객체에 대한 키포인트를 추출한다.

[0065] 단계 S220에서 프로세서는 키포인트 기반의 모션 생성 모델을 통해 키포인트가 변화된 키포인트 시퀀스를 생성한다.

[0066] 단계 S230에서 프로세서는 하나의 이미지와 키포인트 시퀀스를 이용하여 키포인트 기반의 이미지 변환 모델을 통해 비디오를 생성한다.

[0067] 키포인트 검출 모델은 키포인트 기반의 이미지 변환 모델과 함께 학습된다. 키포인트 검출 모델을 통해 기준 이미지(Reference Image)로부터 기준 키포인트를 추출하고, 키포인트 검출 모델을 통해 대상 이미지(Target Image)로부터 대상 키포인트를 추출한다.

[0068] 키포인트 기반의 이미지 변환 모델은 기준 키포인트 및 대상 키포인트 간의 차이를 기반으로 기준 이미지로부터 합성 이미지(Synthesized Image)를 생성한다.

[0069] 키포인트 기반의 이미지 변환 모델은 기준 이미지의 객체의 새로운 외형(Appearance)과 배경 마스크를 이용하여 상기 생성된 합성 이미지의 동적 영역을 처리한다.

[0070] 키포인트 기반의 이미지 변환 모델은 기준 이미지 및 합성 이미지를 혼합(Blend)하여 변형 이미지(Translated Image)를 생성한다.

[0071] 키포인트 기반의 이미지 변환 모델은 (i) 기준 이미지에 합성 이미지에 대한 배경 마스크를 적용한 제1 마스크 적용 이미지 및 (ii) 합성 이미지에 상기 배경 마스크를 전환한 마스크를 적용한 제2 마스크 적용 이미지를 혼합(Blend)하여 변형 이미지(Translated Image)를 생성한다.

[0072] 키포인트 기반의 이미지 변환 모델과 상호 작용하는 이미지 판별기를 통해 대상 이미지 및 변형 이미지를 비교하여 구별한다.

[0073] 키포인트 기반의 모션 생성 모델은 슈도 레이블과 함께 학습된다. 키포인트 기반의 모션 생성 모델은 실제 이미지 시퀀스로부터 상기 키포인트 검출 모델을 통해 추출한 실제 키포인트 시퀀스를 슈도 레이블로 설정한다.

[0074] 키포인트 기반의 모션 생성 모델과 상호 작용하는 키포인트 시퀀스 판별기를 통해 초기 키포인트로부터 생성한 키포인트 시퀀스 및 실제 키포인트 시퀀스를 비교하여 구별한다.

[0075] 키포인트 기반의 모션 생성 모델은 인코더와 디코더로 구성된 오토 인코더이며, 인코더의 입력 레이어가 은닉 레이어에 매핑되고, 은닉 레이어가 상기 디코더의 출력 레이어에 매핑된다. 은닉 레이어에 추가된 노이즈는 다양한 움직임을 생성하도록 한다.

[0076] 인코더에 슈도 레이블, 초기 키포인트 및 액션 클래스가 입력되고, 디코더에 초기 키포인트 및 액션 클래스가 입력된다. 인코더 및 디코더에 은닉 노드가 방향을 가진 엣지로 연결된 순환 구조를 갖고 시계열 데이터를 처리

하는 모델이 적용될 수 있다.

- [0077] 도 7 내지 도 10은 본 발명의 실시예들에 따른 시뮬레이션 결과를 예시한 도면이다.
- [0078] 종래 기술은 비디오 프레임만으로 구성된 데이터 세트에서 한 장의 사진을 샘플링하여 입력 이미지로 한 후 다중 컨볼루션 연산으로 구성된 심층 신경망을 학습하는 방식이 있고, 비디오 프레임과 비디오 내부 객체의 키포인트를 사람이 직접 레이블링한 데이터 세트에서 한 장의 사진과 그 사진의 레이블링된 키포인트의 프레임을 입력으로 하여 비디오를 생성하는 방식이 있다.
- [0079] 첫 번째 방식은 다중 컨볼루션 연산만으로는 비디오 내부 객체의 움직임을 모델링하는데 한계가 있고, 두 번째 방식은 비디오 모든 프레임에서의 객체 키포인트를 사람이 직접 레이블링 해야 한다는 한계가 있다.
- [0080] 이와 달리 본 실시예에 따른 비디오 생성 방법은 한 장의 객체 사진만을 입력으로 하여, 심층 신경망 내부에서 키포인트 프레임들을 자동으로 생성하고 이를 통해 효과적으로 비디오를 생성한다. 합성된 비디오의 품질이 우수하며 키포인트 레이블이 없는 다양한 데이터 (예컨대, 사람, 동물, 애니메이션 등)에 적용할 수 있다.
- [0081] 도 7은 비디오 생성 장치에 의해 수행된 풀업 동작에 대한 입력 이미지, 검증자료(Ground Truth) 비디오, 마스크 적용 전의 합성 이미지 세트, 마스크가 적용된 합성 이미지 세트, 배경 마스크 세트, 및 키포인트 세트를 도시한다.
- [0082] 도 8은 키포인트 기반의 이미지 변환 모델에 의해 수행된 변형(Translation), 복원(Inpainting), 및 객체 제거(Object Removal)에 대해서 기준 이미지, 대상 이미지, 기준 이미지로부터 검출된 키포인트 세트, 대상 이미지로부터 검출된 키포인트 세트, 배경 마스크, 합성 이미지, 및 최종 변형 결과를 도시한다.
- [0083] 도 9은 키포인트 기반의 이미지 변환 모델에 의해 수행된 변형(Translation), 복원(Inpainting), 및 객체 제거(Object Removal)에 대해서 기준 이미지, 대상 이미지, 기준 이미지로부터 검출된 키포인트 세트, 대상 이미지로부터 검출된 키포인트 세트, 배경 마스크, 합성 이미지, 및 최종 변형 결과를 도시한다.
- [0084] 도 9의 (a) 및 (b)의 결과를 보면, 키포인트 검출 모델은 기준 키포인트 및 대상 키포인트를 적용하여 키포인트 검출 모델의 성능을 향상시킨다. 키포인트 기반의 이미지 변환 모델은 기준 이미지의 키포인트, 대상 이미지의 키포인트, 및 기준 이미지로부터 관계를 추론하여 대상 자세를 갖는 전경 객체를 합성한다. 대상 키포인트를 적용하지 않으면, 이미지 변환 모델은 독립적으로 변형할 지역을 검출해야 하며, 이는 중복되고 비효율적인 설정이다.
- [0085] 도 9의 (c)의 결과를 보면, 키포인트 기반의 이미지 변환 모델은 배경 마스크 생성을 키포인트 유도 이미지 변형(Keypoints-Guided Image Translation)에 통합하는 방식으로, 변형된 이미지의 품질을 크게 향상시킨다. 이미지 변환 모델은 마스크 생성을 통해 이미지의 특정 부분만 변형하고 전경 객체만 합성할 수 있다. 전체 장면을 합성하는 것에 비해 모델링의 복잡성을 감소시킴으로써 네트워크가 이미지 판별기를 속이는 데 유용하다.
- [0086] 도 10을 참조하면, 본 실시예들에 의하면 키포인트 검출, 키포인트 시퀀스 생성, 및 이미지 변형을 통해 하나의 이미지로부터 자연스러운 비디오를 생성할 수 있음을 쉽게 파악할 수 있다.
- [0087] 도 11은 본 발명의 실시예들을 실시하는 컴퓨팅 디바이스를 예시한 블록도이다.
- [0088] 컴퓨팅 디바이스(310)는 적어도 하나의 프로세서(320), 컴퓨터 판독 가능한 저장매체(330) 및 통신 버스(370)를 포함한다.
- [0089] 프로세서(320)는 컴퓨팅 디바이스(310)를 동작하도록 제어할 수 있다. 예컨대, 프로세서(320)는 컴퓨터 판독 가능한 저장 매체(330)에 저장된 하나 이상의 프로그램들을 실행할 수 있다. 하나 이상의 프로그램들은 하나 이상의 컴퓨터 실행 가능 명령어를 포함할 수 있으며, 컴퓨터 실행 가능 명령어는 프로세서(320)에 의해 실행되는 경우 컴퓨팅 디바이스(310)로 하여금 예시적인 실시예에 따른 동작들을 수행하도록 구성될 수 있다.
- [0090] 컴퓨터 판독 가능한 저장 매체(330)는 컴퓨터 실행 가능 명령어 내지 프로그램 코드, 프로그램 데이터 및/또는 다른 적합한 형태의 정보를 저장하도록 구성된다. 컴퓨터 판독 가능한 저장 매체(330)에 저장된 프로그램(340)은 프로세서(320)에 의해 실행 가능한 명령어의 집합을 포함한다. 일 실시예에서, 컴퓨터 판독한 가능 저장 매체(330)는 메모리(랜덤 액세스 메모리와 같은 휘발성 메모리, 비휘발성 메모리, 또는 이들의 적절한 조합), 하나 이상의 자기 디스크 저장 디바이스들, 광학 디스크 저장 디바이스들, 플래시 메모리 디바이스들, 그 밖에 컴퓨팅 디바이스(310)에 의해 액세스되고 원하는 정보를 저장할 수 있는 다른 형태의 저장 매체, 또는 이들의 적합한 조합일 수 있다.

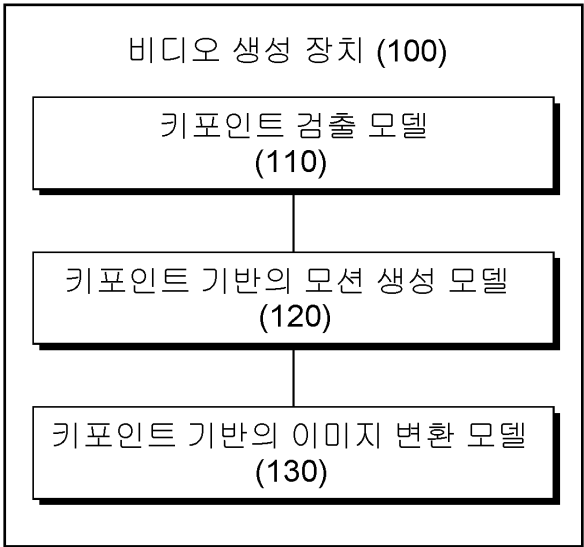
- [0091] 통신 버스(370)는 프로세서(320), 컴퓨터 판독 가능한 저장 매체(340)를 포함하여 컴퓨팅 디바이스(310)의 다른 다양한 컴포넌트들을 상호 연결한다.
- [0092] 컴퓨팅 디바이스(310)는 또한 하나 이상의 입출력 장치를 위한 인터페이스를 제공하는 하나 이상의 입출력 인터페이스(350) 및 하나 이상의 통신 인터페이스(360)를 포함할 수 있다. 입출력 인터페이스(350) 및 통신 인터페이스(360)는 통신 버스(370)에 연결된다. 입출력 장치(미도시)는 입출력 인터페이스(350)를 통해 컴퓨팅 디바이스(310)의 다른 컴포넌트들에 연결될 수 있다.
- [0093] 비디오 생성 장치에 포함된 구성요소들이 도 1에서는 분리되어 도시되어 있으나, 복수의 구성요소들은 상호 결합되어 적어도 하나의 모듈로 구현될 수 있다. 구성요소들은 장치 내부의 소프트웨어적인 모듈 또는 하드웨어적인 모듈을 연결하는 통신 경로에 연결되어 상호 간에 유기적으로 동작한다. 이러한 구성요소들은 하나 이상의 통신 버스 또는 신호선을 이용하여 통신한다.
- [0094] 비디오 생성 장치는 하드웨어, 펌웨어, 소프트웨어 또는 이들의 조합에 의해 로직회로 내에서 구현될 수 있고, 범용 또는 특정 목적 컴퓨터를 이용하여 구현될 수도 있다. 장치는 고정배선형(Hardwired) 기기, 필드 프로그램 가능한 게이트 어레이(Field Programmable Gate Array, FPGA), 주문형 반도체(Application Specific Integrated Circuit, ASIC) 등을 이용하여 구현될 수 있다. 또한, 장치는 하나 이상의 프로세서 및 컨트롤러를 포함한 시스템온칩(System on Chip, SoC)으로 구현될 수 있다.
- [0095] 비디오 생성 장치는 하드웨어적 요소가 마련된 컴퓨팅 디바이스 또는 서버에 소프트웨어, 하드웨어, 또는 이들의 조합하는 형태로 탑재될 수 있다. 컴퓨팅 디바이스 또는 서버는 각종 기기 또는 유무선 통신망과 통신을 수행하기 위한 통신 모듈 등의 통신장치, 프로그램을 실행하기 위한 데이터를 저장하는 메모리, 프로그램을 실행하여 연산 및 명령하기 위한 마이크로프로세서 등을 전부 또는 일부 포함한 다양한 장치를 의미할 수 있다.
- [0096] 도 6에서는 각각의 과정을 순차적으로 실행하는 것으로 기재하고 있으나 이는 예시적으로 설명한 것에 불과하고, 이 분야의 기술자라면 본 발명의 실시예의 본질적인 특성에서 벗어나지 않는 범위에서 도 6에 기재된 순서를 변경하여 실행하거나 또는 하나 이상의 과정을 병렬적으로 실행하거나 다른 과정을 추가하는 것으로 다양하게 수정 및 변형하여 적용 가능할 것이다.
- [0097] 본 실시예들에 따른 동작은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능한 매체에 기록될 수 있다. 컴퓨터 판독 가능한 매체는 실행을 위해 프로세서에 명령어를 제공하는 데 참여한 임의의 매체를 나타낸다. 컴퓨터 판독 가능한 매체는 프로그램 명령, 데이터 파일, 데이터 구조 또는 이들의 조합을 포함할 수 있다. 예를 들면, 자기 매체, 광기록 매체, 메모리 등이 있을 수 있다. 컴퓨터 프로그램은 네트워크로 연결된 컴퓨터 시스템 상에 분산되어 분산 방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수도 있다. 본 실시예를 구현하기 위한 기능적인(Functional) 프로그램, 코드, 및 코드 세그먼트들은 본 실시예가 속하는 기술분야의 프로그래머들에 의해 용이하게 추론될 수 있을 것이다.
- [0098] 본 실시예들은 본 실시예의 기술 사상을 설명하기 위한 것이고, 이러한 실시예에 의하여 본 실시예의 기술 사상의 범위가 한정되는 것은 아니다. 본 실시예의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 실시예의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

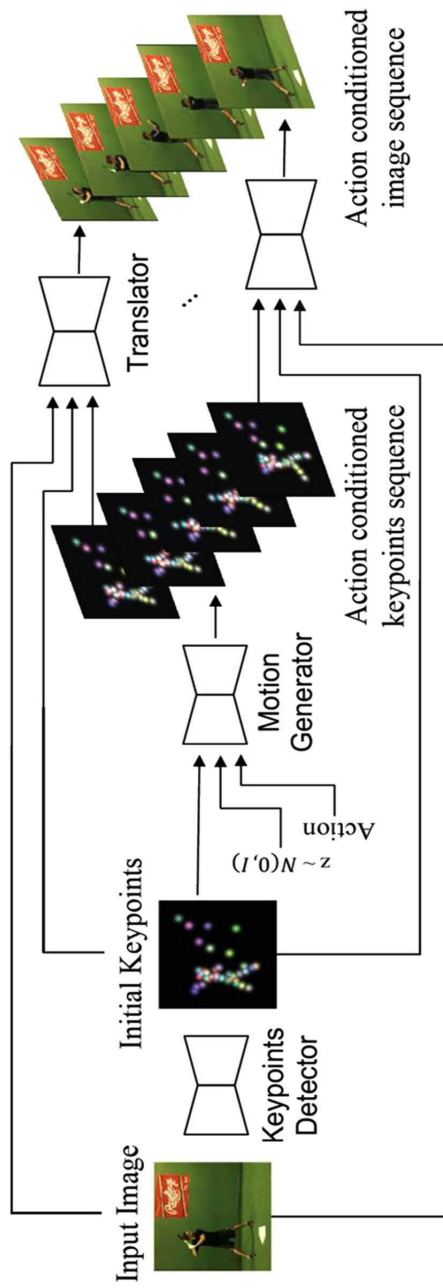
- [0099] 100: 비디오 생성 장치
- 110: 키포인트 검출 모듈
- 120: 키포인트 기반의 모션 생성 모듈
- 130: 키포인트 기반의 이미지 변환 모듈

도면

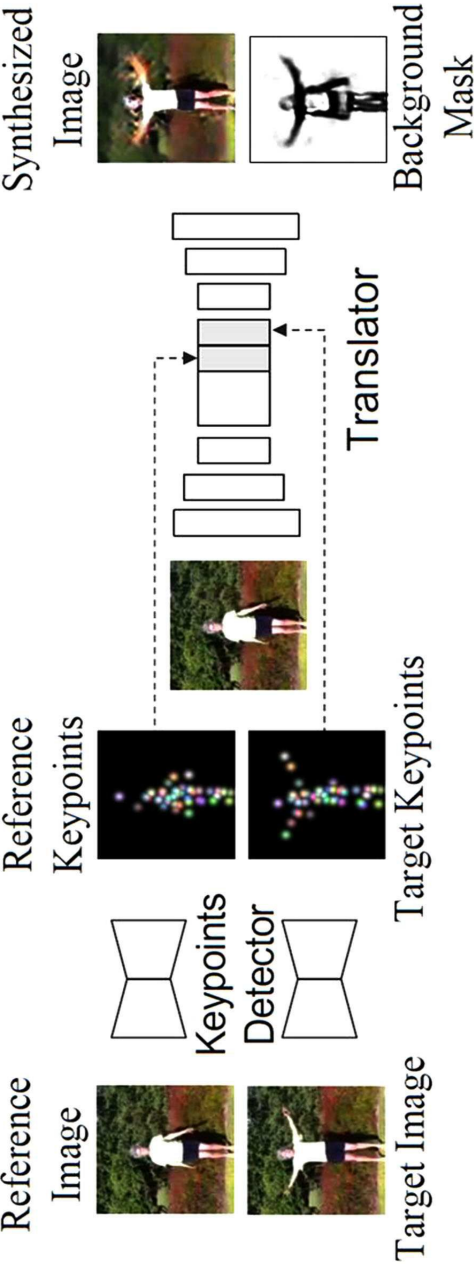
도면1



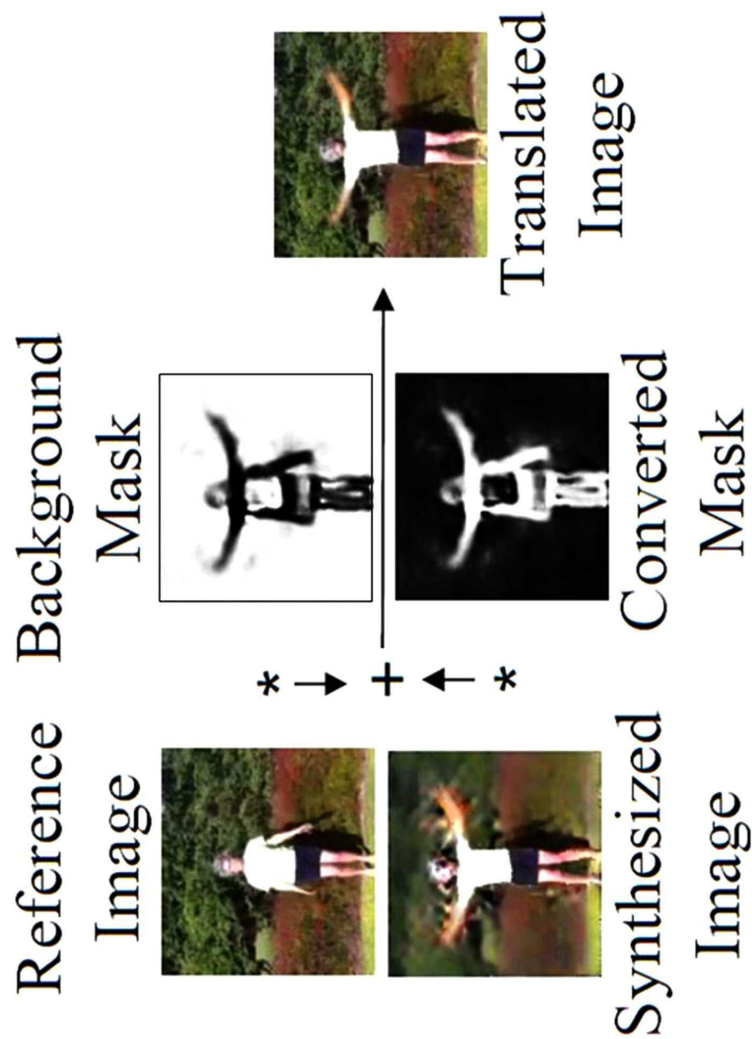
도면2



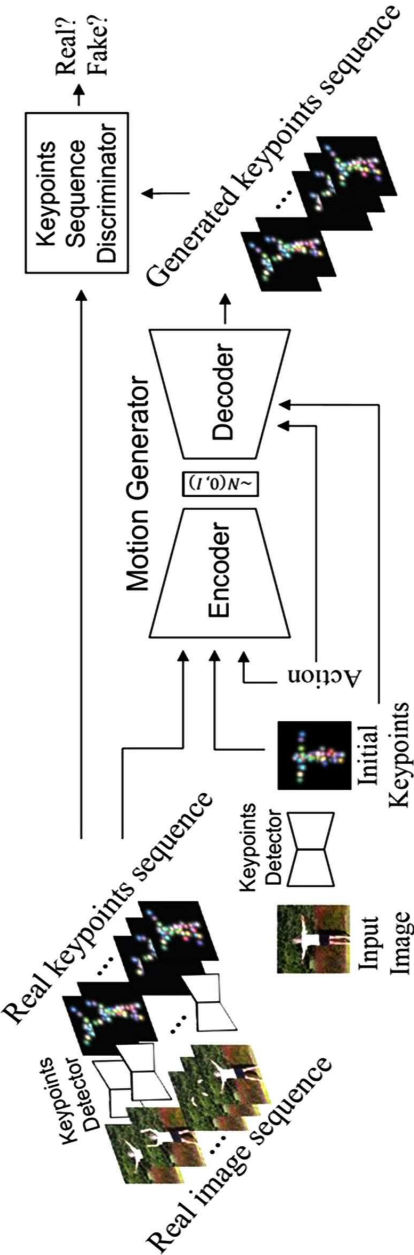
도면3



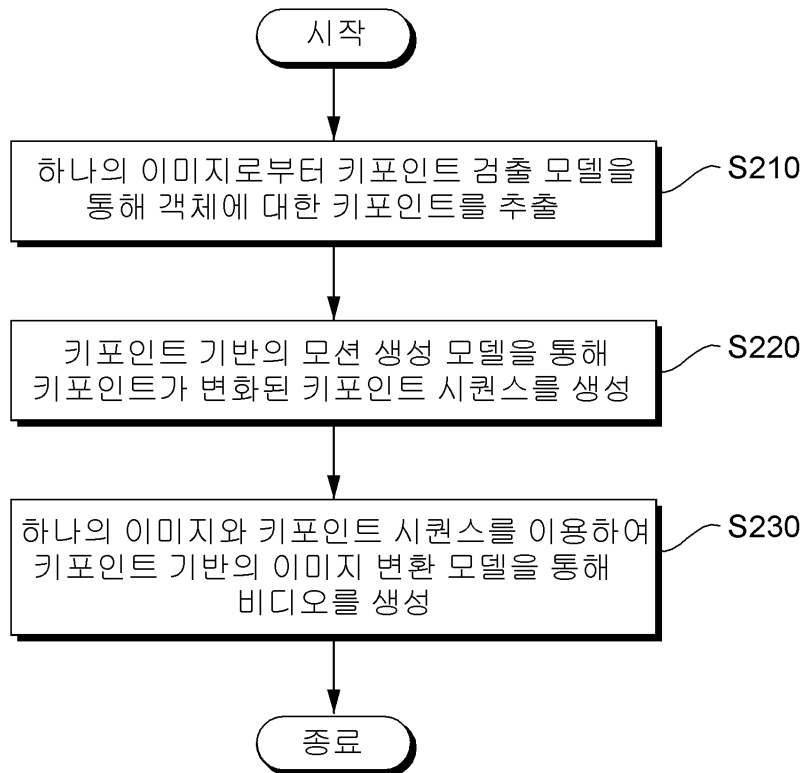
도면4



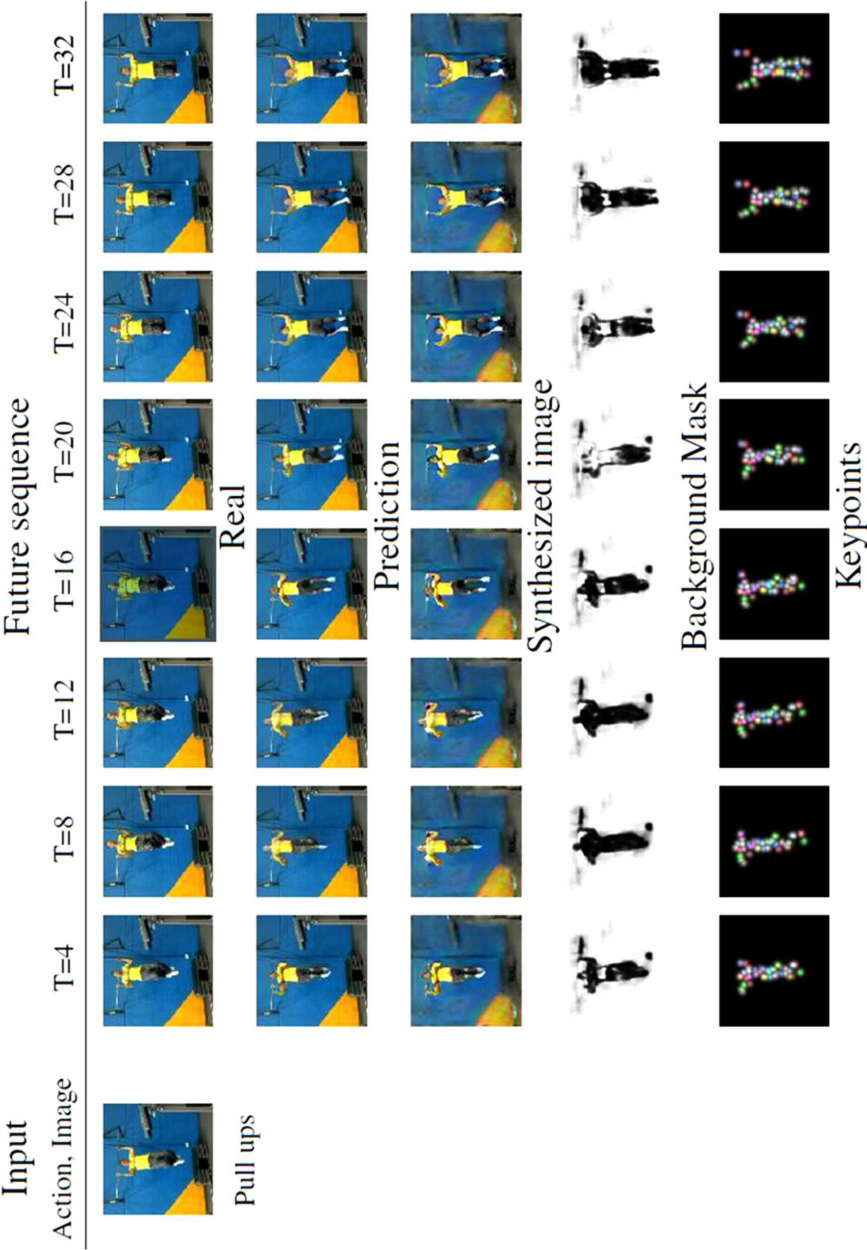
도면5



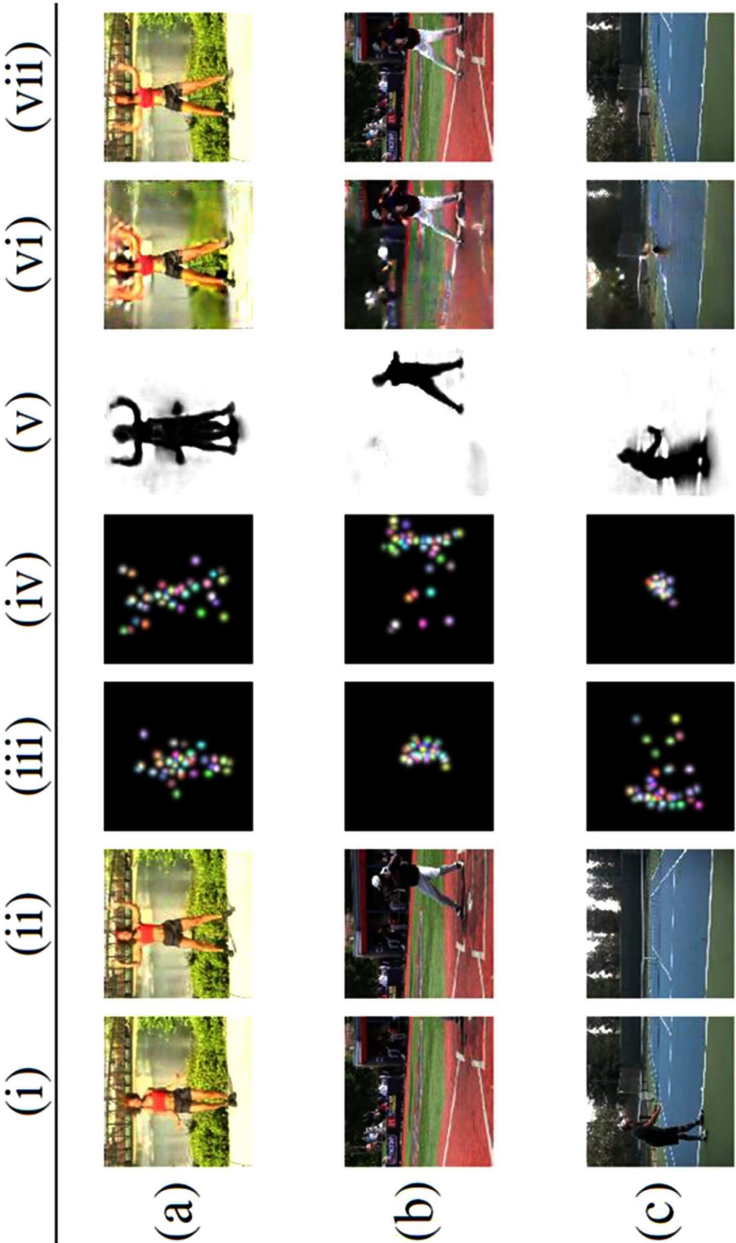
도면6



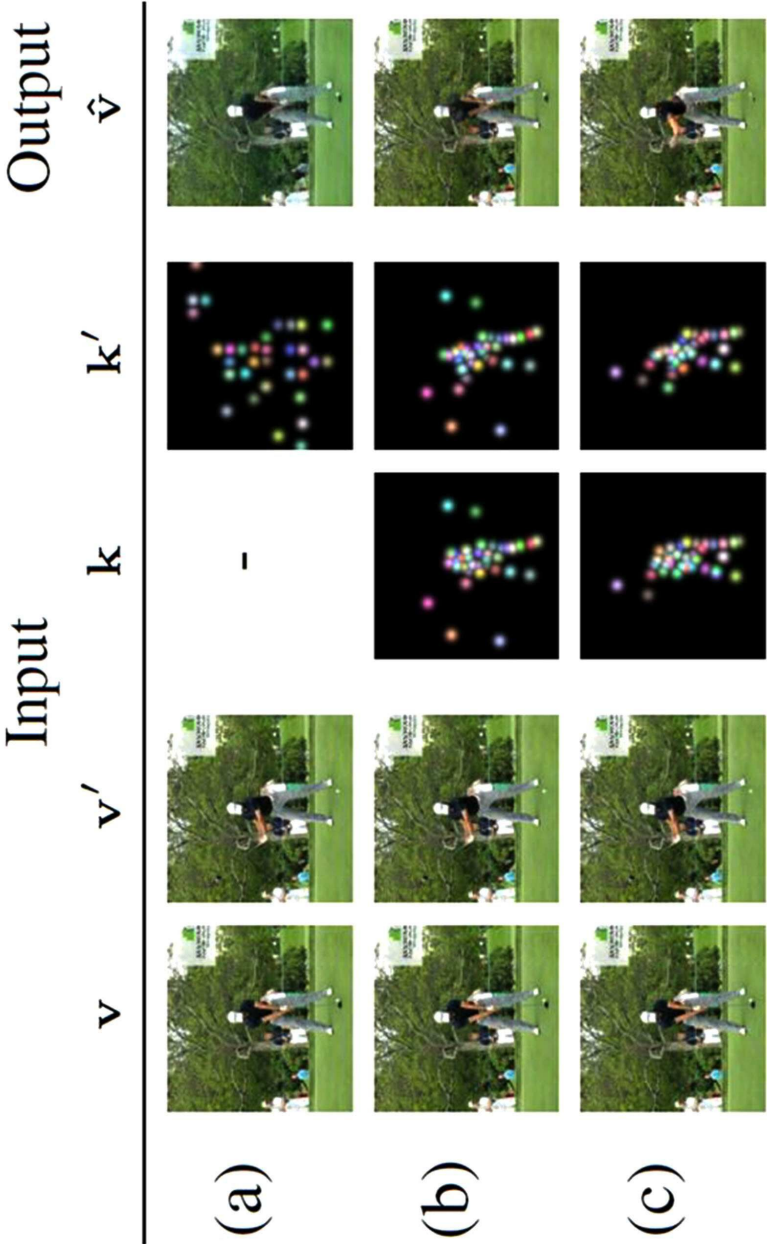
도면7



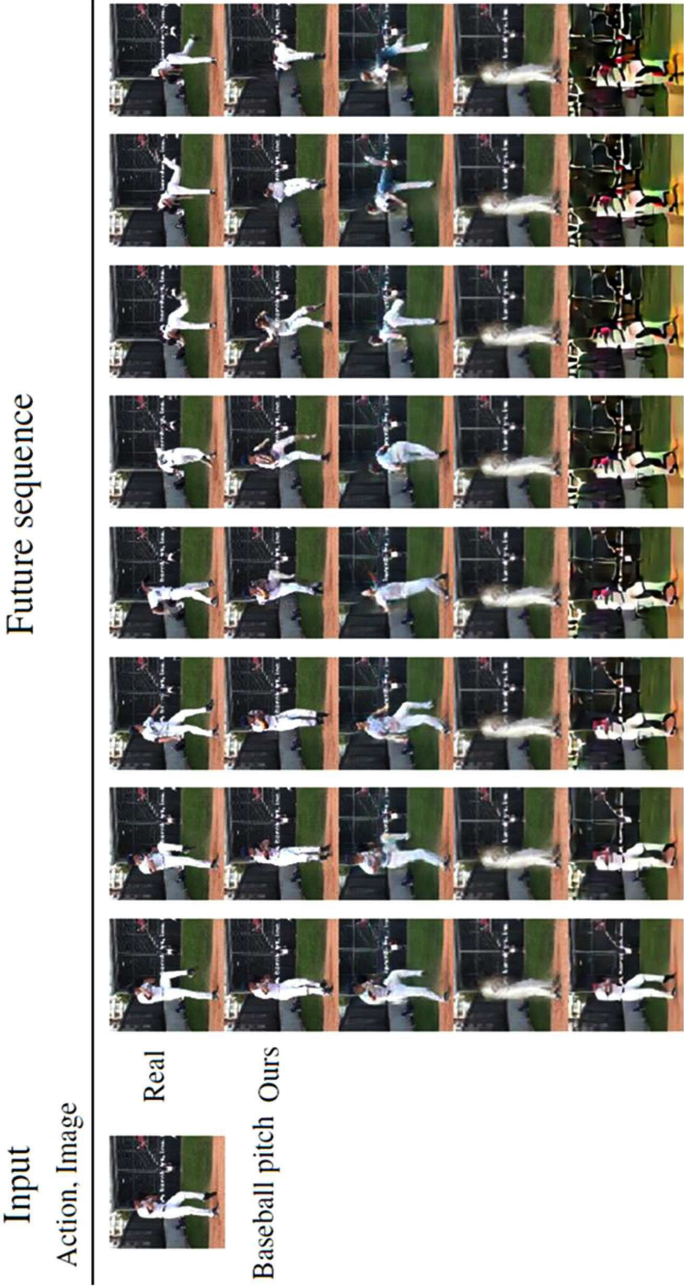
도면8



도면9



도면10



도면11

