



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0074547
(43) 공개일자 2025년05월27일

- | | |
|--|---|
| <p>(51) 국제특허분류(Int. Cl.)
 <i>G06N 3/0475</i> (2023.01) <i>G06N 3/045</i> (2023.01)
 <i>G06N 3/047</i> (2023.01) <i>H04N 21/854</i> (2011.01)</p> <p>(52) CPC특허분류
 <i>G06N 3/0475</i> (2023.01)
 <i>G06N 3/045</i> (2023.01)</p> <p>(21) 출원번호 10-2024-0026308
 (22) 출원일자 2024년02월23일
 심사청구일자 2024년02월23일</p> <p>(30) 우선권주장
 102030160217 2023년11월20일 대한민국(KR)</p> | <p>(71) 출원인
 울산과학기술원
 울산광역시 울주군 언양읍 유니스트길 50</p> <p>(72) 발명자
 김태환
 울산광역시 울주군 언양읍 유니스트길 50
 한성희
 울산광역시 울주군 언양읍 유니스트길 50</p> <p>(74) 대리인
 특허법인 무한</p> |
|--|---|

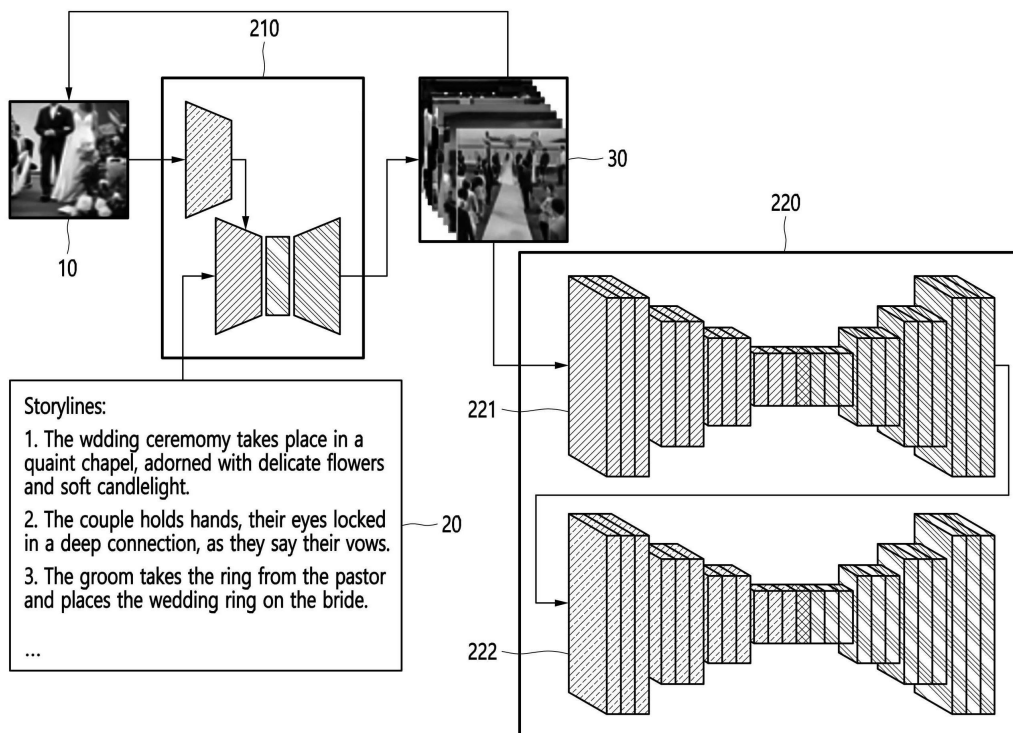
전체 청구항 수 : 총 18 항

(54) 발명의 명칭 **비디오 생성 모델을 이용하여 스토리 기반으로 비디오를 생성하는 방법 및 장치**

(57) 요약

일 실시예에 따른 비디오 생성 모델을 이용하여 스토리 기반으로 비디오를 생성하는 방법 및 장치는 이미지 생성 모델에 기본 이미지와 하나 이상의 스토리 텍스트를 입력(input)함으로써, 입력 이미지들을 생성하는 단계, 2차원 비디오 생성 모델에 상기 생성된 입력 이미지들을 입력함으로써, 각 청크(chunk)에 전체적인 비디오 정보를 (뒷면에 계속)

대표도



보존하는 인터(inter) 항목, 이전 청크에서 생성한 비디오 정보를 보존하는 인트라(intra) 항목 및 해당 청크에서 생성하는 연속성 비디오의 연속된 프레임 중 직전 프레임 정보를 나타내는 콘스(cons) 항목을 포함하는 청크 단위의 초기 연속성 비디오를 생성하는 단계 및 3차원 비디오 생성 모델에 상기 생성된 초기 연속성 비디오를 입력함으로써, 추가적인 모션이 더해져 정제된 수정 연속성 비디오를 생성하는 단계를 포함하고, 상기 청크 단위의 초기 연속성 비디오는, 상기 인터 항목, 상기 인트라 항목, 및 상기 콘스 항목의 퓨전 비율(fusion ratio)이 청크의 생성된 순서에 따라 점진적으로 변경되는 것에 기초하여 생성될 수 있다.

(52) CPC특허분류

G06N 3/047 (2023.01)

H04N 21/854 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193081
과제번호	2022-0-00608-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	사람중심인공지능핵심원천기술개발
연구과제명	(1세부) 인간과 교감하는 멀티모달 인터랙션 인공지능 기술
과제수행기관명	한국전자기술연구원
연구기간	2023.01.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193983
과제번호	2020-0-01336-004
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	인공지능대학원지원(울산과학기술원)
과제수행기관명	울산과학기술원
연구기간	2023.01.01 ~ 2023.12.31

명세서

청구범위

청구항 1

프로세서에 의해 비디오를 생성하는 방법에 있어서,

이미지 생성 모델에 기본 이미지와 하나 이상의 스토리 텍스트를 입력(input)함으로써, 입력 이미지들을 생성하는 단계;

2차원 비디오 생성 모델에 상기 생성된 입력 이미지들을 입력함으로써, 각 청크(chunk)에 전체적인 비디오 정보를 보존하는 인터(inter) 항목, 이전 청크에서 생성한 비디오 정보를 보존하는 인트라(intra) 항목 및 해당 청크에서 생성하는 연속성 비디오의 연속된 프레임 중 직전 프레임 정보를 나타내는 콘스(cons) 항목을 포함하는 청크 단위의 초기 연속성 비디오를 생성하는 단계; 및

3차원 비디오 생성 모델에 상기 생성된 초기 연속성 비디오를 입력함으로써, 추가적인 모션이 더해져 정제된 수정 연속성 비디오를 생성하는 단계

를 포함하고,

상기 청크 단위의 초기 연속성 비디오는,

상기 인터 항목, 상기 인트라 항목, 및 상기 콘스 항목의 퓨전 비율(fusion ratio)이 청크의 생성된 순서에 따라 점진적으로 변경되는 것에 기초하여 생성되는,

방법.

청구항 2

제1항에 있어서,

상기 청크 단위의 초기 연속성 비디오를 생성하는 단계는,

상기 생성된 입력 이미지들로부터 청크 순서에 따른 상기 인터 항목 및 상기 인트라 항목에 대한 이미지 정보를 추출하는 단계;

상기 인터 항목, 상기 인트라 항목 및 상기 콘스 항목에 대한 어텐션(attention) 가중치를 계산하는 단계;

미리 설정된 상기 인터 항목, 상기 인트라 항목 및 상기 콘스 항목에 대한 퓨전(fusion) 비율을 조정하는 단계;

상기 어텐션 가중치와 퓨전 비율에 기초하여 상기 인터 항목, 상기 인트라 항목 및 상기 콘스 항목에 대한 히든(hidden) 값을 산출하는 단계;

상기 인터 항목, 인트라 항목 및 콘스 항목에 대한 히든 값에 기초하여 히든 상태를 결정하는 단계;

상기 결정된 히든 상태에 기초하여 새로운 이미지를 생성하는 단계; 및

상기 새로운 이미지를 인트라 이미지로 교체하는 단계

를 포함하는 방법.

청구항 3

제2항에 있어서,

상기 어텐션 가중치를 계산하는 단계는,

상기 인터 항목의 인터 이미지 정보에 기초하여 쿼리(query), 키(key), 값(value)에 기반한 인터 어텐션

(attention) 가중치를 계산하는 단계;

상기 인트라 항목의 인트라 이미지 정보에 기초하여 쿼리, 키, 값에 기반한 인트라 어텐션 가중치를 계산하는 단계; 및

상기 콘스 항목의 직전 프레임 정보에 기초하여 쿼리, 키, 값에 기반한 콘스 어텐션 가중치를 계산하는 단계를 포함하는 방법.

청구항 4

제2항에 있어서,

상기 퓨전 비율을 조정하는 단계는,

시간에 따라 상기 콘스 항목에 대한 퓨전 비율을 조정하는 단계; 및

상기 콘스 항목에 대한 퓨전 비율에 기초하여 상기 인트라 항목과 상기 인터 항목에 대한 퓨전 비율을 조정하는 단계

를 포함하는 방법.

청구항 5

제1항에 있어서,

상기 입력 이미지들을 생성하는 단계는,

상기 이미지 생성 모델을 통해 기본 이미지와 제1 스토리 텍스트에 현재 이미지 프레임을 생성하는 단계;

상기 현재 이미지 프레임과 상기 제1 스토리 텍스트에 기초하여 다음 이미지 프레임을 생성하는 단계; 및

상기 다음 이미지 프레임을 상기 현재 이미지 프레임으로 업데이트하여 각 스토리 텍스트 별로 미리 결정된 개수의 이미지 프레임들을 생성하여 상기 입력 이미지들을 구성하는 단계

를 포함하는 방법.

청구항 6

제1항에 있어서,

상기 수정 연속성 비디오를 생성하는 단계는,

상기 3차원 비디오 생성 모델의 디퓨전(diffusion) 모델을 이용하여 상기 생성된 초기 연속성 비디오에 노이즈를 더하는 단계; 및

확률적 미분 방정식을 이용하여 디노이징(denoising)함에 따라 상기 노이즈가 더해진 초기 연속성 비디오를 정제하는 단계

를 포함하는 방법.

청구항 7

제1항에 있어서,

상기 인터 항목은 상기 연속성의 비디오의 첫번째 프레임에 대한 정보를 포함하고,

상기 인트라 항목은 이전 청크의 마지막 프레임에 대한 정보를 포함하는,

방법.

청구항 8

제1항에 있어서,
상기 비디오 정보는,
비디오의 테마, 장소, 배경 및 등장인물 중 하나 이상을 포함하는,
방법.

청구항 9

제1항에 있어서,
상기 스토리 텍스트는 하나의 청크에 대한 스토리 텍스트가 이어지는 다음 청크에 대한 스토리 텍스트와 일부 중복되는,
방법.

청구항 10

이미지 생성 모델에 기본 이미지와 하나 이상의 스토리 텍스트를 입력함으로써, 입력 이미지들을 생성하고, 2차원 비디오 생성 모델에 상기 생성된 입력 이미지들을 입력함으로써, 각 청크에 전체적인 비디오 정보를 보존하는 인터 항목, 이전 청크에서 생성한 비디오 정보를 보존하는 인트라 항목 및 해당 청크에서 생성하는 연속성 비디오의 연속된 프레임 중 직전 프레임 정보를 나타내는 콘스 항목을 포함하는 청크 단위의 초기 연속성 비디오를 생성하며, 3차원 비디오 생성 모델에 상기 생성된 초기 연속성 비디오를 입력함으로써, 추가적인 모션이 더해져 정제된 수정 연속성 비디오를 생성하는 프로세서

를 포함하고,
상기 청크 단위의 초기 연속성 비디오는,
상기 인터 항목, 상기 인트라 항목, 및 상기 콘스 항목의 퓨전 비율이 청크의 생성된 순서에 따라 점진적으로 변경되는 것에 기초하여 생성되는,
장치.

청구항 11

제10항에 있어서,
상기 프로세서는,
상기 청크 단위의 초기 연속성 비디오를 생성함에 있어서, 상기 2차원 비디오 생성 모델을 이용하여 상기 생성된 입력 이미지들로부터 청크 순서에 따른 상기 인터 항목 및 상기 인트라 항목에 대한 이미지 정보를 추출하고, 상기 인터 항목, 상기 인트라 항목 및 상기 콘스 항목에 대한 어텐션 가중치를 계산하고, 미리 설정된 상기 인터 항목, 상기 인트라 항목 및 상기 콘스 항목에 대한 퓨전 비율을 조정하고, 상기 어텐션 가중치와 퓨전 비율에 기초하여 상기 인터 항목, 상기 인트라 항목 및 상기 콘스 항목에 대한 히든 값을 산출하고, 상기 인터 항목, 인트라 항목 및 콘스 항목에 대한 히든 값에 기초하여 히든 상태를 결정하고, 상기 결정된 히든 상태에 기초하여 새로운 이미지를 생성하며, 상기 새로운 이미지를 인트라 이미지로 교체하는,
장치.

청구항 12

제11항에 있어서,

상기 프로세서는,

상기 어텐션 가중치를 계산함에 있어서, 상기 인터 항목의 인터 이미지 정보에 기초하여 쿼리, 키, 값에 기반한 인터 어텐션 가중치를 계산하고, 상기 인트라 항목의 인트라 이미지 정보에 기초하여 쿼리, 키, 값에 기반한 인트라 어텐션 가중치를 계산하며, 상기 콘스 항목의 직전 프레임 정보에 기초하여 쿼리, 키, 값에 기반한 콘스 어텐션 가중치를 계산하는,

장치.

청구항 13

제11항에 있어서,

상기 프로세서는,

상기 퓨전 비율을 조정함에 있어서, 시간에 따라 상기 콘스 항목에 대한 퓨전 비율을 조정하며, 상기 콘스 항목에 대한 퓨전 비율에 기초하여 상기 인트라 항목과 상기 인터 항목에 대한 퓨전 비율을 조정하는,

장치.

청구항 14

제10항에 있어서,

상기 프로세서는,

상기 입력 이미지들을 생성함에 있어서, 상기 이미지 생성 모델을 통해 기본 이미지와 제1 스토리 텍스트에 현재 이미지 프레임을 생성하고, 상기 현재 이미지 프레임과 상기 제1 스토리 텍스트에 기초하여 다음 이미지 프레임을 생성하며, 상기 다음 이미지 프레임을 상기 현재 이미지 프레임으로 업데이트하여 각 스토리 텍스트 별로 미리 결정된 개수의 이미지 프레임들을 생성하여 상기 입력 이미지들을 구성하는,

장치.

청구항 15

제10항에 있어서,

상기 프로세서는,

상기 수정 연속성 비디오를 생성함에 있어서, 상기 3차원 비디오 생성 모델의 디퓨전 모델을 이용하여 상기 생성된 초기 연속성 비디오에 노이즈를 더하고, 상기 노이즈가 더해진 초기 연속성 비디오를 확률적 미분 방정식을 이용하여 디노이징함에 따라 정제하는,

장치.

청구항 16

제10항에 있어서,

상기 인터 항목은 상기 연속성의 비디오의 첫번째 프레임에 대한 정보를 포함하고,

상기 인트라 항목은 이전 청크의 마지막 프레임에 대한 정보를 포함하는,

장치.

청구항 17

제10항에 있어서,
상기 비디오 정보는,
비디오의 테마, 장소, 배경 및 등장인물 중 하나 이상을 포함하는,
장치.

청구항 18

제10항에 있어서,
상기 스토리 텍스트는 하나의 청크에 대한 스토리 텍스트가 이어지는 다음 청크에 대한 스토리 텍스트와 일부 중복되는,
장치.

발명의 설명

기술 분야

[0001] 이하, 비디오 생성 모델을 이용하여 하나 이상의 스토리 텍스트로부터 연속성의 비디오를 생성하는 것에 관한 기술이 제공된다.

배경 기술

[0003] 생성형(Generative) AI는 대규모 데이터 세트를 기반으로 트레이닝된 딥러닝 모델을 사용하여 새로운 콘텐츠를 생성하는 일종의 인공지능 기술을 의미한다. 생성형 AI를 통해 생성되는 새로운 콘텐츠는 대화(chat), 이야기(story), 이미지(image), 비디오(video), 음악(music) 등을 포함할 수 있다. 생성형 AI는 방대한 경험 데이터 베이스에 인공지능을 이용하여 새롭고 독창적인 정보를 생성할 수 있어, 금융 서비스, 의료와 생명 과학, 자동차와 제조, 미디어와 엔터테인먼트, 전기통신 및 에너지 등 다양한 분야에 적용될 수 있다. 예를 들어, 생성형 AI는 약물의 효과를 최적화한 화합물을 생성하여 신약을 개발할 수 있고, 이미지 또는 텍스트 입력을 기반으로 이미지를 생성할 수 있다.

[0004] 위에서 설명한 배경기술은 발명자가 본원의 개시 내용을 도출하는 과정에서 보유하거나 습득한 것으로서, 반드시 본 출원 전에 일반 공중에 공개된 공지기술이라고 할 수는 없다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 한국 공개 특허공보 10-2023-0072454(2023년 05월 24일 공고)에는 이미지 텍스트 양방향 생성 장치, 방법 및 프로그램이 제시된다.

(특허문헌 0002) 한국 등록 특허공보 10-2192015(2020년 12월 10일 공고)에는 자연어 문장을 기반으로 하는 변환 이미지 생성 장치, 방법 및 이를 이용하는 변환 이미지 생성 시스템이 제시된다.

발명의 내용

해결하려는 과제

- [0007] 일 실시예에 따른 비디오 생성 모델을 이용하여 스토리 기반으로 비디오를 생성하는 방법 및 장치는 스토리에 기초하여 청크 단위로 비디오를 생성할 수 있다.
- [0008] 일 실시예에 따른 비디오 생성 모델을 이용하여 스토리 기반으로 비디오를 생성하는 방법 및 장치는 하나의 청크에 대한 스토리 텍스트가 이어지는 다음 청크에 대한 스토리 텍스트와 일부 중복되게 대응될 수 있다.
- [0009] 다만, 기술적 과제는 상술한 기술적 과제들로 한정되는 것은 아니며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

- [0011] 일 실시예에 따른 비디오 생성 모델을 이용하여 스토리 기반으로 비디오를 생성하는 방법은 이미지 생성 모델에 기본 이미지와 하나 이상의 스토리 텍스트를 입력(input)함으로써, 입력 이미지들을 생성하는 단계, 2차원 비디오 생성 모델에 상기 생성된 입력 이미지들을 입력함으로써, 각 청크(chunk)에 전체적인 비디오 정보를 보존하는 인터(inter) 항목, 이전 청크에서 생성한 비디오 정보를 보존하는 인트라(intra) 항목 및 해당 청크에서 생성하는 연속성 비디오의 연속된 프레임 중 직전 프레임 정보를 나타내는 콘스(cons)항목을 포함하는 청크 단위의 초기 연속성 비디오를 생성하는 단계 및 3차원 비디오 생성 모델에 상기 생성된 초기 연속성 비디오를 입력함으로써, 추가적인 모션이 더해져 정제된 수정 연속성 비디오를 생성하는 단계를 포함하고, 상기 청크 단위의 초기 연속성 비디오는, 상기 인터 항목, 상기 인트라 항목, 및 상기 콘스 항목의 퓨전 비율(fusion ratio)이 청크의 생성된 순서에 따라 점진적으로 변경되는 것에 기초하여 생성될 수 있다.
- [0012] 상기 청크 단위의 초기 연속성 비디오를 생성하는 단계는, 상기 생성된 입력 이미지들로부터 청크 순서에 따른 상기 인터 항목 및 상기 인트라 항목에 대한 이미지 정보를 추출하는 단계, 상기 인터 항목, 상기 인트라 항목 및 상기 콘스 항목에 대한 어텐션(attention) 가중치를 계산하는 단계, 미리 설정된 상기 인터 항목, 상기 인트라 항목 및 상기 콘스 항목에 대한 퓨전(fusion) 비율을 조정하는 단계, 상기 어텐션 가중치와 퓨전 비율에 기초하여 상기 인터 항목, 상기 인트라 항목 및 상기 콘스 항목에 대한 히든(hidden) 값을 산출하는 단계, 상기 인터 항목, 인트라 항목 및 콘스 항목에 대한 히든 값에 기초하여 히든 상태를 결정하는 단계, 상기 결정된 히든 상태에 기초하여 새로운 이미지를 생성하는 단계 및 상기 새로운 이미지를 인트라 이미지로 교체하는 단계를 포함할 수 있다.
- [0013] 상기 어텐션 가중치를 계산하는 단계는, 상기 인터 항목의 인터 이미지 정보에 기초하여 쿼리(query), 키(key), 값(value)에 기반한 인터 어텐션(attention) 가중치를 계산하는 단계, 상기 인트라 항목의 인트라 이미지 정보에 기초하여 쿼리, 키, 값에 기반한 인트라 어텐션 가중치를 계산하는 단계 및 상기 콘스 항목의 직전 프레임 정보에 기초하여 쿼리, 키, 값에 기반한 콘스 어텐션 가중치를 계산하는 단계를 포함할 수 있다.
- [0014] 상기 퓨전 비율을 조정하는 단계는, 시간에 따라 콘스 항목에 대한 퓨전 비율을 조정하는 단계 및 상기 콘스 항목에 대한 퓨전 비율에 기초하여 상기 인트라 항목과 상기 인터 항목에 대한 퓨전 비율을 조정하는 단계를 포함할 수 있다.
- [0015] 상기 입력 이미지들을 생성하는 단계는, 상기 이미지 생성 모델을 통해 기본 이미지와 제1 스토리 텍스트에 현재 이미지 프레임을 생성하는 단계, 상기 현재 이미지 프레임과 상기 제1 스토리 텍스트에 기초하여 다음 이미지 프레임을 생성하는 단계 및 상기 다음 이미지 프레임을 상기 현재 이미지 프레임으로 업데이트하여 각 스토리 텍스트 별로 미리 결정된 개수의 이미지 프레임들을 생성하여 상기 입력 이미지들을 구성하는 단계를 포함할 수 있다.
- [0016] 상기 수정 연속성 비디오를 생성하는 단계는, 상기 3차원 비디오 생성 모델의 디퓨전(diffusion) 모델을 이용하여 상기 생성된 초기 연속성 비디오에 노이즈를 더하는 단계 및 확률적 미분 방정식을 이용하여 디노이징(denoising)함에 따라 상기 노이즈가 더해진 초기 연속성 비디오를 정제하는 단계를 포함할 수 있다.
- [0017] 상기 인터 항목은 상기 연속성의 비디오의 첫번째 프레임에 대한 정보를 포함하고, 상기 인트라 항목은 이전 청크의 마지막 프레임에 대한 정보를 포함할 수 있다.
- [0018] 상기 비디오 정보는, 비디오의 테마, 장소, 배경 및 등장인물 중 하나 이상을 포함할 수 있다.
- [0019] 상기 스토리 텍스트는 하나의 청크에 대한 스토리 텍스트가 이어지는 다음 청크에 대한 스토리 텍스트와 일부

중복될 수 있다.

- [0020] 일 실시예에 따른 비디오 생성 모델을 이용하여 스토리 기반으로 비디오를 생성하는 장치는 이미지 생성 모델에 기본 이미지와 하나 이상의 스토리 텍스트를 입력함으로써, 입력 이미지들을 생성하고, 2차원 비디오 생성 모델에 상기 생성된 입력 이미지들을 입력함으로써, 각 청크에 전체적인 비디오 정보를 보존하는 인터 항목, 이전 청크에서 생성한 비디오 정보를 보존하는 인트라 항목 및 해당 청크에서 생성하는 연속성 비디오의 연속된 프레임 중 직전 프레임 정보를 나타내는 콘스 항목을 포함하는 청크 단위의 초기 연속성 비디오를 생성하며, 3차원 비디오 생성 모델에 상기 생성된 초기 연속성 비디오를 입력함으로써, 추가적인 모션이 더해져 정제된 수정 연속성 비디오를 생성하는 프로세서를 포함하고, 상기 청크 단위의 초기 연속성 비디오는, 인터 항목, 인트라 항목, 및 콘스 항목의 퓨전 비율이 청크의 생성된 순서에 따라 점진적으로 변경되는 것에 기초하여 생성될 수 있다.
- [0021] 상기 프로세서는, 상기 청크 단위의 초기 연속성 비디오를 생성함에 있어서, 상기 2차원 비디오 생성 모델을 이용하여 상기 생성된 입력 이미지들로부터 청크 순서에 따른 상기 인터 항목 및 상기 인트라 항목에 대한 이미지 정보를 추출하고, 상기 인터 항목, 상기 인트라 항목 및 상기 콘스 항목에 대한 어텐션 가중치를 계산하고, 미리 설정된 상기 인터 항목, 상기 인트라 항목 및 상기 콘스 항목에 대한 퓨전 비율을 조정하고, 상기 어텐션 가중치와 퓨전 비율에 기초하여 상기 인터 항목, 상기 인트라 항목 및 상기 콘스 항목에 대한 히든 값을 산출하고, 상기 인터 항목, 인트라 항목 및 콘스 항목에 대한 히든 값에 기초하여 히든 상태를 결정하고, 상기 결정된 히든 상태에 기초하여 새로운 이미지를 생성하며, 상기 새로운 이미지를 인트라 이미지로 교체할 수 있다.
- [0022] 상기 프로세서는, 상기 어텐션 가중치를 계산함에 있어서, 상기 인터 항목의 인터 이미지 정보에 기초하여 쿼리, 키, 값에 기반한 인터 어텐션 가중치를 계산하고, 상기 인트라 항목의 인트라 이미지 정보에 기초하여 쿼리, 키, 값에 기반한 인트라 어텐션 가중치를 계산하며, 상기 콘스 항목의 직전 프레임 정보에 기초하여 쿼리, 키, 값에 기반한 콘스 어텐션 가중치를 계산할 수 있다.
- [0023] 상기 프로세서는, 상기 퓨전 비율을 조정함에 있어서, 시간에 따라 콘스 항목에 대한 퓨전 비율을 조정하며, 상기 콘스 항목에 대한 퓨전 비율에 기초하여 상기 인트라 항목과 상기 인터 항목에 대한 퓨전 비율을 조정할 수 있다.
- [0024] 상기 프로세서는, 상기 입력 이미지들을 생성함에 있어서, 상기 이미지 생성 모델을 통해 기본 이미지와 제1 스토리 텍스트에 현재 이미지 프레임을 생성하고, 상기 현재 이미지 프레임과 상기 제1 스토리 텍스트에 기초하여 다음 이미지 프레임을 생성하며, 상기 다음 이미지 프레임을 상기 현재 이미지 프레임으로 업데이트하여 각 스토리 텍스트 별로 미리 결정된 개수의 이미지 프레임들을 생성하여 상기 입력 이미지들을 구성할 수 있다.
- [0025] 상기 프로세서는, 상기 수정 연속성 비디오를 생성함에 있어서, 상기 3차원 비디오 생성 모델의 디퓨전 모델을 이용하여 상기 생성된 초기 연속성 비디오에 노이즈를 더하고, 확률적 미분 방정식을 이용하여 디노이징함에 따라 상기 노이즈가 더해진 초기 연속성 비디오를 정제할 수 있다.
- [0026] 상기 인터 항목은 상기 연속성의 비디오의 첫번째 프레임에 대한 정보를 포함하고, 상기 인트라 항목은 이전 청크의 마지막 프레임에 대한 정보를 포함할 수 있다.
- [0027] 상기 비디오 정보는, 비디오의 테마, 장소, 배경 및 등장인물 중 하나 이상을 포함할 수 있다.
- [0028] 상기 스토리 텍스트는 하나의 청크에 대한 스토리 텍스트가 이어지는 다음 청크에 대한 스토리 텍스트와 일부 중복될 수 있다.

발명의 효과

- [0030] 일 실시예에 따른 비디오 생성 모델을 이용하여 스토리 기반으로 비디오를 생성하는 방법 및 장치는 스토리에 기초하여 청크 단위로 비디오를 생성함에 따라 GPU에 영향을 받지 않고 이어지는 배경 및/또는 등장인물을 포함하는 무제한의 길이를 가진 비디오를 계속 이어가며 생성할 수 있다.
- [0031] 일 실시예에 따른 비디오 생성 모델을 이용하여 스토리 기반으로 비디오를 생성하는 방법 및 장치는 하나의 청크에 대한 스토리 텍스트가 이어지는 다음 청크에 대한 스토리 텍스트와 일부 중복되게 대응됨에 따라 프로세서가 배경과 등장인물이 일관된 연속적인 비디오를 생성할 수 있다.

도면의 간단한 설명

- [0033] 도 1은 일 실시예에 따른 비디오 생성 모델을 이용하여 비디오를 생성하는 장치의 블록도를 도시한다.
- 도 2는 일 실시예에 따른 비디오 생성 모델의 아키텍처를 예시적으로 나타낸다.
- 도 3은 일 실시예에 따른 비디오 생성 모델을 이용하여 비디오를 생성하는 장치에서 생성된 입력 이미지들을 예시적으로 나타낸다.
- 도 4는 일 실시예에 따른 비디오 생성 모델의 아키텍처 중 2차원 비디오 생성 모델의 구조를 예시적으로 나타낸다.
- 도 5(a)는 일 실시예에 따른 비디오 생성 모델을 이용하여 비디오를 생성하는 장치에서 생성된 청크의 시계열적 구조를 예시적으로 나타내고, 도 5(b)는 일 실시예에 따른 비디오 생성 모델을 이용하여 비디오를 생성하는 장치에서 생성된 청크의 시계열적 퓨전 비율의 그래프를 예시적으로 나타낸다.
- 도 6은 일 실시예에 따른 비디오 생성 모델을 이용하여 비디오를 생성하는 장치에서 생성된 청크에 스토리 텍스트가 대응된 구조를 예시적으로 나타낸다.
- 도 7은 일 실시예에 따른 비디오 생성 모델을 이용하여 비디오를 생성하는 장치에서 생성된 초기 연속성 비디오와 수정 연속성 비디오를 예시적으로 나타낸다.
- 도 8은 일 실시예에 따른 비디오 생성 모델을 이용하여 비디오를 생성하는 방법의 흐름도를 도시한다.
- 도 9는 일 실시예에 따른 비디오 생성 모델을 이용하여 비디오를 생성하는 방법에서 생성된 수정 연속성 비디오와 이전의 비디오 생성 모델을 이용하여 생성된 비디오를 예시적으로 나타낸다.

발명을 실시하기 위한 구체적인 내용

- [0034] 실시예들에 대한 특정한 구조적 또는 기능적 설명들은 단지 예시를 위한 목적으로 개시된 것으로서, 다양한 형태로 변경되어 구현될 수 있다. 따라서, 실제 구현되는 형태는 개시된 특정 실시예로만 한정되는 것이 아니며, 본 명세서의 범위는 실시예들로 설명한 기술적 사상에 포함되는 변경, 균등물, 또는 대체물을 포함한다.
- [0035] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 이런 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 해석되어야 한다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.
- [0036] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다.
- [0037] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설명된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함으로써 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0038] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 해당 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.
- [0039] 이하, 실시예들을 첨부된 도면들을 참조하여 상세하게 설명한다. 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조 부호를 부여하고, 이에 대한 중복되는 설명은 생략하기로 한다.
- [0040] 자연어 처리(Natural Language Processing, NLP)는 인간이 일상적으로 사용하는 자연어를 기계적으로 처리하는 기술을 의미할 수 있다. 이러한 자연어 처리는 자연어 분석, 자연어 이해, 자연어 생성 등의 기술을 포함할 수 있다. 자연어 분석은 자연어의 형태소, 통사, 의미 및 화용을 분석하는 것을 의미할 수 있고, 자연어 이해는

컴퓨터가 자연어로 주어진 입력에 따라 동작하게 하는 것을 의미할 수 있다. 또한, 자연어 생성은 특정 내용을 사람이 이해할 수 있는 자연어로 변환하는 것을 의미할 수 있다. 이러한 자연어 처리는 인공 지능(AI)의 주요한 개발 분야 중 하나로 머신 러닝(Machine Learning, ML) 및/또는 딥 러닝(Deep learning) 기술 등을 이용하여 수행될 수 있다. 예를 들어, 챗봇은 대화 사용자의 입력되는 대화를 이해하고 이에 적절한 답변을 제공하고 있으며, 음성 인식 플랫폼은 사용자의 음성이 입력되면 입력된 음성을 텍스트로 변환하여 사용자에게 제공할 수 있다.

[0041] 한편, 비디오 생성 모델은 생성형 AI로서, 전술한 자연어 처리 기술을 이용하여 입력된 자연어에 기초한 비디오를 생성할 수 있다. 그러나, 이전의 비디오 생성 모델은 한 문장 정도로 이루어진 텍스트(예: 자연어)에 기초하여 길이가 짧은 비디오 정도만을 생성할 수 있을 뿐, 영화와 같이 연속된 스토리에 기초하여 길이가 긴 비디오를 생성하지 못하는 문제가 있다.

[0042] 전술한 문제를 해결하기 위하여 일 실시예에 따른 비디오 생성 모델을 이용하여 비디오를 생성하는 장치는 스토리에 기초하여 청크(chunk) 단위로 비디오를 생성함에 따라 GPU에 영향을 받지 않고 무제한의 길이를 가진 비디오를 계속 이어가며 생성할 수 있다.

[0043] 도 1은 일 실시예에 따른 비디오 생성 모델을 이용하여 비디오를 생성하는 장치의 블록도를 도시한다.

[0044] 도 2는 일 실시예에 따른 비디오 생성 모델의 아키텍처를 예시적으로 나타낸다.

[0045] 장치(100)는 프로세서(110) 및 메모리(120)를 포함할 수 있다.

[0046] 프로세서(110)는 이미지 생성 모델(210)에 기본 이미지(10)(예: 입력 이미지)와 하나 이상의 스토리 텍스트(20)를 입력(input)함으로써, 입력 이미지들을 생성할 수 있고, 2차원 비디오 생성 모델(221)에 생성된 입력 이미지들을 입력함으로써, 각 청크(chunk)에 전체적인 비디오 정보를 보존하는 인터(inter) 항목, 이전 청크에서 생성한 비디오 정보를 보존하는 인트라(intra) 항목 및 해당 청크에서 생성하는 연속성 비디오의 연속된 프레임 중 직전 프레임 정보를 나타내는 콘스(consecutive, cons) 항목을 포함하는 청크 단위의 초기 연속성 비디오를 생성하며, 3차원 비디오 생성 모델(222)에 생성된 초기 연속성 비디오를 입력함으로써, 추가적인 모션이 더해져 정제된 수정 연속성 비디오를 생성할 수 있다. 비디오 생성 모델(220)은 2차원 비디오 생성 모델(221) 및 3차원 비디오 생성 모델(222)을 포함할 수 있다. 비디오 정보는 비디오의 테마, 장소, 배경 및 등장인물 중 하나 이상을 포함할 수 있다. 또한, 청크 단위의 초기 연속성 비디오는 인터 항목, 인트라 항목, 및 콘스 항목의 퓨전 비율(fusion ratio)이 청크의 생성된 순서에 따라 점진적으로 변경되는 것에 기초하여 생성될 수 있다. 퓨전 비율은 도 5에서 상세히 후술한다.

[0047] 이를 통해 프로세서(110)는 이어지는 배경 및/또는 등장인물을 길이 제한 없이 만들 수 있다.

[0048] 이미지 생성 모델(210)은 2차원 비디오 생성 모델(221)에 기초할 수 있다. 이미지 생성 모델(210)은 컨트롤넷(ControlNet)-AR(auto-regressive)을 포함하고, 컨트롤넷-AR은 인코더만으로 구성될 수 있다. 컨트롤넷-AR에 적용되는 제로 컨볼루션 레이어들은 인코더를 제외한 디코더 부분에 작용하기 때문에 디코더의 개수와 동일할 수 있다. 또한, 2차원 비디오 생성 모델(221)은 컨트롤넷-AR을 포함하고 컨트롤넷-AR이 복사한 원 모델을 포함할 수 있다. 다시 말하자면, 2차원 비디오 생성 모델(221)은 컨트롤넷-AR의 원 모델에 컨트롤넷-AR이 결합된 구조를 포함할 수 있다. 이에 트레이닝 시 프로세서(110)는 2차원 비디오 손실을 최소화함에 따라 결정된 2차원 비디오 파라미터를 동결(freezing)(예: 저장)하여 더 이상의 2차원 비디오 파라미터를 업데이트하지 않을 수 있다. 이후, 프로세서(110)는 2차원 비디오 생성 모델(221)을 복사한 트레이닝 가능한 이미지 생성 모델(210)을 이용하여 2차원 비디오 파라미터에 기초한 이미지 생성 파라미터를 결정할 수 있다. 이에 프로세서(110)는 이미지 생성 모델(210)을 통해 생성한 이미지들로부터 비디오를 자기 회귀(auto-regressive)하게 만들 수 있고 이는 더 나은 초기화를 하겠다는 것임을 의미할 수 있다. 이미지 생성 모델(210)은 제1 제로 컨볼루션 레이어(Zero Convolution layer), 트레이닝 가능한 이미지 생성 모델(210) 및 제2 제로 컨볼루션 레이어로 구성되고, 제로 컨볼루션 레이어는 트레이닝할수록 더욱 고품질의 결과를 생성할 수 있도록 한다. 이에 이미지 생성 모델(210)은 제로 컨볼루션 레이어로 2차원 비디오 생성 모델(221)과 연결됨에 따라 프로세서(110)가 이미지 생성 모델(210)을 트레이닝시키면 연결된 제로 컨볼루션 레이어 또한 트레이닝될 수 있다.

[0049] 프로세서(110)는 자동 회귀(Auto regressive) 방식으로 이미지 생성 모델(210)을 트레이닝할 수 있다. 이미지 생성 모델(210)은 수학적 1을 이용하여 산출된 이미지 생성 손실을 최소화하면서 이미지 생성 파라미터를 결정할 수 있다.

수학식 1

$$\mathcal{L} = \mathbb{E}_{z_0, t, c_t, c_f, \epsilon \sim \mathcal{N}(0,1)} [\|\epsilon - \epsilon_\theta(z_t, t, c_t, c_f)\|_2^2]$$

[0050]

[0051]

수학식 1에서, Z_0 는 입력 이미지를 나타내고, Z_t 는 노이즈가 추가된 이미지를 나타낼 수 있다. 또한, t 는 시간 스텝(time step)을 나타내고, c_t 는 조건으로서, 이미지 프레임들(예: 입력 이미지들)을 나타내고, c_f 는 텍스트 프롬프트(text prompt)를 나타낼 수 있다. 텍스트 프롬프트는 하나의 스토리를 구성하는 복수개의 스토리 텍스트 중 어느 하나를 의미할 수 있다.

[0052]

도 3은 일 실시예에 따른 비디오 생성 모델을 이용하여 비디오를 생성하는 장치에서 생성된 입력 이미지들을 예시적으로 나타낸다.

[0053]

프로세서(110)는 입력 이미지들(30)(예: (312,322))을 생성함에 있어서, 이와 같이 트레이닝된 이미지 생성 모델(210)을 통해 기본($n=0$) 이미지(10)와 제1 스토리 텍스트(20)(예: (311,321))에 현재($n=1$) 이미지 프레임을 생성하고, 현재($n=1$) 이미지 프레임과 제1 스토리 텍스트(311,321)에 기초하여 다음($n=2$) 이미지 프레임을 생성하며, 다음($n=2$) 이미지 프레임을 현재($n=1$) 이미지 프레임으로 업데이트하여 각 스토리 텍스트 별로 미리 결정된 개수의 이미지 프레임들을 생성하여 입력 이미지들(30)을 구성할 수 있다. 이는 이전($n-1$) 이미지 프레임과 현재 주기(m) 스토리 텍스트에 기초하여 현재(n) 이미지 프레임을 생성할 수 있음을 의미할 수 있다. 이에 생성된 이미지 프레임들은 연속적인 스토리의 각 시퀀스와 관련하여 일관성(consistency)을 가질 수 있다. 결국, 프로세서(110)는 컨트롤넷(ControlNet)-AR을 활용한 이미지 생성 모델(210)을 통해 이미지 준비 과정으로 입력 이미지들(30)을 생성할 수 있다. 프로세서(110)는 생성된 입력 이미지들(30)을 비디오 생성 모델(220)에 입력하며, 이는 입력 이미지들(30)이 2차원 비디오 생성 모델(221)에 입력됨에 따라 더 나은 초기화가 수행될 수 있음을 의미할 수 있다.

[0054]

프로세서(110)는 청크 단위의 초기 연속성 비디오를 생성함에 있어서, 2차원 비디오 생성 모델(221)을 이용하여 생성된 입력 이미지들로부터 청크 순서에 따른 인터 항목 및 인트라 항목에 대한 이미지 정보를 추출하고, 인터 항목, 인트라 항목 및 콘스 항목에 대한 어텐션 가중치를 계산하고, 미리 설정된 인터 항목, 인트라 항목 및 콘스 항목에 대한 퓨전 비율을 조정하고, 어텐션 가중치와 퓨전 비율에 기초하여 인터 항목, 인트라 항목 및 콘스 항목에 대한 히든 값을 산출하고, 인터 항목, 인트라 항목 및 콘스 항목에 대한 히든 값에 기초하여 히든 상태를 결정하고, 결정된 히든 상태에 기초하여 새로운 이미지를 생성하며, 새로운 이미지를 인트라 이미지로 교체할 수 있다. 프로세서(110)는 2차원 비디오 생성 모델(221)을 이용하여 비디오를 생성하는 경우, 예를 들어, 첫번째 청크에 사용되는 이미지는 청크의 크기가 10이라고 한다면, 첫 두 개의 프레임을 인터 항목과 인트라 항목으로 사용될 수 있다.

[0055]

도 4는 일 실시예에 따른 비디오 생성 모델의 아키텍처 중 2차원 비디오 생성 모델의 구조를 예시적으로 나타낸다.

[0056]

프로세서(110)는 2차원 비디오 생성 모델(221)의 컨볼루션 레이어(Convolution Layer)를 이용하여 입력 이미지들(30)로부터 이미지 정보(예: 특징(feature))를 추출할 수 있다. 그런 후, 프로세서(110)는 어텐션 레이어(attention layer)를 통해 인터 항목, 인트라 항목 및 콘스 항목에 대한 어텐션 가중치를 계산할 수 있다. 프로세서(110)는 어텐션 가중치를 계산함에 있어서, 인터 항목의 인터 이미지 정보에 기초하여 쿼리(query, Q), 키(key, K), 값(value, V)에 기반한 인터 어텐션 가중치를 계산하고, 인트라 항목의 인트라 이미지 정보에 기초하여 쿼리, 키, 값에 기반한 인트라 어텐션 가중치를 계산하며, 콘스 항목의 직전 프레임 정보에 기초하여 쿼리, 키, 값에 기반한 콘스 어텐션 가중치를 계산할 수 있다. 쿼리는 해당 프레임의 스토리를 구성하는 텍스트가 인코딩되어 임베딩으로 들어가는 것을 나타낸다. 키와 값은 참고하고 싶은 이미지를 인코딩하여 사용하는 것을 나타내고, 예를 들어, 인터 항목, 인트라 항목 및 콘스 항목을 나타낼 수 있다. 선형 프로젝션(linear projection)의 경우, 패치단위로 입력된 항목들을 플랫텐(flatten)(예: 1차원의 텐서(Tensor) 행렬로 변환)한 후 사용할 수 있다. 또한, 이에 기초하여 어텐션(Attention)을 계산할 수 있다. 추가로, 소프트 맥스(Softmax)는 쿼리, 키, 값으로부터 확률 분포를 변환하는 과정 중 크로스 어텐션 계산에서 사용될 수 있다. 또한, 하나의 청크에는 제1 스토리 텍스트에 기초한 제1 이미지들(410)과 제2 스토리 텍스트에 기초한 제2 이미지들(420)을 포함할 수 있다. 예를 들어, 프로세서(110)는 도 2에서의 스토리 텍스트(20) 중 첫번째 문장에 기초

하여 제1 이미지들(410)을 생성할 수 있고, 스토리 텍스트(20) 중 두번째 문장에 기초하여 제2 이미지들(420)을 생성할 수 있다. 일 실시예에 따른 비디오 생성 모델을 이용하여 비디오를 생성하는 장치는 생성된 제1 이미지들(410)과 제2 이미지들(420)이 각기 다른 스토리 텍스트에 기초하여 생성됨에 따라 발생할 수 있는 불연속성을 인터 항목 및 인트라 항목을 이용하여 해소하는 것뿐만 아니라 추가적인 방안을 가질 수 있는데, 이와 관련된 상세한 설명은 도 6에서 후술한다.

[0057] 프로세서(110)는 히든 상태 퓨전(Hidden State Fusion) 레이어를 통해 인터 항목, 인트라 항목 및 콘스 항목에 대한 퓨전 비율과 인터 항목, 인트라 항목 및 콘스 항목에 대한 히든 값에 기초하여 히든 상태를 결정할 수 있다. 프로세서(110)는 크로스-어텐션 레이어(Cross-Attention) 레이어를 통해 서로 다른 모달리티(modality)(예: 텍스트와 이미지) 간의 상호작용을 고려하여 결정된 히든 상태에 기초하여 새로운 이미지를 생성할 수 있다. 프로세서(110)는 피드 포워드 네트워크(Feed Forward Network) 레이어를 통해 새로운 이미지를 인트라 이미지로 교체할 수 있다.

[0058] 전술한 히든 상태 퓨전 레이어는 스케줄러(scheduler)를 이용하여 보간(interpolate)하는 정도가 조정될 수 있다.

[0059] 도 5(a)는 일 실시예에 따른 비디오 생성 모델을 이용하여 비디오를 생성하는 장치에서 생성된 청크의 시계열적 구조를 예시적으로 나타내고, 도 5(b)는 일 실시예에 따른 비디오 생성 모델을 이용하여 비디오를 생성하는 장치에서 생성된 청크의 시계열적 퓨전 비율의 그래프를 예시적으로 나타낸다.

[0060] 각 청크는 인터 항목, 인트라 항목 및 콘스 항목을 포함하는데, 프로세서(110)는 인터 항목, 인트라 항목 및 콘스 항목(예: 입력 이미지(30))에 기초하여 새로운 콘스 항목(예: 복수개의 이미지 프레임)을 생성할 수 있다.

[0061] 인터 항목은 연속성의 비디오의 첫번째 프레임에 대한 정보를 포함할 수 있다. 이에 인터 항목은 초기 연속성 비디오 내 모든 청크에 동일한 정보를 포함할 수 있다. 또한, 인트라 항목은 이전 청크의 마지막 프레임에 대한 정보를 포함할 수 있다. 이에 인트라 항목은 이전 청크의 콘스 항목을 다음 청크의 인트라 항목의 정보로 하여 포함할 수 있다. 또한, 콘스 항목은, 예를 들어, 8개의 프레임 이미지를 포함할 수 있으나, 이에 한정되지 않는다.

[0062] 또한, 프로세서(110)는 퓨전 비율을 조정함에 있어서, 시간에 따라 콘스 항목에 대한 퓨전 비율을 조정하며, 콘스 항목에 대한 퓨전 비율에 기초하여 인트라 항목과 인터 항목에 대한 퓨전 비율을 조정할 수 있다. 예를 들어, 콘스 항목에 대한 퓨전 비율은 시간에 한정되지 않고 청크의 개수와 길이에도 기초하여 조정될 수 있다. 이에 프로세서(110)는 스케줄러를 통해 결정된 퓨전 비율에 기초하여 각 청크의 비디오를 생성할 수 있다. 예를 들어, 제0(zero) 청크에서의 청크 개수는 영(zero)이기 때문에 인터 항목의 비율이 1이 되고, 이에 따라 프로세서(110)는 인터 항목만을 참조하여 제0(zero) 청크를 생성할 수 있다. 또한, 도 5(b)에서 볼 수 있는 바와 같이, 제1 청크에서 제5 청크가 생성될수록 점진적으로 인터 항목의 퓨전 비율이 증가한다. 만약 시간이 지나도(예: 여러 개의 청크가 생성되어도) 여전히 인터 항목의 퓨전 비율이 높은 경우, 프로세서(110)는 일그러진 비디오를 생성할 수 있다. 추가로, 콘스 항목(예: 복수개의 이미지 프레임)을 생성하는 경우에도 각각의 이미지 프레임의 퓨전 비율도 다르게 설정될 수 있다. 예를 들어, 하나의 청크에 시간 순으로 $t=1$ 보다 $t=5$ 에 있는 이미지 프레임의 인터 항목을 참조하는 비율이 더 낮을 수 있다.

[0063] 도 6은 일 실시예에 따른 비디오 생성 모델을 이용하여 비디오를 생성하는 장치에서 생성된 청크에 스토리 텍스트가 대응된 구조를 예시적으로 나타낸다.

[0064] 프로세서(110)는 (610)와 같이 스토리 텍스트 하나에만 대응한 이미지에 기초하여 청크를 생성할 수 있다. 그러나 (610)과 같은 구조는 제1 청크와 제2 청크의 비디오를 합치는 경우, 배경과 등장인물이 변경되는 문제가 발생할 수 있다. 이러한 문제를 해결하기 위하여 스토리 텍스트는 하나의 청크를 생성하는 이미지에 대한 스토리 텍스트가 이어지는 다음 청크를 생성하는 이미지에 대한 스토리 텍스트와 일부 중복될 수 있다. 예를 들어, 프로세서(110)는 (620)과 같이 제1 스토리 텍스트와 제2 스토리 텍스트에 기초하여 제1 청크를 생성하고, 제2 스토리 텍스트, 제3 스토리 텍스트 및 제4 스토리 텍스트에 기초하여 제2 청크를 생성할 수 있다. 이를 통해 프로세서(110)는 배경과 등장인물이 일관된 연속적인 비디오를 생성할 수 있다.

[0065] 도 7은 일 실시예에 따른 비디오 생성 모델을 이용하여 비디오를 생성하는 장치에서 생성된 초기 연속성 비디오와 수정 연속성 비디오를 예시적으로 나타낸다.

[0066] 프로세서(110)는 수정 연속성 비디오(720)를 생성함에 있어서, 3차원 비디오 생성 모델(222)의 디퓨전 모델을 이용하여 생성된 초기 연속성 비디오(710)에 노이즈를 더하고, 확률적 미분 방정식을 이용하여 디노이징함에 따

라 노이즈가 더해진 초기 연속성 비디오를 정제할 수 있다. 도 7에서 도시된 바와 같이, 수정 연속성 비디오 (720)는 초기 연속성 비디오(710)의 표기된 위치의 대상물에 움직임의 더 부여하여 이미지가 개선될 수 있다.

[0067] 도 8은 일 실시예에 따른 비디오 생성 모델을 이용하여 비디오를 생성하는 방법의 흐름도를 도시한다.

[0068] 단계(810)에서, 프로세서는 이미지 생성 모델에 기본 이미지와 하나 이상의 스토리 텍스트를 입력(input)함으로써, 입력 이미지들을 생성할 수 있다. 프로세서는 이미지 생성 모델을 통해 기본 이미지와 제1 스토리 텍스트에 현재 이미지 프레임을 생성하고, 현재 이미지 프레임과 제1 스토리 텍스트에 기초하여 다음 이미지 프레임을 생성하며, 다음 이미지 프레임을 현재 이미지 프레임으로 업데이트하여 각 스토리 텍스트 별로 미리 결정된 개수의 이미지 프레임들을 생성하여 입력 이미지들을 구성할 수 있다.

[0069] 단계(820)에서, 프로세서는 2차원 비디오 생성 모델에 생성된 입력 이미지들을 입력함으로써, 각 청크(chunk)에 전체적인 비디오 정보를 보존하는 인터(inter) 항목, 이전 청크에서 생성한 비디오 정보를 보존하는 인트라(intra) 항목 및 해당 청크에서 생성하는 연속성 비디오의 연속된 프레임 중 직전 프레임을 나타내는 콘스(cons) 항목을 포함하는 청크 단위의 초기 연속성 비디오를 생성할 수 있다. 비디오 정보는 비디오의 테마, 장소, 배경 및 등장인물 중 하나 이상을 포함할 수 있다. 또한, 인터 항목은 연속성의 비디오의 첫번째 프레임에 대한 정보를 포함하고, 인트라 항목은 이전 청크의 마지막 프레임에 대한 정보를 포함할 수 있다.

[0070] 단계(830)에서, 프로세서는 3차원 비디오 생성 모델에 생성된 초기 연속성 비디오를 입력함으로써, 추가적인 모션이 더해져 정제된 수정 연속성 비디오를 생성할 수 있다. 프로세서는 3차원 비디오 생성 모델의 디퓨전(diffusion) 모델을 이용하여 생성된 초기 연속성 비디오에 노이즈를 더하고, 확률적 미분 방정식을 이용하여 디노이징(denoising)(예: 노이즈 제거)함에 따라 노이즈가 더해진 초기 연속성 비디오를 정제할 수 있다.

[0071] 청크 단위의 초기 연속성 비디오는, 인터 항목, 인트라 항목, 및 콘스 항목의 퓨전 비율(fusion ratio)이 청크의 생성된 순서에 따라 점진적으로 변경되는 것에 기초하여 생성될 수 있다. 또한, 스토리 텍스트는 하나의 청크에 대한 스토리 텍스트가 이어지는 다음 청크에 대한 스토리 텍스트와 일부 중복될 수 있다.

[0072] 청크 단위의 초기 연속성 비디오를 생성함에 있어서, 프로세서는 표 1의 알고리즘을 이용할 수 있다.

표 1

Algorithm 1 Algorithm for Cross Attention Fusion and Decay Scheduler

```

1: procedure GENERATE VIDEO(s:story)
2:   n = 0
3:   Inter_img = C0[0]
4:   Intra_img = Cn-1[0]
5:   Global_step = len(Cn) * Nc
6:   while n = Nc do
7:     Cn = Inter_img + Intra_img + Cn
8:     Inter=attn(Q(s[n]), K(Cn[0]),V(Cn[0]))           ▷ Inter Attn
9:     Intra=attn(Q(s[n]), K(Cn[1]),V(Cn[1]))           ▷ Intra Attn
10:    Hidden_states = []
11:    for i = 0, ..., len(Cn) - 1 do
12:      Cons= attn(Q(s[n]), K(Cn[i]),V(Cn[i]))           ▷ Cons Attn
13:      α = (n / Global_step) / 7                         ▷ Cons Fusion ratio
14:      β = α * 6                                           ▷ Intra Fusion ratio
15:      γ = 1 - α                                           ▷ Inter Fusion ratio
16:      Cons_hidden = Cons * (α*0.75 + 0.05)
17:      Intra_hidden = Intra * (β*0.75 + 0.05)
18:      Inter_hidden = Inter * (γ*0.75 + 0.15)
19:      Hidden_state = Cons_hidden+Intra_hidden+Inter_hidden
20:      Hidden_states += Hidden_state
21:    n += 1
22:    new = θ(Hidden_states)
23:    result = new[2:]
24:    Intra_img = new[0]                                   ▷ replace Intra_img
25:    results += result
26:  return results

```

[0073]

- [0074] 프로세서는 생성된 입력 이미지들로부터 청크 순서에 따른 인터 항목 및 인트라 항목에 대한 이미지 정보를 추출할 수 있다. 2번째 줄 내지 5번째 줄에서, 청크는 0번째($n=0$)부터 생성하고, 인터 항목(예: 인터 이미지(*inter_img*))은 제0 청크의 0프레임 이미지로 설정될 수 있다. 또한, 인트라 항목(예: 인트라 이미지(*intra_img*))은 이전($n-1$) 청크의 0프레임 이미지로 설정되고, 글로벌 스텝(*Global_step*)은 시간에 기초하여 산출될 수 있다.
- [0075] 7번째 줄에서, n 은 청크 개수에 대응되고, n 번째 청크는 인터 이미지와 인트라 이미지에 기초하여 결정될 수 있다.
- [0076] 프로세서는 인터 항목, 인트라 항목 및 콘스 항목에 대한 어텐션(*attention*) 가중치를 계산할 수 있다. 이에 8번째 줄에서, 프로세서는 인터 항목의 인터 이미지 정보에 기초하여 쿼리(*query*), 키(*key*), 값(*value*)에 기반한 인터 어텐션(*attention*) 가중치를 계산할 수 있다.
- [0077] 9번째 줄에서, 프로세서는 인트라 항목의 인트라 이미지 정보에 기초하여 쿼리, 키, 값에 기반한 인트라 어텐션 가중치를 계산할 수 있다.
- [0078] 12번째 줄에서, 프로세서는 콘스 항목의 직전 프레임 정보에 기초하여 쿼리, 키, 값에 기반한 콘스 어텐션 가중치를 계산할 수 있다.
- [0079] 프로세서는 미리 설정된 인터 항목, 인트라 항목 및 콘스 항목에 대한 퓨전(*fusion*) 비율을 조정할 수 있다. 이에 13번째 줄 내지 15번째 줄에서, 프로세서는 퓨전 비율에 기초하여 인트라 항목, 인터 항목 및 콘스 항목에 대한 퓨전 비율을 조정할 수 있다. 이는 프로세서가 처음 사용자가 제공한 퓨전 비율에 따라 시간이 지날수록 인터 항목의 참고 비율을 줄이고 인트라 항목 및 콘스 항목의 참고 비율을 높이는 방향으로 진행됨을 의미할 수 있다.
- [0080] 16번째 줄 내지 18번째 줄에서, 프로세서는 어텐션 가중치와 퓨전 비율에 기초하여 인터 항목, 인트라 항목 및 콘스 항목에 대한 히든(*hidden*) 값을 산출할 수 있다.
- [0081] 19번째 줄에서, 프로세서는 인터 항목, 인트라 항목 및 콘스 항목에 대한 히든 값에 기초하여 히든 상태를 결정할 수 있다. 프로세서는 인터 항목, 인트라 항목 및 콘스 항목에 대한 히든 상태들을 합하여 전체의 히든 상태를 결정할 수 있다. 다시 말하자면, 전체의 히든 상태는 각각의 비율로 비례하게 조절된 인터 항목, 인트라 항목 및 콘스 항목의 값들의 합에 기초할 수 있다.
- [0082] 23번째 줄에서, 프로세서는 결정된 히든 상태에 기초하여 새로운 이미지를 생성할 수 있다.
- [0083] 24번째 줄에서, 프로세서는 새로운 이미지를 인트라 이미지로 교체할 수 있다.
- [0084] 도 9는 일 실시예에 따른 비디오 생성 모델을 이용하여 비디오를 생성하는 방법에서 생성된 수정 연속성 비디오와 이전의 비디오 생성 모델을 이용하여 생성된 비디오를 예시적으로 나타낸다.
- [0085] 비디오(911, 921)는 스토리 텍스트들(913, 923)에 기초하여 이전의 비디오 생성 모델을 이용하여 생성된 비디오를 나타낼 수 있다. 또한, 수정 연속성 비디오(912, 922)는 문장(913, 923)에 기초하여 일 실시예에 따른 비디오 생성 모델(예: 2차원 비디오 생성 모델 및 3차원 비디오 생성 모델)을 이용하여 생성된 비디오를 나타낼 수 있다. 수정 연속성 비디오(912, 922)는 비디오(911, 921)보다 스토리를 더 정확하게 반영하고 일관성 있으며 정제된 비디오일 수 있다.
- [0086] 이상에서 설명된 실시예들은 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치, 방법 및 구성요소는, 예를 들어, 프로세서, 컨트롤러, ALU(*arithmetic logic unit*), 디지털 신호 프로세서(*digital signal processor*), 마이크로컴퓨터, FPGA(*field programmable gate array*), PLU(*programmable logic unit*), 마이크로프로세서, 또는 명령(*instruction*)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(*processing element*) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하

나의 프로세서 및 하나의 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0087] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 저장될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0088] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있으며 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

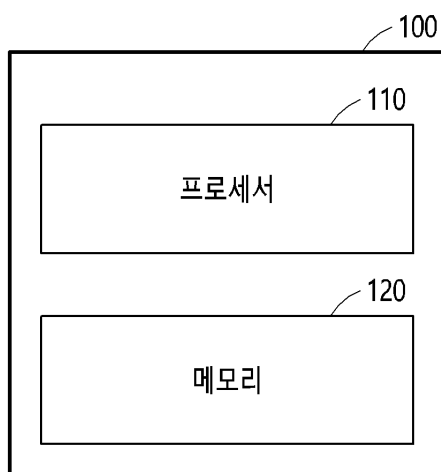
[0089] 위에서 설명한 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 또는 복수의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0090] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 이를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

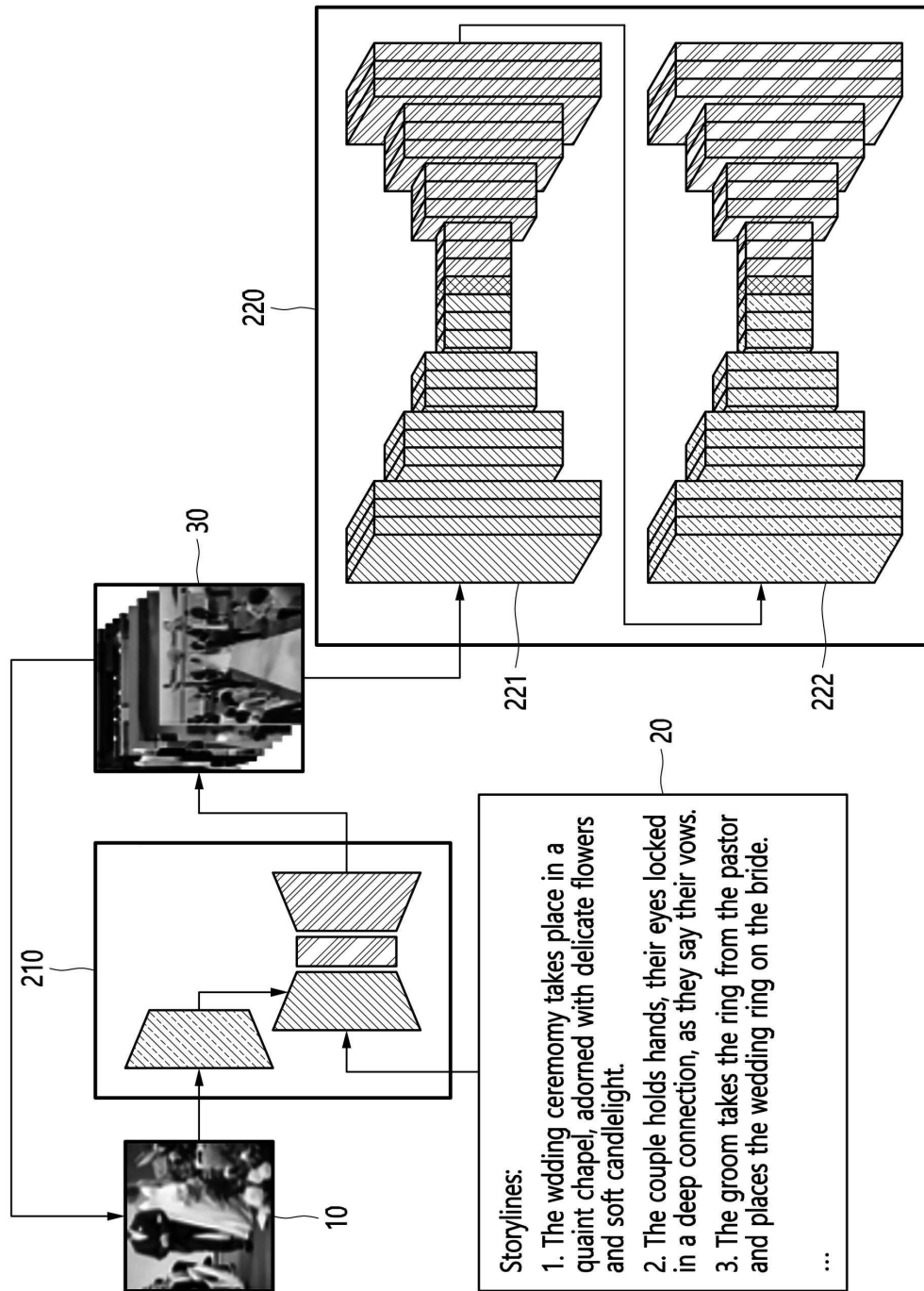
[0091] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

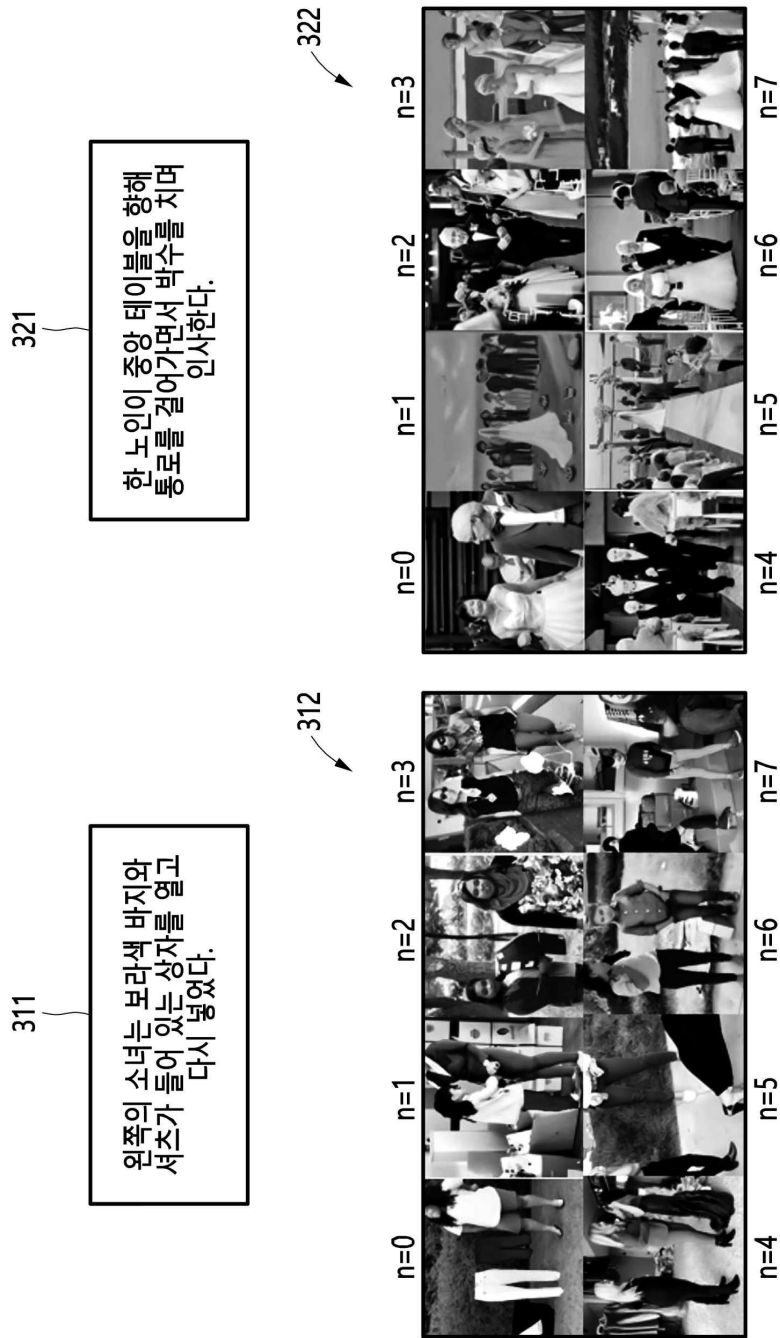
도면1



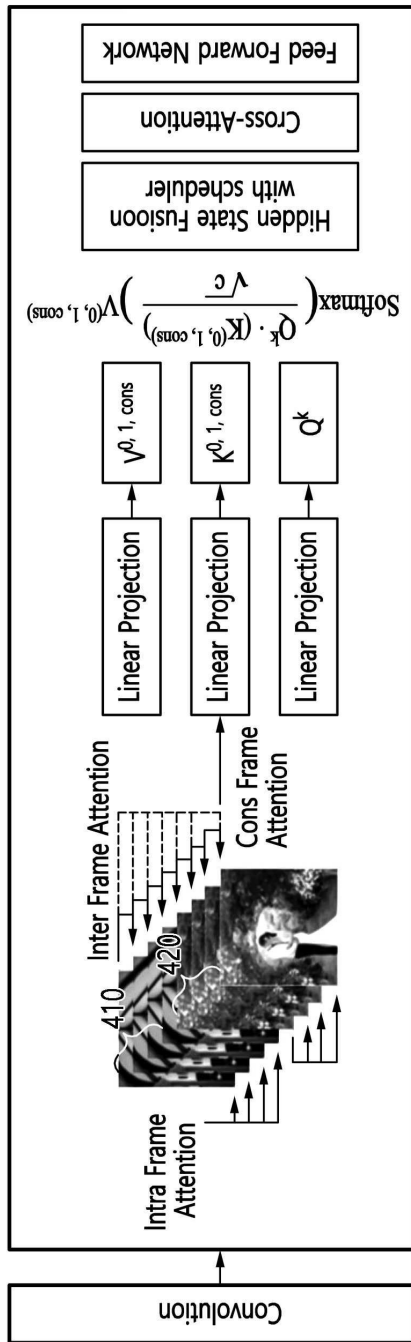
도면2



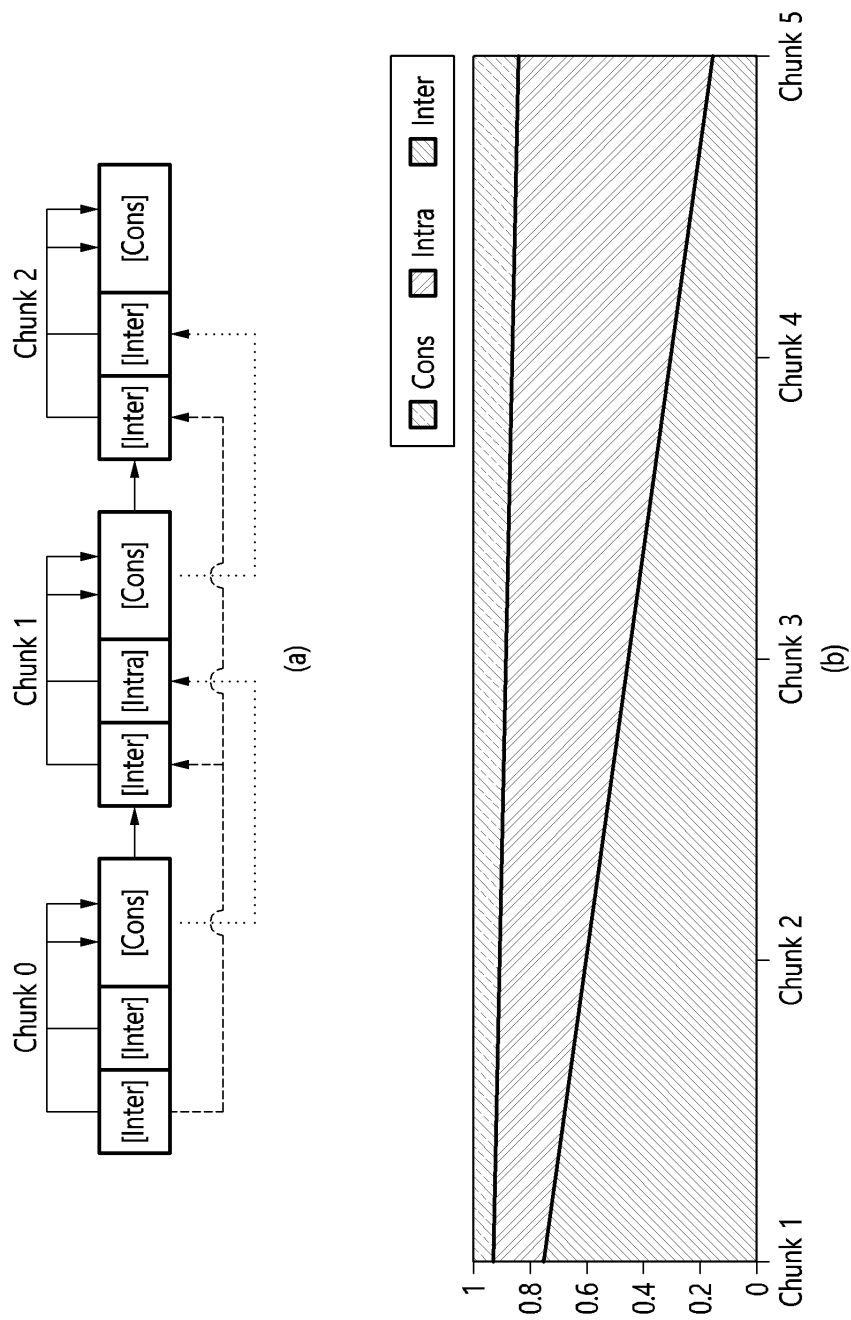
도면3



도면4

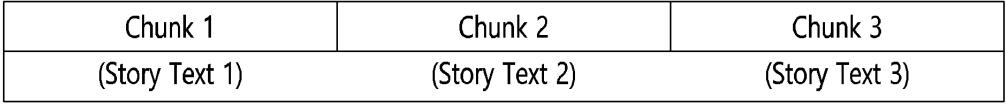


도면5



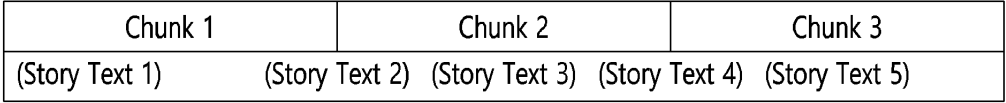
도면6

610



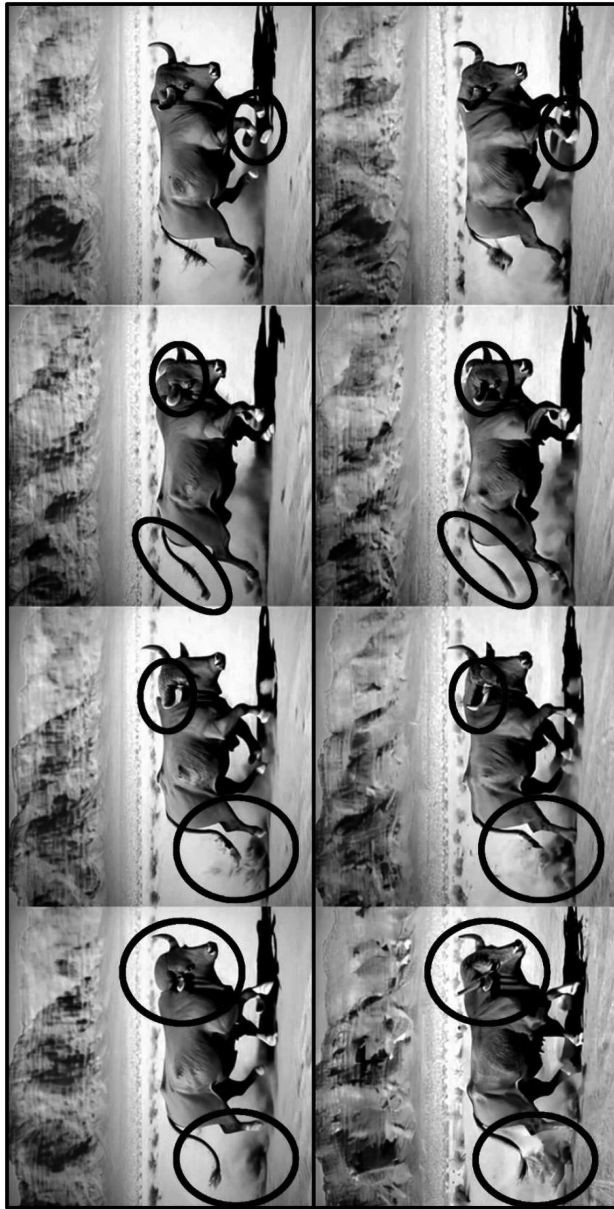
Chunk 1	Chunk 2	Chunk 3
(Story Text 1)	(Story Text 2)	(Story Text 3)

620

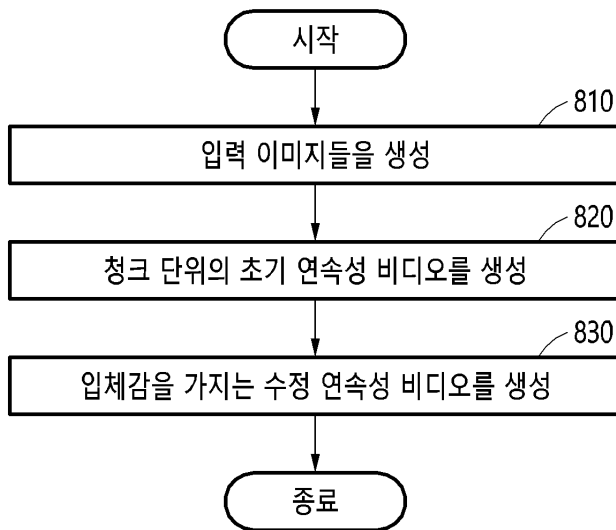


Chunk 1	Chunk 2	Chunk 3		
(Story Text 1)	(Story Text 2)	(Story Text 3)	(Story Text 4)	(Story Text 5)

도면7



도면8



도면9

