(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0153887
(43) 공개일자 2024년10월24일

(51) 국제특허분류(Int. Cl.)
G10L 21/10 (2013.01) *G06F 16/64* (2019.01)
G06T 11/00 (2006.01) *G06T 13/20* (2011.01)
G10L 15/02 (2006.01) *G10L 15/06* (2006.01)
G10L 21/0316 (2013.01)
(52) CPC특허분류
G10L 21/10 (2013.01)
G06F 16/64 (2019.01)
(21) 출원번호 10-2023-0100634
(22) 출원일자 2023년08월01일
심사청구일자 2023년08월01일
(30) 우선권주장
1020230050310 2023년04월17일 대한민국(KR)

(71) 출원인
포항공과대학교 산학협력단
경상북도 포항시 남구 청암로 77 (지곡동)
(72) 발명자
오태현
경상북도 포항시 남구 청암로 77 포항공과대학교 전자전기공학과공학2동 415호
하현우
경상남도 양산시 물금읍 신주로 35, 114동 1905호 (양산2차e-편한세상)
김성빈
세종특별자치시 달빛로 80, 1220-2203 (종촌동, 가재마을12단지)
(74) 대리인
제일특허법인(유)

전체 청구항 수 : 총 23 항

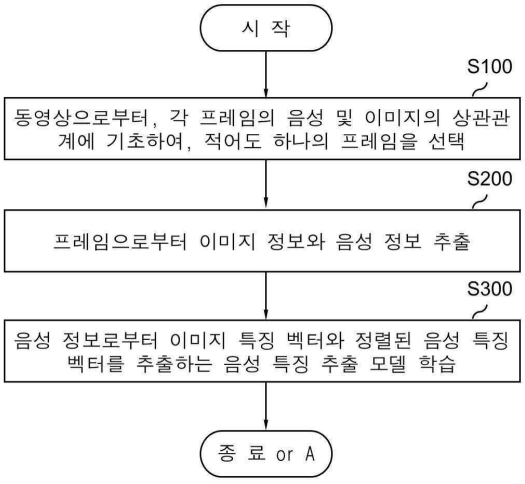
(54) 발명의 명칭 **이미지를 생성하는 장치 및 딥러닝 학습 방법**

(57) 요 약

일 실시예에 따른 이미지 생성 모델 학습 방법은, 복수의 프레임으로 구성된 동영상으로부터, 각 프레임의 음성 및 이미지의 상관관계에 기초하여, 적어도 하나의 프레임을 선택하는 단계, 상기 선택된 적어도 하나의 각 프레임으로부터 이미지 정보와 음성 정보를 추출하는 단계 및 상기 음성 정보로부터 음성 특징 벡터를 추출하는 음성 특징 벡터 추출 모델을 학습시키는 단계를 포함한다.

여기서, 상기 음성 특징 벡터는, 기 학습된 이미지 특징 벡터 추출 모델에 의해 상기 이미지 정보로부터 추출된 이미지 특징 벡터와 소정 임베딩 공간 내에서 정렬된 것일 수 있다.

대 표 도 - 도3



(52) CPC특허분류
G06T 11/00 (2013.01)
G06T 13/205 (2013.01)
G10L 15/02 (2013.01)
G10L 15/063 (2013.01)
G10L 21/0316 (2021.08)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711193622
과제번호 2021-0-02068-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 인공지능 혁신 허브 연구 개발
기 여 율 50/100
과제수행기관명 고려대학교산학협력단
연구기간 2023.01.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711195728
과제번호 00225630
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 실감콘텐츠핵심기술개발
연구과제명 문장으로부터의 3차원 동영상 자동 생성 기술
기 여 율 50/100
과제수행기관명 한국과학기술연구원
연구기간 2023.04.01 ~ 2023.12.31

명 세 서

청구범위

청구항 1

음성으로부터 이미지를 생성하는 이미지 생성 모델 학습 방법에 있어서,

복수의 프레임으로 구성된 동영상으로부터, 각 프레임의 음성 및 이미지의 상관관계에 기초하여, 적어도 하나의 프레임을 선택하는 단계;

상기 선택된 적어도 하나의 각 프레임으로부터 이미지 정보와 음성 정보를 추출하는 단계; 및

상기 음성 정보로부터 음성 특징 벡터를 추출하는 음성 특징 벡터 추출 모델을 학습시키는 단계를 포함하고,

상기 음성 특징 벡터는, 기 학습된 이미지 특징 벡터 추출 모델에 의해 상기 이미지 정보로부터 추출된 이미지 특징 벡터와 소정 임베딩 공간 내에서 정렬된 것인,

이미지 생성 모델 학습 방법.

청구항 2

제1항에 있어서,

상기 적어도 하나의 프레임을 선택하는 단계에서,

프레임 셀렉션 방법(frame selection method)에 의해 상기 동영상으로부터 프레임이 선택되는,

이미지 생성 모델 학습 방법.

청구항 3

제1항에 있어서,

상기 음성 특징 벡터를, 상기 이미지 특징 벡터를 기초로 이미지를 생성하도록 기 학습된 이미지 생성기에 입력하여 이미지를 생성하는 단계를 더 포함하는,

이미지 생성 모델 학습 방법.

청구항 4

제1항에 있어서,

상기 음성 특징 벡터 추출 모델을 학습시키는 단계는,

대조 학습(contrastive learning)의 방법에 의해 수행되는,

이미지 생성 모델 학습 방법.

청구항 5

제4항에 있어서,

상기 대조 학습의 방법은, InfoNCE(noise contrastive estimation)를 사용하는 것을 포함하는,

이미지 생성 모델 학습 방법.

청구항 6

이미지 생성 장치에 있어서,

제1 음성을 입력 받는 입력부;

상기 제1 음성으로부터 제1 음성 특징 벡터를 추출하는 음성 특징 벡터 추출부; 및

상기 제1 음성 특징 벡터를 기초로 제1 이미지를 생성하는 이미지 생성기를 포함하고,

상기 음성 특징 벡터 추출부는,

복수의 프레임으로 구성된 동영상으로부터, 각 프레임의 음성 및 이미지의 상관관계에 기초하여, 적어도 하나의 프레임이 선택되고, 상기 선택된 적어도 하나의 각 프레임으로부터 제2 이미지와 제2 음성이 추출되면, 상기 제2 음성으로부터 제2 음성 특징 벡터를 추출하도록 학습된 것이고,

상기 제2 음성 특징 벡터는,

기 학습된 이미지 특징 벡터 추출 모델에 의해 상기 제2 이미지로부터 추출된 제2 이미지 특징 벡터와 소정 임베딩 공간 내에서 정렬된 것이고,

상기 이미지 생성기는,

상기 제2 이미지 특징 벡터를 기초로 제2 이미지를 생성하도록 기 학습된 것인,

이미지 생성 장치.

청구항 7

제6항에 있어서,

상기 제1 음성은 상기 제2 음성과 다른 것인,

이미지 생성 장치.

청구항 8

제6항에 있어서,

상기 제1 음성의 볼륨 크기가 달라지는 경우, 상기 달라진 볼륨 크기를 반영하여 상기 제1 이미지가 생성되는,

이미지 생성 장치.

청구항 9

제6항에 있어서,

상기 입력부는,

상기 제1 음성 또는 제3 이미지를 입력 받을 수 있고,

상기 이미지 생성기는,

상기 제1 음성과 제3 이미지가 함께 입력되면, 상기 제3 이미지에 상기 제1 이미지가 반영된 제4 이미지를 생성하는,

이미지 생성 장치.

청구항 10

제9항에 있어서,

상기 제4 이미지는,

상기 제3 이미지에 상기 제1 음성에 대응하는 새로운 객체가 더해져서 생성된 것인,

이미지 생성 장치.

청구항 11

제9항에 있어서,

상기 제4 이미지는,

상기 제1 음성에 대응하여 상기 제3 이미지가 변경된 것인,

이미지 생성 장치.

청구항 12

제9항에 있어서,

상기 제1 음성의 볼륨 크기가 달라지는 경우, 상기 달라진 볼륨 크기를 반영하여 상기 제4 이미지가 생성되는,

이미지 생성 장치.

청구항 13

제6항에 있어서,

상기 제1 음성은, 복수의 개체로부터 발생된 복수의 음원을 포함하고,

상기 제1 이미지는, 상기 복수의 개체를 구성하는 각 개체에 대응하는 이미지가 포함된 것인,

이미지 생성 장치.

청구항 14

제13항에 있어서,

상기 제1 음성에 포함된 복수의 음원에 대응하는 각각의 볼륨 크기가 상대적으로 달라지는 경우, 상기 상대적으로 달라진 각각의 볼륨 크기를 반영하여 상기 제1 이미지가 생성되는,

이미지 생성 장치.

청구항 15

제6항에 있어서,

상기 제1 이미지를 기초로 동영상을 생성하는 동영상 생성기를 포함하고,

상기 입력부는,

상기 제1 음성 또는 제1 동영상을 입력 받을 수 있고,

상기 동영상 생성기는,

상기 제1 동영상을 구성하는 제1 복수의 이미지를 기초로, 상기 제1 음성에 대응하는 새로운 객체가 더해져서 생성된 제2 복수의 이미지로부터 제2 동영상을 생성하는,

이미지 생성 장치.

청구항 16

제6항에 있어서,

상기 제1 이미지를 기초로 동영상을 생성하는 동영상 생성기를 포함하고,

상기 입력부는,

상기 제1 음성 또는 제1 동영상을 입력 받을 수 있고,

상기 동영상 생성기는,

상기 제1 음성의 볼륨 크기가 달라지는 경우, 상기 제1 동영상을 구성하는 제1 복수의 이미지를 기초로, 상기 달라진 볼륨 크기를 반영하여 생성된 제2 복수의 이미지로부터 제2 동영상을 생성하는,

이미지 생성 장치.

청구항 17

이미지 생성 장치에 있어서,

제1 음성을 입력 받는 입력부;

컴퓨터 실행 가능한 명령어를 포함하는 메모리; 및

상기 명령어를 실행함으로써,

기 학습된 음성 특징 벡터 추출 모델을 이용하여 상기 제1 음성으로부터 제1 음성 특징 벡터를 추출하고, 이미지 생성기를 이용하여 상기 제1 음성 특징 벡터를 기초로 제1 이미지를 생성하도록 제어하는 프로세서를 포함하되,

상기 음성 특징 추출 모델은,

복수의 프레임으로 구성된 동영상으로부터, 각 프레임의 음성 및 이미지의 상관관계에 기초하여, 적어도 하나의 프레임이 선택되고, 상기 선택된 적어도 하나의 각 프레임으로부터 제2 이미지와 제2 음성이 추출되면, 상기 제2 음성으로부터 제2 음성 특징 벡터를 추출하도록 학습된 것이고,

상기 제2 음성 특징 벡터는,

기 학습된 이미지 특징 벡터 추출 모델에 의해 상기 제2 이미지로부터 추출된 제2 이미지 특징 벡터와 소정 임베딩 공간 내에서 정렬된 것이고,

상기 이미지 생성기는,

상기 제2 이미지 특징 벡터를 기초로 제2 이미지를 생성하도록 기 학습된 것인,

이미지 생성 장치.

청구항 18

섬네일 생성 장치에 의해 수행되는 섬네일 생성 방법에 있어서,

음성 파일을 입력하는 단계;

상기 음성 파일 내 기 정해진 시간 간격에 따라 적어도 하나의 음성 정보를 추출하는 단계;

상기 추출된 적어도 하나의 음성 정보를 기 학습된 음성 특징 벡터 추출 모델에 입력하여 적어도 하나의 음성 특징 벡터를 추출하는 단계; 및

상기 음성 특징 벡터를, 기 학습된 이미지 특징 벡터 추출 모델에 의해 상기 음성 정보에 대응하는 이미지 정보로부터 추출된 이미지 특징 벡터를 기초로 이미지를 생성하도록 기 학습된 이미지 생성기에 입력하여 적어도 하나의 섬네일을 생성하는 단계를 포함하고,

상기 음성 특징 벡터는 상기 이미지 특징 벡터와 소정 임베딩 공간 내에서 정렬된 것인,

섬네일 생성 방법.

청구항 19

제18항에 있어서,

상기 적어도 하나의 음성 특징 벡터에 대해, 군집화(clustering)를 통해 상기 음성 특징 벡터를 각 군집으로 분류하는 단계; 및

상기 각 군집 내 대표 음성 특징 벡터를 결정하는 단계를 더 포함하고,

상기 섬네일을 생성하는 단계는,

상기 대표 음성 특징 벡터를 상기 이미지 생성기에 입력하여 섬네일을 생성하는,

섬네일 생성 방법.

청구항 20

제19항에 있어서,

상기 섬네일이 복수개인 경우, 생성된 상기 섬네일을 차례대로 출력하는 방식으로 최종 섬네일을 선택하는 단계

를 더 포함하는,

섬네일 생성 방법.

청구항 21

제19항에 있어서,

상기 섬네일이 복수개인 경우, 생성된 상기 섬네일 중 하나를 최종 섬네일로서 선택하는 단계를 더 포함하는,

섬네일 생성 방법.

청구항 22

컴퓨터 실행 가능한 명령어를 저장하고 있는 컴퓨터 판독 가능 기록매체로서, 상기 컴퓨터 실행 가능한 명령어는, 프로세서에 의해 실행되면,

복수의 프레임으로 구성된 동영상으로부터, 각 프레임의 음성 및 이미지의 상관관계에 기초하여, 적어도 하나의 프레임을 선택하는 단계;

상기 선택된 적어도 하나의 각 프레임으로부터 이미지 정보와 음성 정보를 추출하는 단계; 및

상기 음성 정보로부터 음성 특징 벡터를 추출하는 음성 특징 벡터 추출 모델을 학습시키는 단계를 포함하되,

상기 음성 특징 벡터는, 기 학습된 이미지 특징 벡터 추출 모델에 의해 상기 이미지 정보로부터 추출된 이미지 특징 벡터와 소정 임베딩 공간 내에서 정렬된 것인, 방법을 상기 프로세서가 수행하도록 하는,

컴퓨터 판독 가능한 기록매체.

청구항 23

컴퓨터 판독 가능한 기록매체에 저장되어 있는 컴퓨터 프로그램으로서,

상기 컴퓨터 프로그램은, 프로세서에 의해 실행되면,

복수의 프레임으로 구성된 동영상으로부터, 각 프레임의 음성 및 이미지의 상관관계에 기초하여, 적어도 하나의 프레임을 선택하는 단계;

상기 선택된 적어도 하나의 각 프레임으로부터 이미지 정보와 음성 정보를 추출하는 단계; 및

상기 음성 정보로부터 음성 특징 벡터를 추출하는 음성 특징 벡터 추출 모델을 학습시키는 단계를 포함하되,

상기 음성 특징 벡터는, 기 학습된 이미지 특징 벡터 추출 모델에 의해 상기 이미지 정보로부터 추출된 이미지 특징 벡터와 소정 임베딩 공간 내에서 정렬된 것인, 방법을 상기 프로세서가 수행하도록 하기 위한 명령어를 포함하는,

컴퓨터 판독 가능한 기록매체에 저장되어 있는 컴퓨터 프로그램.

발명의 설명

기 술 분 야

[0001] 본 발명은 이미지를 생성하는 장치 및 딥러닝 학습 방법에 관한 것이다.

배 경 기 술

[0002] 음성-이미지 교차 모달 생성 분야는 이미지에서 음성으로의 생성과 음성에서 이미지로의 생성의 두 가지 방향으로 탐구되고 있다. 이미지에서 음성으로의 생성은 악기, 음악 및 개방형 도메인 범용 오디오 생성 관점에서 활발히 연구되어 왔다. 반면에, 음성에서 이미지로의 생성은 특정 제한된 음성 도메인에 대해서만 연구되었고, 생성된 이미지의 품질 또한 높지 않은 문제가 있다.

[0003] 모달리티를 다른 모달리티로 번역하는 것을 의미하는 교차 모달 생성에 대한 기존 연구는, 텍스트에서 이미지 또는 비디오로의 번역, 음성에서 얼굴 또는 동작으로의 번역, 이미지 또는 오디오에서 캡션으로의 번역 등 다양한 도메인에서 연구가 수행되었다. 교차 모달 생성에서 이질적인 모달리티를 연결하기 위해 기존의 사전 훈련된

모델을 활용하거나 사전 훈련된 CLIP 임베딩 공간을 확장하여 텍스트-이미지 모달리티에 맞게 조정하는 방법이 있다.

[0004] 음성을 이용하여 이미지를 조작하는 것에 대한 기존 연구는, 텍스트 기반 이미지 편집 모델을 사용하고 이를 음성-이미지 모달리티로 확장하여 임베딩 공간을 확장한 것이 있다. 이와 유사하게, 조건부 생성적 적대 신경망을 사용하여 이미지의 시각적 스타일을 음성과 일치하도록 편집하여, 음성의 볼륨 크기를 조절하거나 여러 음성을 혼합하여 이미지를 조작한 것이 있다. 다만, 이미지의 스타일만 조작할 수 있고, 텍스트 기반의 임베딩 공간이 필요하다는 한계가 있다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) 한국공개특허공보, 10-2020-0145701(2020. 12. 30. 공개)

발명의 내용

해결하려는 과제

[0006] 본 발명의 해결하고자 하는 과제는, 음성으로부터 이미지를 생성하는데 있어서 제한된 음성의 종류밖에 다룰 수 없거나 음성과 함께 항상 이미지를 입력해야 하는 종래 기술의 문제를 해결하기 위해, 음성의 종류에 상관없이 오직 음성만을 입력으로 하여 이미지를 생성하거나, 이미지와 음성을 함께 입력하여 이미지를 수정하거나, 음성에 대응하는 별도의 객체를 생성하는 등, 종래 기술과는 다른 이미지를 생성하는 장치 및 학습 방법을 제안하는 것이다.

[0007] 다만, 본 발명의 해결하고자 하는 과제는 이상에서 언급한 것으로 제한되지 않으며, 언급되지 않은 또 다른 해결하고자 하는 과제는 아래의 기재로부터 본 발명이 속하는 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0008] 일 실시예에 따른 이미지 생성 모델 학습 방법은, 복수의 프레임으로 구성된 동영상으로부터, 각 프레임의 음성 및 이미지의 상관관계에 기초하여, 적어도 하나의 프레임을 선택하는 단계, 상기 선택된 적어도 하나의 각 프레임으로부터 이미지 정보와 음성 정보를 추출하는 단계 및 상기 음성 정보로부터 음성 특징 벡터를 추출하는 음성 특징 벡터 추출 모델을 학습시키는 단계를 포함한다.

[0009] 여기서, 상기 음성 특징 벡터는, 기 학습된 이미지 특징 벡터 추출 모델에 의해 상기 이미지 정보로부터 추출된 이미지 특징 벡터와 소정 임베딩 공간 내에서 정렬된 것일 수 있다.

[0010] 상기 적어도 하나의 프레임을 선택하는 단계에서, 프레임 셀렉션 방법(frame selection method)에 의해 상기 동영상으로부터 프레임이 선택될 수 있다.

[0011] 일 실시예에 따른 이미지 생성 모델 학습 방법은, 상기 음성 특징 벡터를, 상기 이미지 특징 벡터를 기초로 이미지를 생성하도록 기 학습된 이미지 생성기에 입력하여 이미지를 생성하는 단계를 더 포함할 수 있다.

[0012] 상기 음성 특징 벡터 추출 모델을 학습시키는 단계는, 대조 학습(contrastive learning)의 방법에 의해 수행될 수 있다.

[0013] 상기 대조 학습의 방법은, InfoNCE(noise contrastive estimation)를 사용하는 것을 포함할 수 있다.

[0014] 일 실시예에 따른 이미지 생성 장치는, 제1 음성을 입력 받는 입력부, 상기 제1 음성으로부터 제1 음성 특징 벡터를 추출하는 음성 특징 벡터 추출부 및 상기 제1 음성 특징 벡터를 기초로 제1 이미지를 생성하는 이미지 생성기를 포함한다.

[0015] 여기서, 상기 음성 특징 벡터 추출부는, 복수의 프레임으로 구성된 동영상으로부터, 각 프레임의 음성 및 이미지의 상관관계에 기초하여, 적어도 하나의 프레임이 선택되고, 상기 선택된 적어도 하나의 각 프레임으로부터 제2 이미지와 제2 음성이 추출되면, 상기 제2 음성으로부터 제2 음성 특징 벡터를 추출하도록 학습된 것일 수

있다.

[0016] 여기서, 상기 제2 음성 특징 벡터는, 기 학습된 이미지 특징 벡터 추출 모델에 의해 상기 제2 이미지로부터 추출된 제2 이미지 특징 벡터와 소정 임베딩 공간 내에서 정렬된 것일 수 있다.

[0017] 여기서, 상기 이미지 생성기는, 상기 제2 이미지 특징 벡터를 기초로 제2 이미지를 생성하도록 기 학습된 것일 수 있다.

[0018] 상기 제1 음성은 상기 제2 음성과 다른 것일 수 있다.

[0019] 상기 제1 음성의 볼륨 크기가 달라지는 경우, 상기 달라진 볼륨 크기를 반영하여 상기 제1 이미지가 생성될 수 있다.

[0020] 상기 입력부는, 상기 제1 음성 또는 제3 이미지를 입력 받을 수 있고, 상기 이미지 생성기는, 상기 제1 음성과 제3 이미지가 함께 입력되면, 상기 제3 이미지에 상기 제1 이미지가 반영된 제4 이미지를 생성할 수 있다.

[0021] 상기 제4 이미지는, 상기 제3 이미지에 상기 제1 음성에 대응하는 새로운 객체가 더해져서 생성된 것일 수 있다.

[0022] 상기 제4 이미지는, 상기 제1 음성에 대응하여 상기 제3 이미지가 변경된 것일 수 있다.

[0023] 상기 제1 음성의 볼륨 크기가 달라지는 경우, 상기 달라진 볼륨 크기를 반영하여 상기 제4 이미지가 생성될 수 있다.

[0024] 상기 제1 음성은, 복수의 개체로부터 발생된 복수의 음원을 포함할 수 있고, 상기 제1 이미지는, 상기 복수의 개체를 구성하는 각 개체에 대응하는 이미지가 포함된 것일 수 있다.

[0025] 상기 제1 음성에 포함된 복수의 음원에 대응하는 각각의 볼륨 크기가 상대적으로 달라지는 경우, 상기 상대적으로 달라진 각각의 볼륨 크기를 반영하여 상기 제1 이미지가 생성될 수 있다.

[0026] 일 실시예에 따른 이미지 생성 장치는, 상기 제1 이미지를 기초로 동영상을 생성하는 동영상 생성기를 포함할 수 있다.

[0027] 여기서, 상기 입력부는, 상기 제1 음성 또는 제1 동영상을 입력 받을 수 있다.

[0028] 여기서, 상기 동영상 생성기는, 상기 제1 동영상을 구성하는 제1 복수의 이미지를 기초로, 상기 제1 음성에 대응하는 새로운 객체가 더해져서 생성된 제2 복수의 이미지로부터 제2 동영상을 생성할 수 있다.

[0029] 상기 동영상 생성기는, 상기 제1 음성의 볼륨 크기가 달라지는 경우, 상기 제1 동영상을 구성하는 제1 복수의 이미지를 기초로, 상기 달라진 볼륨 크기를 반영하여 생성된 제2 복수의 이미지로부터 제2 동영상을 생성할 수 있다.

[0030] 일 실시예에 따른 이미지 생성 장치는, 제1 음성을 입력 받는 입력부 컴퓨터 실행 가능한 명령어를 포함하는 메모리 및 상기 명령어를 실행함으로써, 기 학습된 음성 특징 벡터 추출 모델을 이용하여 상기 제1 음성으로부터 제1 음성 특징 벡터를 추출하고, 이미지 생성기를 이용하여 상기 제1 음성 특징 벡터를 기초로 제1 이미지를 생성하도록 제어하는 프로세서를 포함한다.

[0031] 여기서, 상기 음성 특징 추출 모델은, 복수의 프레임으로 구성된 동영상으로부터, 각 프레임의 음성 및 이미지의 상관관계에 기초하여, 적어도 하나의 프레임이 선택되고, 상기 선택된 적어도 하나의 각 프레임으로부터 제2 이미지와 제2 음성이 추출되면, 상기 제2 음성으로부터 제2 음성 특징 벡터를 추출하도록 학습된 것일 수 있다.

[0032] 여기서, 상기 제2 음성 특징 벡터는, 기 학습된 이미지 특징 벡터 추출 모델에 의해 상기 제2 이미지로부터 추출된 제2 이미지 특징 벡터와 소정 임베딩 공간 내에서 정렬된 것일 수 있다.

[0033] 여기서, 상기 이미지 생성기는, 상기 제2 이미지 특징 벡터를 기초로 제2 이미지를 생성하도록 기 학습된 것일 수 있다.

[0034] 일 실시예에 따른 섬네일 생성 방법은, 음성 파일을 입력하는 단계, 상기 음성 파일 내 기 정해진 시간 간격에 따라 적어도 하나의 음성 정보를 추출하는 단계, 상기 추출된 적어도 하나의 음성 정보를 기 학습된 음성 특징 벡터 추출 모델에 입력하여 적어도 하나의 음성 특징 벡터를 추출하는 단계 및 상기 음성 특징 벡터를, 기 학습된 이미지 특징 벡터 추출 모델에 의해 상기 음성 정보에 대응하는 이미지 정보로부터 추출된 이미지 특징 벡터를 기초로 이미지를 생성하도록 기 학습된 이미지 생성기에 입력하여 적어도 하나의 섬네일을 생성하는 단계를

포함한다.

[0035] 여기서, 상기 음성 특징 벡터는 상기 이미지 특징 벡터와 소정 임베딩 공간 내에서 정렬된 것일 수 있다.

[0036] 일 실시예에 따른 섬네일 생성 방법은, 상기 적어도 하나의 음성 특징 벡터에 대해, 군집화(clustering)를 통해 상기 음성 특징 벡터를 각 군집으로 분류하는 단계 및 상기 각 군집 내 대표 음성 특징 벡터를 결정하는 단계를 더 포함할 수 있다.

[0037] 여기서, 상기 섬네일을 생성하는 단계는, 상기 대표 음성 특징 벡터를 상기 이미지 생성기에 입력하여 섬네일을 생성할 수 있다.

[0038] 일 실시예에 따른 섬네일 생성 방법은, 상기 섬네일이 복수개인 경우, 생성된 상기 섬네일을 차례대로 출력하는 방식으로 최종 섬네일을 선택하는 단계를 더 포함할 수 있다.

[0039] 일 실시예에 따른 섬네일 생성 방법은, 상기 섬네일이 복수개인 경우, 생성된 상기 섬네일 중 하나를 최종 섬네일로서 선택하는 단계를 더 포함할 수 있다.

[0040] 일 실시예에 따른 컴퓨터 실행 가능한 명령어를 저장하고 있는 컴퓨터 판독 가능 기록매체는, 복수의 프레임으로 구성된 동영상으로부터, 각 프레임의 음성 및 이미지의 상관관계에 기초하여, 적어도 하나의 프레임을 선택하는 단계, 상기 선택된 적어도 하나의 각 프레임으로부터 이미지 정보와 음성 정보를 추출하는 단계 및 상기 음성 정보로부터 음성 특징 벡터를 추출하는 음성 특징 벡터 추출 모델을 학습시키는 단계를 포함하는 방법을 상기 프로세서가 수행하도록 하는 명령어를 포함한다.

[0041] 여기서, 상기 음성 특징 벡터는, 기 학습된 이미지 특징 벡터 추출 모델에 의해 상기 이미지 정보로부터 추출된 이미지 특징 벡터와 소정 임베딩 공간 내에서 정렬된 것일 수 있다.

[0042] 일 실시예에 따른 컴퓨터 판독 가능한 기록매체에 저장되어 있는 컴퓨터 프로그램은, 프로세서에 의해 실행되면, 복수의 프레임으로 구성된 동영상으로부터, 각 프레임의 음성 및 이미지의 상관관계에 기초하여, 적어도 하나의 프레임을 선택하는 단계, 상기 선택된 적어도 하나의 각 프레임으로부터 이미지 정보와 음성 정보를 추출하는 단계 및 상기 음성 정보로부터 음성 특징 벡터를 추출하는 음성 특징 벡터 추출 모델을 학습시키는 단계를 포함하는 방법을 상기 프로세서가 수행하도록 하기 위한 명령어를 포함한다.

[0043] 여기서, 상기 음성 특징 벡터는, 기 학습된 이미지 특징 벡터 추출 모델에 의해 상기 이미지 정보로부터 추출된 이미지 특징 벡터와 소정 임베딩 공간 내에서 정렬된 것일 수 있다.

발명의 효과

[0044] 본 발명의 일 실시예에 따른 이미지 생성 방법에 의하면, 음성을 입력함으로써 음성에 대응되도록 이미지를 수정하거나, 이미지를 기초로 음성을 반영한 이미지를 생성할 수 있다.

[0045] 또한, 음성에 대응 새로운 이미지를 생성하는 콘텐츠 제작 도구로 활용될 수 있다.

[0046] 또한, 텍스트 기반의 이미지-언어 임베딩 공간을 필요로 하지 않고, 레이블되지 않은 음성-이미지 쌍만을 사용하여 모델을 학습시키고, 음성에 대응하는 이미지를 생성할 수 있다.

[0047] 본 발명에서 얻을 수 있는 효과는 이상에서 언급한 효과들로 제한되지 않으며, 언급하지 않은 또 다른 효과들은 아래의 기재로부터 본 개시가 속하는 기술 분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0048] 도 1은 본 발명의 일 실시예에 따른 이미지 생성 모델 학습 장치(100)의 예시도이다.

도 2는 본 발명의 일 실시예에 따른 이미지 생성 모델 학습 장치(100)에 포함된 제어부(130)의 예시도이다.

도 3은 본 발명의 일 실시예에 따른 이미지 생성 모델 학습 방법을 예시적으로 보여주는 순서도이다.

도 4는 본 발명의 일 실시예에 따른 이미지 생성 모델 학습 방법에 따라 학습된 이후, 이미지를 생성하는 방법을 예시적으로 보여주는 순서도이다.

도 5는 본 발명의 일 실시예에 따른 이미지 생성 장치(200)의 예시도이다.

도 6은 본 발명의 또 다른 일 실시예에 따른 이미지 생성 장치(300)의 예시도이다.

도 7은 본 발명의 일 실시예에 따른 이미지 생성 방법을 구현하는 시스템의 예시도이다.

도 8은 본 발명의 일 실시예에 따른 프레임 셀렉션 방법의 개념을 보여주는 예시도이다.

도 9는 본 발명의 일 실시예에 따른 이미지 생성 방법에 따라 입력될 수 있는 입력 정보와 이를 기초로 생성될 수 있는 이미지를 보여주는 예시도이다.

도 10은 본 발명의 일 실시예에 따른 이미지 생성 방법에 따라, 동일한 이미지에 대해 다른 음성을 입력했을 때 생성되는 이미지가 다른 것과, 음성의 볼륨 크기가 달라지는 경우 이를 반영하여 이미지가 생성되는 것을 보여주는 예시도이다.

도 11은 본 발명의 일 실시예에 따른 이미지 생성 장치에 대해, 음성과 동영상이 함께 입력되고 음성의 볼륨 크기가 달라지는 경우, 이를 반영하여 동영상이 수정되는 것을 개념적으로 보여주는 예시도이다.

도 12는 본 발명의 일 실시예에 따른 이미지 생성 장치에 대해, 음성과 동영상이 함께 입력되고 음성의 볼륨 크기가 달라지는 경우, 이를 반영하여 동영상이 수정되는 것을 구체적으로 보여주는 예시도이다.

도 13은 본 발명의 일 실시예에 따른 섬네일 생성 방법을 예시적으로 보여주는 순서도이다.

도 14는 본 발명의 또 다른 일 실시예에 따른 섬네일 생성 방법을 예시적으로 보여주는 순서도이다.

도 15는 본 발명의 일 실시예에 따른 섬네일 생성 방법의 예시도이다.

발명을 실시하기 위한 구체적인 내용

[0049] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다.

[0050] 본 발명의 실시예들을 설명함에 있어서 공지 기능 또는 구성에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략할 것이다. 그리고 후술되는 용어들은 본 발명의 실시예에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.

[0051] 본 명세서에서 사용되는 용어에 대해 간략히 설명하고, 본 발명에 대해 구체적으로 설명하기로 한다.

[0052] 본 명세서에서 사용되는 용어는 본 발명의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 당 분야에 종사하는 기술자의 의도 또는 판례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 발명의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서 본 발명에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 발명의 전반에 걸친 내용을 토대로 정의되어야 한다.

[0053] 명세서 전체에서 어떤 부분이 어떤 구성요소를 '포함'한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있음을 의미한다.

[0054] 또한, 명세서에서 사용되는 '부'라는 용어는 소프트웨어 또는 FPGA나 ASIC과 같은 하드웨어 구성요소를 의미하며, '부'는 어떤 역할들을 수행한다. 그렇지만 '부'는 소프트웨어 또는 하드웨어에 한정되는 의미는 아니다. '부'는 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 따라서, 일 예로서 '부'는 소프트웨어 구성요소들, 객체지향 소프트웨어 구성요소들, 클래스 구성요소들 및 태스크 구성요소들과 같은 구성요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로 코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들 및 변수들을 포함한다. 구성요소들과 '부'들 안에서 제공되는 기능은 더 작은 수의 구성요소들 및 '부'들로 결합되거나 추가적인 구성요소들과 '부'들로 더 분리될 수 있다.

[0055] 아래에서는 첨부한 도면을 참고하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다.

[0056] 도 1은 본 발명의 일 실시예에 따른 이미지 생성 모델 학습 장치(100)의 예시도이다.

[0057] 도 1에 나타난 바와 같이, 이미지 생성 모델 학습 장치(100)는 입력부(110), 제어부(130) 및 메모리(140)를 포함하고, 통신부(140) 또는 출력부(120)를 선택적으로 더 포함할 수 있다.

[0058] 입력부(110)는 이미지 생성 모델 학습에 필요한 이미지 정보, 음성 정보 또는 동영상 정보를 입력 받을 수 있다.

[0059] 입력부(110)는 이미지 생성에 필요한 이미지 정보, 음성 정보 또는 동영상 정보를 입력 받을 수 있다.

[0060] 입력부(110)는 입력 받은 정보를 제어부(130)에 제공할 수 있다.

[0061] 입력부(110)는 이미지 생성 모델 학습에 필요한 정보 또는 이미지 생성에 필요한 정보를 입력 받을 수 있는 입력 인터페이스 또는 통신망을 통하여 수신할 수 있는 통신모듈 등을 포함할 수 있다.

[0062] 입력부(110)가 이미지 생성 모델 학습에 필요한 정보 또는 이미지 생성에 필요한 정보를 입력 받는 방법은 다양할 수 있으며 특정 수단으로 한정되지 않는다.

[0063] 출력부(120)는 생성된 이미지 또는 동영상을 사용자 인터페이스 또는 디스플레이 수단을 통해 시각적인 정보로서 표시할 수 있다.

[0064] 출력부(120)는 생성된 이미지 또는 동영상을 출력할 수 있는 출력 인터페이스 또는 데이터를 통신망을 통하여 송신할 수 있는 통신모듈 등을 포함할 수 있다.

[0065] 출력부(120)가 생성된 이미지 또는 동영상을 표시하는 방법은 다양할 수 있으며 특정 수단으로 한정되지 않는다.

[0066] 제어부(130)는 프로세서에 의해 구현될 수 있으며, 프로세서는 프로그램 내에 포함된 코드 또는 명령으로 표현된 기능을 수행하기 위해 물리적으로 구조화된 회로를 갖는, 하드웨어에 내장된 데이터 처리 장치를 의미할 수 있다. 제어부(130)는 마이크로프로세서(microprocessor), 중앙처리장치(central processing unit: CPU), 프로세서 코어(processor core), 멀티프로세서(multiprocessor), ASIC(application-specific integrated circuit), FPGA(field programmable gate array) 등의 처리 장치를 의미할 수 있으나, 상술한 실시예에 한정되지 않는다.

[0067] 제어부(130)는 이미지 생성 모델 학습 방법을 수행하도록 제어할 수 있고, 이에 대한 상세한 설명은, 도 2의 설명에서 하기로 한다.

[0068] 제어부(130)는 이미지 생성 모델 학습에 필요한 정보, 이미지 생성에 필요한 정보, 생성된 이미지 또는 동영상을 포함하는 정보를 송수신하도록 통신부(140)를 제어할 수 있다. 통신부(140)는 CDMA, GSM, W-CDMA, TD-SCDMA, WiBro, LTE, EPC, 무선 랜(wireless lan), 와이파이(wi-fi), 블루투스(bluetooth), 지그비(zigbee), WFD(wi-fi direct), UWB(ultra wide band), 적외선 통신(IrDA; infrared data association), BLE(bluetooth low energy) 또는 NFC(near field communication)와 같은 통신 방법을 채택하여 무선 통신을 수행할 수 있는 무선 통신 모듈일 수 있으나, 상술한 실시예에 한정되지 않는다.

[0069] 이미지 생성 모델 학습에 필요한 정보, 이미지 생성에 필요한 정보를 포함하는 정보를 통신부(140)를 통해 입력 받거나 내부 장치를 이용하여 직접 입력 받을 수 있다. 또한, 생성된 이미지 또는 동영상을 포함하는 정보를 통신부(140)를 통해 외부 장치로 전송하거나 내부 장치를 이용하여 직접 전송할 수 있다. 이미지 생성 모델 학습에 필요한 정보, 이미지 생성에 필요한 정보, 생성된 이미지 또는 동영상을 포함하는 정보를 입력 받거나 전송하는 방법은 상술한 실시예에 한정되지 않는다.

[0070] 메모리(150)에는 이미지 생성 모델 학습 프로그램(160) 및 이미지 생성 모델 학습 프로그램(160)의 실행에 필요한 정보가 저장될 수 있고, 제어부(130)에 의한 처리 결과가 저장될 수도 있다.

[0071] 메모리(150)에는 이미지 생성 모델 학습에 필요한 정보, 이미지 생성에 필요한 정보, 생성된 이미지 또는 동영상을 포함하는 정보가 저장될 수 있다.

[0072] 메모리(150)는 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), 플래시 메모리(flash memory)와 같은 프로그램 명령어들을 저장하고 수행하도록 특별히 구성된 하드웨어 장치 등과 같이 컴퓨터 판독 가능한 기록매체를 의미할 수 있으나, 상술한 실시예에 한정되지 않는다.

[0073] 도 2는 본 발명의 일 실시예에 따른 이미지 생성 모델 학습 장치(100)에 포함된 제어부(130)의 예시도이다.

[0074] 도 2에 나타난 바와 같이, 제어부(130)는 프레임 선택부(131), 이미지 정보 및 음성 정보 추출부(132), 이미지 특징 벡터 추출부(133), 음성 특징 벡터 추출부(134) 및 이미지 생성부(135)를 포함할 수 있다.

[0075] 제어부(130)는 본 발명의 일 실시예에 따라, 복수의 프레임으로 구성된 동영상으로부터, 각 프레임의 음성 및 이미지의 상관관계에 기초하여, 적어도 하나의 프레임을 선택하는 단계, 상기 선택된 적어도 하나의 각 프레임으로부터 이미지 정보와 음성 정보를 추출하는 단계 및 상기 음성 정보로부터 음성 특징 벡터를 추출하는 음성 특징 벡터 추출 모델을 학습시키는 단계를 포함하는 이미지 생성 모델 학습 방법을 수행하도록 제어할 수 있다.

[0076] 프레임 선택부(131)는 복수의 프레임으로 구성된 동영상으로부터, 각 프레임의 음성 및 이미지의 상관관계에 기초하여, 적어도 하나의 프레임을 선택할 수 있다. 각 프레임의 음성 및 이미지의 상관관계에 기초한다는 것은, 동영상 내에서 상관관계가 높은 음성 정보 및 이미지 정보가 포함된 프레임을 선택한다는 것을 의미할 수 있다. 프레임을 선택하는 방법으로서, 프레임 셀렉션 방법이 사용될 수 있다.

[0077] 이미지 정보 및 음성 정보 추출부(132)는 복수의 프레임으로 구성된 동영상에서 선택된 적어도 하나의 각 프레임으로부터 이미지 정보와 음성 정보를 추출할 수 있다. 각 프레임의 음성 및 이미지의 상관관계에 기초하여 프레임을 선택하였으므로, 동영상 내에서 상관관계가 높은 음성 정보 및 이미지 정보가 추출될 수 있다.

[0078] 이미지 특징 벡터 추출부(133)는 이미지 정보를 입력하여, 이에 대응하는 이미지 특징 벡터를 추출할 수 있다.

[0079] 이미지 특징 벡터 추출부(133)는 복수의 프레임으로 구성된 동영상에서 선택된 적어도 하나의 각 프레임으로부터 추출된 이미지 정보를 입력하여, 이에 대응하는 이미지 특징 벡터를 추출할 수 있다.

[0080] 이미지 특징 벡터 추출부(133)는 복수의 프레임으로 구성된 동영상에서 선택된 적어도 하나의 각 프레임으로부터 추출된 이미지 정보를 입력하여, 이에 대응하는 이미지 특징 벡터를 추출하도록 기 학습된 이미지 특징 벡터 추출 모델을 포함할 수 있다. 이미지 특징 벡터 추출 모델을 학습시키기 위한 방법으로 자기 지도 학습(self-supervised learning)이 사용될 수 있다.

[0081] 음성 특징 벡터 추출부(134)는 음성 정보를 입력하여, 이에 대응하는 음성 특징 벡터를 추출할 수 있다.

[0082] 음성 특징 벡터 추출부(134)는 복수의 프레임으로 구성된 동영상에서 선택된 적어도 하나의 각 프레임으로부터 추출된 음성 정보를 입력하여, 이에 대응하는 음성 특징 벡터를 추출할 수 있다.

[0083] 여기서, 음성 특징 벡터는, 이미지 특징 벡터 추출 모델에 의해 이미지 정보로부터 추출된 이미지 특징 벡터와 소정 임베딩 공간 내에서 정렬된 것일 수 있다.

[0084] 이미지 특징 벡터 추출부(133)에 의해 이미지 정보로부터 추출된 이미지 특징 벡터와, 음성 특징 벡터 추출부(134)에 의해 음성 정보로부터 추출된 음성 특징 벡터가 소정 임베딩 공간 내에서 잘 정렬이 되도록 음성 특징 벡터 추출부(134)를 학습시킬 수 있다. 동영상 내에서 상관관계가 높은 음성 정보 및 이미지 정보가 추출되었기 때문에, 음성 특징 벡터 추출부(134)가 이미지 특징 벡터와 잘 정렬이 된 음성 특징 벡터를 추출하고, 추출된 음성 특징 벡터를 이미지 생성부(135)에 입력하면, 음성 특징 벡터를 입력하는 것만으로도 이미지 특징 벡터를 입력한 것과 동일하거나 유사한 이미지가 생성될 수 있다.

[0085] 서로 다른 모달리티에 의해 정의된 임베딩 공간 내에서, 이미지 특징 벡터 Z^V와 음성 특징 벡터 Z^A를 정렬시키기 위한 방법으로, 양 벡터간 거리($\|Z^V - Z^A\|_2$)를 최소화하는 방법을 사용할 수 있다.

[0086] 서로 다른 모달리티에 의해 정의된 임베딩 공간 내에서, 이미지 특징 벡터 Z^V와 음성 특징 벡터 Z^A를 정렬시키기 위한 방법으로, 하기 수학식 1과 같이 InfoNCE를 대조 학습으로서 사용할 수 있다.

수학식 1

[0087]
$$\text{InfoNCE}(a_j, \{b\}_{k=1}^{N}) = -log\frac{\exp(-d(a_j, b_j))}{\sum_{k=1}^{N}\exp(-d(a_j, b_k))}$$

[0088] 여기서, ^{a}와 ^{b}는 동일 차원에서 임의의 벡터를 의미하고, $d(a,b) = \|a-b\|_2$ 이다. 이를 이용하여, 이미지 특

징 벡터 추출부(133)는 이미지 특징 벡터와 유사도가 높은 음성 특징 벡터를 추출하도록 학습될 수 있다.

[0089] 이미지 생성부(135)는 이미지 특징 벡터 추출부(133)에서 추출된 이미지 특징 벡터를 입력하여, 이에 대응하는 이미지를 생성할 수 있다.

[0090] 이미지 생성부(135)는 이미지 특징 벡터 추출부(133)에서 추출된 이미지 특징 벡터를 입력하여, 이에 대응하는 이미지를 생성하도록 기 학습될 수 있다. 이미지 생성부(135)는 조건부 생성적 적대 신경망(conditional generative adversarial network)을 포함할 수 있다. 또한 이미지 생성부(135)는 확산 모델(diffusion model)을 포함할 수 있다. 이미지 생성부(135)에 포함된 조건부 생성적 적대 신경망 또는 확산 모델을 기 학습시킬 수 있다. 이미지 생성부(135), 조건부 생성적 적대 신경망 또는 확산 모델을 학습시키기 위한 방법으로 자기 지도 학습이 사용될 수 있다.

[0091] 이미지 생성부(135)는 음성 특징 벡터 추출부(134)에서 추출된 음성 특징 벡터를 입력하여, 이에 대응하는 이미지를 생성할 수 있다.

[0092] 동영상 내에서 상관관계가 높은 음성 정보 및 이미지 정보가 추출되었기 때문에, 음성 특징 벡터 추출부(134)가 이미지 특징 벡터와 잘 정렬이 된 음성 특징 벡터를 추출하고, 추출된 음성 특징 벡터를 이미지 생성부(135)에 입력하면, 음성 특징 벡터를 입력하는 것만으로도 이미지 특징 벡터를 입력한 것과 동일하거나 유사한 이미지가 생성될 수 있다.

[0093] 도 3은 본 발명의 일 실시예에 따른 이미지 생성 모델 학습 방법을 예시적으로 보여주는 순서도이다.

[0094] 도 3에 나타난 바와 같이, 일 실시예에 따른 이미지 생성 모델 학습 방법은, 복수의 프레임으로 구성된 동영상으로부터, 각 프레임의 음성 및 이미지의 상관관계에 기초하여, 적어도 하나의 프레임을 선택하는 단계(S100), 상기 선택된 적어도 하나의 각 프레임으로부터 이미지 정보와 음성 정보를 추출하는 단계(S200), 상기 음성 정보로부터 음성 특징 벡터를 추출하는 음성 특징 벡터 추출 모델을 학습시키는 단계(S300)을 포함한다.

[0095] 여기서, 음성 특징 벡터는, 이미지 특징 벡터 추출 모델에 의해 이미지 정보로부터 추출된 이미지 특징 벡터와 소정 임베딩 공간 내에서 정렬된 것일 수 있다.

[0096] 도 4는 본 발명의 일 실시예에 따른 이미지 생성 모델 학습 방법에 따라 학습된 이후, 이미지를 생성하는 방법을 예시적으로 보여주는 순서도이다.

[0097] 도 4에 나타난 바와 같이, 일 실시예에 따른 이미지 생성 모델 학습 방법은, 음성 특징 벡터를, 이미지 특징 벡터를 기초로 이미지를 생성하도록 기 학습된 이미지 생성기에 입력하여 이미지를 생성하는 단계를 더 포함할 수 있다.

[0098] 도 5는 본 발명의 또 다른 일 실시예에 따른 이미지 생성 장치(200)의 예시도이다.

[0099] 도 5에 나타난 바와 같이, 일 실시예에 따른 이미지 생성 장치(200)는, 입력부(210), 음성 특징 벡터 추출부(230) 및 이미지 생성기(240)를 포함하고, 출력부(220)를 선택적으로 더 포함할 수 있다.

[0100] 입력부(210)는 이미지 생성 모델 학습에 필요한 이미지 정보, 음성 정보 또는 동영상 정보를 입력 받을 수 있다.

[0101] 입력부(210)는 이미지 생성에 필요한 이미지 정보, 음성 정보 또는 동영상 정보를 입력 받을 수 있다.

[0102] 입력부(210)는 이미지 생성 모델 학습에 필요한 정보 또는 이미지 생성에 필요한 정보를 입력 받을 수 있는 입력 인터페이스 또는 통신망을 통하여 수신할 수 있는 통신모듈 등을 포함할 수 있다.

[0103] 입력부(210)가 이미지 생성 모델 학습에 필요한 정보 또는 이미지 생성에 필요한 정보를 입력 받는 방법은 다양할 수 있으며 특정 수단으로 한정되지 않는다.

[0104] 출력부(220)는 생성된 이미지 또는 동영상을 사용자 인터페이스 또는 디스플레이 수단을 통해 시각적인 정보로서 표시할 수 있다.

[0105] 출력부(220)는 생성된 이미지 또는 동영상을 출력할 수 있는 출력 인터페이스 또는 데이터를 통신망을 통하여 송신할 수 있는 통신모듈 등을 포함할 수 있다.

[0106] 출력부(220)가 생성된 이미지 또는 동영상을 표시하는 방법은 다양할 수 있으며 특정 수단으로 한정되지 않는다.

[0107] 음성 특징 벡터 추출부(230)는 음성 정보를 입력하여, 이에 대응하는 음성 특징 벡터를 추출할 수 있다.

[0108] 음성 특징 벡터 추출부(230)는 복수의 프레임으로 구성된 동영상에서 선택된 적어도 하나의 각 프레임으로부터 추출된 음성 정보를 입력하여, 이에 대응하는 음성 특징 벡터를 추출할 수 있다.

[0109] 여기서, 음성 특징 벡터는, 복수의 프레임으로 구성된 동영상에서 선택된 적어도 하나의 각 프레임으로부터 추출된 이미지 정보를, 기 학습된 이미지 특징 벡터 추출 모델 입력하여 추출된 이미지 특징 벡터와 소정 임베딩 공간 내에서 정렬된 것일 수 있다.

[0110] 기 학습된 이미지 특징 벡터 추출 모델에 의해 이미지 정보로부터 추출된 이미지 특징 벡터와, 음성 특징 벡터 추출부(230)에 의해 음성 정보로부터 추출된 음성 특징 벡터가 소정 임베딩 공간 내에서 잘 정렬이 되도록 음성 특징 벡터 추출부(230)를 학습시킬 수 있다. 동영상 내에서 상관관계가 높은 음성 정보 및 이미지 정보가 추출되었기 때문에, 음성 특징 벡터 추출부(230)가 이미지 특징 벡터와 잘 정렬이 된 음성 특징 벡터를 추출하고, 추출된 음성 특징 벡터를 이미지 생성기(240)에 입력하면, 음성 특징 벡터를 입력하는 것만으로도 이미지 특징 벡터를 입력한 것과 동일하거나 유사한 이미지가 생성될 수 있다.

[0111] 서로 다른 모달리티에 의해 정의된 임베딩 공간 내에서, 이미지 특징 벡터 z^V와 음성 특징 벡터 z^A를 정렬시키기 위한 방법으로, 양 벡터간 거리($\|z^V - z^A\|_2$)를 최소화하는 방법을 사용할 수 있다.

[0112] 서로 다른 모달리티에 의해 정의된 임베딩 공간 내에서, 이미지 특징 벡터 z^V와 음성 특징 벡터 z^A를 정렬시키기 위한 방법으로, 상기 수학식 1과 같이 InfoNCE를 대조 학습으로서 사용할 수 있다.

[0113] 이미지 생성기(240)는 기 학습된 이미지 특징 벡터 추출 모델에서 추출된 이미지 특징 벡터를 입력하여, 이에 대응하는 이미지를 생성하도록 기 학습될 수 있다. 이미지 생성기(240)는 조건부 생성적 적대 신경망을 포함할 수 있다. 또한 이미지 생성기(240)는 확산 모델을 포함할 수 있다. 이미지 생성기(240)에 포함된 조건부 생성적 적대 신경망 또는 확산 모델을 기 학습시킬 수 있다. 이미지 생성기(240), 조건부 생성적 적대 신경망 또는 확산 모델을 학습시키기 위한 방법으로 자기 지도 학습이 사용될 수 있다.

[0114] 이미지 생성기(240)는 음성 특징 벡터 추출부(230)에서 추출된 음성 특징 벡터를 입력하여, 이에 대응하는 이미지를 생성할 수 있다.

[0115] 동영상 내에서 상관관계가 높은 음성 정보 및 이미지 정보가 추출되었기 때문에, 음성 특징 벡터 추출부(230)가 이미지 특징 벡터와 잘 정렬이 된 음성 특징 벡터를 추출하고, 추출된 음성 특징 벡터를 이미지 생성기(240)에 입력하면, 음성 특징 벡터를 입력하는 것만으로도 이미지 특징 벡터를 입력한 것과 동일하거나 유사한 이미지가 생성될 수 있다.

[0116] 다시 도 5를 참조하면, 일 실시예에 따른 이미지 생성 장치(200)는, 제1 음성을 입력 받는 입력부(210), 상기 제1 음성으로부터 제1 음성 특징 벡터를 추출하는 음성 특징 벡터 추출부(230) 및 상기 제1 음성 특징 벡터를 기초로 제1 이미지를 생성하는 이미지 생성기(240)를 포함하고, 선택적으로 출력부(220)를 더 포함할 수 있다.

[0117] 여기서, 상기 음성 특징 벡터 추출부(230)는, 복수의 프레임으로 구성된 동영상으로부터, 각 프레임의 음성 및 이미지의 상관관계에 기초하여, 적어도 하나의 프레임이 선택되고, 상기 선택된 적어도 하나의 각 프레임으로부터 제2 이미지와 제2 음성이 추출되면, 상기 제2 음성으로부터 제2 음성 특징 벡터를 추출하도록 학습된 것일 수 있다.

[0118] 여기서, 상기 제2 음성 특징 벡터는, 기 학습된 이미지 특징 벡터 추출 모델에 의해 상기 제2 이미지로부터 추출된 제2 이미지 특징 벡터와 소정 임베딩 공간 내에서 정렬된 것일 수 있다.

[0119] 여기서, 상기 이미지 생성기(240)는, 상기 제2 이미지 특징 벡터를 기초로 제2 이미지를 생성하도록 기 학습된 것일 수 있다.

[0120] 상기 제1 음성은 상기 제2 음성과 다른 것일 수 있다. 따라서, 학습에 사용되지 않은 음성 데이터를, 학습된 모델에 입력하여도 이에 상응하는 이미지를 출력할 수 있다. 이는 레이블링 없이 모델 학습을 시킬 수 있다는 것을 의미할 수 있다.

[0121] 상기 제1 음성의 볼륨 크기가 달라지는 경우, 상기 달라진 볼륨 크기를 반영하여 상기 제1 이미지가 생성될 수 있다.

[0122] 상기 입력부는, 상기 제1 음성 또는 제3 이미지를 입력 받을 수 있고, 상기 이미지 생성기는, 상기 제1 음성과 제3 이미지가 함께 입력되면, 상기 제3 이미지에 상기 제1 이미지가 반영된 제4 이미지를 생성할 수 있다.

[0123] 상기 제4 이미지는, 상기 제3 이미지에 상기 제1 음성에 대응하는 새로운 객체가 더해져서 생성된 것일 수 있다.

[0124] 상기 제4 이미지는, 상기 제1 음성에 대응하여 상기 제3 이미지가 변경된 것일 수 있다.

[0125] 상기 제1 음성의 볼륨 크기가 달라지는 경우, 상기 달라진 볼륨 크기를 반영하여 상기 제4 이미지가 생성될 수 있다.

[0126] 상기 제1 음성은, 서로 다른 주파수 대역을 가지는 복수의 음성 정보를 포함한 것일 수 있다.

[0127] 상기 제1 음성은, 복수의 개체로부터 발생된 복수의 음원을 포함하고, 상기 제1 이미지는, 상기 복수의 개체를 구성하는 각 개체에 대응하는 이미지가 포함된 것일 수 있다.

[0128] 상기 제4 이미지는, 상기 제3 이미지에 상기 제1 음성에 대응하는 새로운 객체가 더해져서 생성된 것일 수 있다. 따라서, 다른 종류의 음성이 혼합된 음성 정보를 입력하여도, 생성된 이미지는 객체가 섞이지 않고 각 객체가 구별되어 표현된 이미지일 수 있다.

[0129] 상기 제4 이미지는, 상기 제1 음성에 대응하여 상기 제3 이미지가 변경된 것일 수 있다. 따라서, 입력된 이미지 정보를 기초로, 입력된 음성 정보를 반영하여 이미지가 생성될 수 있다.

[0130] 상기 제1 음성의 볼륨 크기가 달라지는 경우, 상기 달라진 볼륨 크기를 반영하여 상기 제4 이미지가 생성될 수 있다.

[0131] 상기 제1 음성은, 복수의 개체로부터 발생된 복수의 음원을 포함하고, 상기 제1 이미지는, 상기 복수의 개체를 구성하는 각 개체에 대응하는 이미지가 포함된 것일 수 있다.

[0132] 상기 제1 음성에 포함된 복수의 음원에 대응하는 각각의 볼륨 크기가 상대적으로 달라지는 경우, 상기 상대적으로 달라진 각각의 볼륨 크기를 반영하여 상기 제1 이미지가 생성될 수 있다.

[0133] 일 실시예에 따른 이미지 생성 장치(200)는, 상기 제1 이미지를 기초로 동영상을 생성하는 동영상 생성기를 포함할 수 있다.

[0134] 여기서, 상기 입력부는, 상기 제1 음성 또는 제1 동영상을 입력 받을 수 있다.

[0135] 여기서, 상기 동영상 생성기는, 상기 제1 동영상을 구성하는 제1 복수의 이미지를 기초로, 상기 제1 음성에 대응하는 새로운 객체가 더해져서 생성된 제2 복수의 이미지로부터 제2 동영상을 생성할 수 있다.

[0136] 상기 동영상 생성기는, 상기 제1 음성의 볼륨 크기가 달라지는 경우, 상기 제1 동영상을 구성하는 제1 복수의 이미지를 기초로, 상기 달라진 볼륨 크기를 반영하여 생성된 제2 복수의 이미지로부터 제2 동영상을 생성할 수 있다.

[0137] 음성 정보를 반영하여 동영상을 수정하는 방법은, 도 11 내지 도 12에서 구체적으로 설명하기로 한다.

[0138] 도 6은 본 발명의 또 다른 일 실시예에 따른 이미지 생성 장치(300)의 예시도이다.

[0139] 도 6에 나타난 바와 같이, 일 실시예에 따른 이미지 생성 장치(300)는, 입력부(310), 제어부(330) 및 메모리(350)를 포함하고, 출력부(320) 또는 통신부(340)를 선택적으로 더 포함할 수 있다.

[0140] 입력부(310)는 이미지 생성 모델 학습에 필요한 이미지 정보, 음성 정보 또는 동영상 정보를 입력 받을 수 있다.

[0141] 입력부(310)는 이미지 생성에 필요한 이미지 정보, 음성 정보 또는 동영상 정보를 입력 받을 수 있다.

[0142] 입력부(310)는 이미지 생성 모델 학습에 필요한 정보 또는 이미지 생성에 필요한 정보를 입력 받을 수 있는 입력 인터페이스 또는 통신망을 통하여 수신할 수 있는 통신모듈 등을 포함할 수 있다.

[0143] 입력부(310)가 이미지 생성 모델 학습에 필요한 정보 또는 이미지 생성에 필요한 정보를 입력 받는 방법은 다양할 수 있으며 특정 수단으로 한정되지 않는다.

[0144] 출력부(320)는 생성된 이미지 또는 동영상을 사용자 인터페이스 또는 디스플레이 수단을 통해 시각적인 정보로

서 표시할 수 있다.

[0145] 출력부(320)는 생성된 이미지 또는 동영상을 출력할 수 있는 출력 인터페이스 또는 데이터를 통신망을 통하여 송신할 수 있는 통신모듈 등을 포함할 수 있다.

[0146] 출력부(320)가 생성된 이미지 또는 동영상을 표시하는 방법은 다양할 수 있으며 특정 수단으로 한정되지 않는다.

[0147] 제어부(330)는 프로세서를 의미할 수 있으며, 프로그램 내에 포함된 코드 또는 명령으로 표현된 기능을 수행하기 위해 물리적으로 구조화된 회로를 갖는, 하드웨어에 내장된 데이터 처리 장치를 의미할 수 있다. 제어부(330)는 마이크로프로세서(microprocessor), 중앙처리장치(central processing unit: CPU), 프로세서 코어(processor core), 멀티프로세서(multiprocessor), ASIC(application-specific integrated circuit), FPGA(field programmable gate array) 등의 처리 장치를 의미할 수 있으나, 상술한 실시예에 한정되지 않는다.

[0148] 제어부(330)는 일 실시예에 따른 이미지 생성 방법을 수행하도록 제어할 수 있다.

[0149] 제어부(330)는 이미지 생성 모델 학습에 필요한 정보, 이미지 생성에 필요한 정보, 생성된 이미지 또는 동영상을 포함하는 정보를 송수신하도록 통신부(340)를 제어할 수 있다. 통신부(140)는 CDMA, GSM, W-CDMA, TD-SCDMA, WiBro, LTE, EPC, 무선 랜(wireless lan), 와이파이(wi-fi), 블루투스(bluetooth), 지그비(zigbee), WFD(wi-fi direct), UWB(ultra wide band), 적외선 통신(IrDA; infrared data association), BLE(bluetooth low energy) 또는 NFC(near field communication)와 같은 통신 방법을 채택하여 무선 통신을 수행할 수 있는 무선 통신 모듈일 수 있으나, 상술한 실시예에 한정되지 않는다.

[0150] 이미지 생성 모델 학습에 필요한 정보, 이미지 생성에 필요한 정보를 포함하는 정보를 통신부(340)를 통해 입력받거나 내부 장치를 이용하여 직접 입력 받을 수 있다. 또한, 생성된 이미지 또는 동영상을 포함하는 정보를 통신부(340)를 통해 외부 장치로 전송하거나 내부 장치를 이용하여 직접 전송할 수 있다. 이미지 생성 모델 학습에 필요한 정보, 이미지 생성에 필요한 정보, 생성된 이미지 또는 동영상을 포함하는 정보를 입력 받거나 전송하는 방법은 상술한 실시예에 한정되지 않는다.

[0151] 메모리(350)에는 이미지 생성 프로그램(360) 및 이미지 생성 프로그램(360)의 실행에 필요한 정보가 저장될 수 있고, 제어부(330)에 의한 처리 결과가 저장될 수도 있다.

[0152] 메모리(350)에는 이미지 생성 모델 학습에 필요한 정보, 이미지 생성에 필요한 정보, 생성된 이미지 또는 동영상을 포함하는 정보가 저장될 수 있다.

[0153] 메모리(350)는 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), 플래시 메모리(flash memory)와 같은 프로그램 명령어들을 저장하고 수행하도록 특별히 구성된 하드웨어 장치 등과 같이 컴퓨터 판독 가능한 기록매체를 의미할 수 있으나, 상술한 실시예에 한정되지 않는다.

[0154] 도 7은 본 발명의 일 실시예에 따른 이미지 생성 방법을 구현하는 시스템의 예시도이다.

[0155] 먼저, 복수의 프레임으로 구성된 동영상으로부터, 각 프레임의 음성 및 이미지의 상관관계에 기초하여, 적어도 하나의 프레임을 선택할 수 있다. 각 프레임의 음성 및 이미지의 상관관계에 기초한다는 것은, 동영상 내에서 상관관계가 높은 음성 정보 및 이미지 정보가 포함된 프레임을 선택한다는 것을 의미할 수 있다. 프레임을 선택하는 방법으로서, 프레임 셀렉션 방법이 사용될 수 있다.

[0156] 다음으로, 복수의 프레임으로 구성된 동영상에서 선택된 적어도 하나의 각 프레임으로부터 이미지 정보와 음성 정보를 추출할 수 있다. 각 프레임의 음성 및 이미지의 상관관계에 기초하여 프레임을 선택하였으므로, 동영상 내에서 상관관계가 높은 음성 정보 및 이미지 정보가 추출될 수 있다.

[0157] 다음으로, 복수의 프레임으로 구성된 동영상에서 선택된 적어도 하나의 각 프레임으로부터 추출된 이미지 정보를 이미지 특징 벡터 추출 모델에 입력하여, 이에 대응하는 이미지 특징 벡터를 추출하도록 기 학습시킬 수 있다.

[0158] 다음으로, 이미지 특징 벡터 추출 모델에서 추출된 이미지 특징 벡터를 이미지 생성기에 입력하여, 이에 대응하는 이미지를 생성하도록 기 학습시킬 수 있다.

[0159] 다음으로, 이미지 특징 벡터 추출 모델에 의해 이미지 정보로부터 추출된 이미지 특징 벡터와, 음성 특징 벡터

추출 모델에 의해 음성 정보로부터 추출된 음성 특징 벡터가 소정 임베딩 공간 내에서 잘 정렬이 되도록 음성 특징 벡터 추출 모델을 학습시킬 수 있다. 동영상 내에서 상관관계가 높은 음성 정보 및 이미지 정보가 추출되었기 때문에, 음성 특징 벡터 추출 모델이 이미지 특징 벡터와 잘 정렬이 된 음성 특징 벡터를 추출하고, 추출된 음성 특징 벡터를 이미지 생성기에 입력하면, 음성 특징 벡터를 입력하는 것만으로도 이미지 특징 벡터를 입력한 것과 동일하거나 유사한 이미지가 생성될 수 있다.

[0160] 도 8은 본 발명의 일 실시예에 따른 프레임 셀렉션 방법의 개념을 보여주는 예시도이다.

[0161] 프레임 셀렉션 방법은 복수의 프레임으로 구성된 동영상으로부터, 각 프레임의 음성 및 이미지의 상관관계에 기초하여, 상관관계가 높은 음성 정보와 이미지 정보 쌍이 포함된 적어도 하나의 프레임을 선택하는 방법으로 사용할 수 있다.

[0162] 도 8에 나타난 바와 같이, 기차의 이미지 정보가 포함된 동영상에서, 기차의 소리가 포함된 음성 정보와 상관관계가 높은 이미지 정보에 기초하여, 기차 소리에 대응하는 기차 이미지가 포함된 프레임을 선택할 수 있다.

[0163] 도 9는 본 발명의 일 실시예에 따른 이미지 생성 방법에 따라 입력될 수 있는 입력 정보와 이를 기초로 생성될 수 있는 이미지를 보여주는 예시도이다.

[0164] 도 9에 나타난 바와 같이, 다양한 특성의 음성을 입력하여 이미지를 생성할 수 있다.

[0165] 단일 음성(single waveform) 정보를 본 발명의 일 실시예에 따른 이미지 생성 장치에 입력하면, 그에 대응하는 이미지들을 생성할 수 있다.

[0166] 혼합된 음성(mixing waveforms) 정보를 본 발명의 일 실시예에 따른 이미지 생성 장치에 입력하면, 두 가지 음성 정보가 모두 반영된 이미지를 생성할 수 있다. 예를 들면, 도 9에 나타난 바와 같이, 강아지와 물소리를 섞은 음성 정보를 이미지 생성 장치에 입력하면, 강아지와 물에 대한 이미지 정보가 모두 포함된 이미지가 생성될 수 있다.

[0167] 본 발명의 일 실시예에 따른 이미지 생성 장치에 입력되는 음성 정보의 볼륨 크기가 달라지는 경우(volume changes), 동일한 객체에 대한 이미지 정보가 존재하지만, 달라진 볼륨 크기가 반영된 이미지를 생성할 수 있다. 예를 들면, 도 9에 나타난 바와 같이, 물 소리가 커지는 경우, 물의 흐름이 세지는 이미지가 생성될 수 있고, 엘크의 소리가 커지는 경우, 엘크의 크기가 커지는 이미지가 생성될 수 있다.

[0168] 혼합된 음성(mixing waveforms) 정보가 본 발명의 일 실시예에 따른 이미지 생성 장치에 입력되는 경우에도, 혼합된 음성 정보의 볼륨 크기가 달라지는 경우, 달라진 볼륨 크기를 반영하여 이미지가 생성될 수 있다. 예를 들면, 도 9에 나타난 바와 같이, 강아지와 바람 소리의 상대적인 볼륨 크기에 따라, 강아지의 크기 및 배경의 크기가 달라진 이미지가 생성될 수 있다.

[0169] 도 10은 본 발명의 일 실시예에 따른 이미지 생성 방법에 따라, 동일한 이미지에 대해 다른 음성을 입력했을 때 생성되는 이미지가 다른 것과, 음성의 볼륨 크기가 달라지는 경우 이를 반영하여 이미지가 생성되는 것을 보여주는 예시도이다.

[0170] 도 10에 나타난 바와 같이, 본 발명의 일 실시예에 따른 이미지 생성 장치에 음성 정보뿐만 아니라, 음성 정보와 이미지 정보를 동시에 입력하여 이미지를 생성할 수 있다.

[0171] 예를 들면, 건물의 이미지와 환호하는 소리를 입력하는 경우, 환호하는 소리에 대응하여 건물이 번쩍이는 이미지가 생성될 수 있다. 이 때, 도 10에 나타난 바와 같이, 동일한 건물의 이미지에 대해, 환호하는 소리의 종류, 위치, 볼륨 크기 등에 따라, 이미지가 다르게 생성될 수 있다.

[0172] 또 다른 예를 들면, 해변가의 이미지와 트랙터 소리를 입력하는 경우, 해변가의 이미지에 트랙터의 이미지가 반영되어 이미지가 생성될 수 있다. 이 때, 도 10에 나타난 바와 같이, 동일한 해변가의 이미지에 대해, 트랙터 소리의 종류, 위치, 볼륨 크기 등에 따라, 이미지가 다르게 생성될 수 있다.

[0173] 또한, 입력된 이미지 정보에 대해, 입력된 음성 정보의 볼륨 크기가 달라지는 경우, 달라진 볼륨 크기를 반영하여 이미지가 생성될 수 있다. 예를 들면, 도 10에 나타난 바와 같이, 입력된 물이 흐르는 소리에 대한 볼륨 크기를 다르게 하는 경우, 달라진 볼륨 크기를 반영하여 물살이 바뀌는 이미지가 생성될 수 있다.

[0174] 도 11은 본 발명의 일 실시예에 따른 이미지 생성 장치에 대해, 음성과 동영상이 함께 입력되고 음성의 볼륨 크기가 달라지는 경우, 이를 반영하여 동영상이 수정되는 것을 개념적으로 보여주는 예시도이다.

[0175] 도 11에 나타난 바와 같이, 본 발명의 일 실시예에 따른 이미지 생성 장치에 동영상과 음성을 입력할 수 있다. 이 때, 입력된 음성의 볼륨 크기가 달라지는 경우, 입력된 동영상을 구성하는 각 프레임의 이미지가, 달라진 볼륨 크기를 반영하여 생성될 수 있고, 생성된 이미지로부터 달라진 볼륨 크기가 반영된 동영상이 생성될 수 있다. 따라서, 입력하는 음성의 볼륨 크기를 다르게 함으로써, 동영상을 수정할 수 있다.

[0176] 도 12는 본 발명의 일 실시예에 따른 이미지 생성 장치에 대해, 음성과 동영상이 함께 입력되고 음성의 볼륨 크기가 달라지는 경우, 이를 반영하여 동영상이 수정되는 것을 구체적으로 보여주는 예시도이다.

[0177] 도 12에 나타난 바와 같이, 입력된 음성의 볼륨 크기가 달라지는 경우, 이를 반영하여 동영상이 수정되는 것을 구체적으로 볼 수 있다.

[0178] 도 13은 본 발명의 일 실시예에 따른 섬네일 생성 방법에 대한 예시도이다.

[0179] 도 13에 나타난 바와 같이, 본 발명의 일 실시예에 따른 섬네일 생성 방법은, 음성 파일을 입력하는 단계(S500), 상기 음성 파일 내 기 정해진 시간 간격에 따라 적어도 하나의 음성 정보를 추출하는 단계(S600), 상기 추출된 적어도 하나의 음성 정보를 기 학습된 음성 특징 벡터 추출 모델에 입력하여 적어도 하나의 음성 특징 벡터를 추출하는 단계(S700) 및 상기 음성 특징 벡터를, 기 학습된 이미지 특징 벡터 추출 모델에 의해 상기 음성 정보에 대응하는 이미지 정보로부터 추출된 이미지 특징 벡터를 기초로 이미지를 생성하도록 기 학습된 이미지 생성기에 입력하여 적어도 하나의 섬네일을 생성하는 단계(S800)를 포함한다.

[0180] 여기서, 상기 음성 특징 벡터는 상기 이미지 특징 벡터와 소정 임베딩 공간 내에서 정렬된 것일 수 있다.

[0181] 도 14는 본 발명의 또 다른 일 실시예에 따른 섬네일 생성 방법에 대한 예시도이다.

[0182] 도 14에 나타난 바와 같이, 본 발명의 일 실시예에 따른 섬네일 생성 방법은, 상기 적어도 하나의 음성 특징 벡터에 대해, 군집화를 통해 상기 음성 특징 벡터를 각 군집으로 분류하는 단계(S900) 및 상기 각 군집 내 대표 음성 특징 벡터를 결정하는 단계(S1000)를 더 포함할 수 있다.

[0183] 여기서, 상기 대표 음성 특징 벡터를 상기 이미지 생성기에 입력하여 섬네일을 생성할 수 있다(S1100).

[0184] 본 발명의 일 실시예에 따른 섬네일 생성 방법은, 상기 섬네일이 복수개인 경우, 생성된 상기 섬네일을 차례대로 출력하는 방식으로 최종 섬네일을 선택하는 단계(S1200)를 더 포함할 수 있다.

[0185] 본 발명의 일 실시예에 따른 섬네일 생성 방법은, 상기 섬네일이 복수개인 경우, 생성된 상기 섬네일 중 하나를 최종 섬네일로서 선택하는 단계(S1200)를 더 포함할 수 있다.

[0186] 도 15는 본 발명의 일 실시예에 따른 섬네일 생성 방법의 예시도이다.

[0187] 먼저, 음성 파일을 수신하면, 수신한 음성 파일 내 음성 정보에 대해, N초씩 겹치도록 T초 단위로 적어도 하나의 음성 정보를 추출할 수 있다.

[0188] 추출한 적어도 하나의 음성 정보를, 기 학습된 음성 특징 벡터 추출 모델에 입력하여 적어도 하나의 음성 특징 벡터를 추출할 수 있다.

[0189] 추출된 적어도 하나의 음성 특징 벡터에 대해, K평균 군집화를 수행하여 음성 특징 벡터를 각 군집으로 분류하고, 군집 내 대표 음성 특징 벡터를 결정할 수 있다.

[0190] 군집 내 대표 음성 특징 벡터를 기 학습된 이미지 생성기에 입력하여 섬네일을 생성할 수 있다.

[0191] 섬네일이 복수개인 경우, 생성된 상기 섬네일을 차례대로 출력하는 방식으로 최종 섬네일을 선택할 수 있다. 또한, 생성된 상기 섬네일 중 하나를 최종 섬네일로서 선택할 수 있다. 이 경우, 최대 군집의 대표 이미지를 최종 섬네일로서 선택할 수 있다.

[0192] 본 발명에 첨부된 블록도의 각 블록과 흐름도의 각 단계의 조합들은 컴퓨터 프로그램 인스트럭션들에 의해 수행될 수도 있다. 이들 컴퓨터 프로그램 인스트럭션들은 범용 컴퓨터, 특수용 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장비의 인코딩 프로세서에 탑재될 수 있으므로, 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장비의 인코딩 프로세서를 통해 수행되는 그 인스트럭션들이 블록도의 각 블록 또는 흐름도의 각 단계에서 설명된 기능들을 수행하는 수단을 생성하게 된다. 이들 컴퓨터 프로그램 인스트럭션들은 특정 방법으로 기능을 구현하기 위해 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장비를 지향할 수 있는 컴퓨터 이용 가능 또는 컴퓨터 판독 가능 메모리에 저장되는 것도 가능하므로, 그 컴퓨터 이용가능 또는 컴퓨터 판독 가능 메모리

에 저장된 인스트럭션들은 블록도의 각 블록 또는 흐름도 각 단계에서 설명된 기능을 수행하는 인스트럭션 수단을 내포하는 제조 품목을 생산하는 것도 가능하다. 컴퓨터 프로그램 인스트럭션들은 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장비 상에 탑재되는 것도 가능하므로, 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장비 상에서 일련의 동작 단계들이 수행되어 컴퓨터로 실행되는 프로세스를 생성해서 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장비를 수행하는 인스트럭션들은 블록도의 각 블록 및 흐름도의 각 단계에서 설명된 기능들을 실행하기 위한 단계들을 제공하는 것도 가능하다.

[0193] 또한, 각 블록 또는 각 단계는 특정된 논리적 기능(들)을 실행하기 위한 하나 이상의 실행 가능한 인스트럭션들을 포함하는 모듈, 세그먼트 또는 코드의 일부를 나타낼 수 있다. 또, 몇 가지 대체 실시예들에서는 블록들 또는 단계들에서 언급된 기능들이 순서를 벗어나서 발생하는 것도 가능함을 주목해야 한다. 예컨대, 잇달아 도시되어 있는 두 개의 블록들 또는 단계들은 사실 실질적으로 동시에 수행되는 것도 가능하고 또는 그 블록들 또는 단계들이 때때로 해당하는 기능에 따라 역순으로 수행되는 것도 가능하다.

[0194] 이상의 설명은 본 발명의 기술 사상을 예시적으로 설명한 것에 불과한 것으로서, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자라면 본 발명의 본질적인 품질에서 벗어나지 않는 범위에서 다양한 수정 및 변형이 가능할 것이다. 따라서, 본 발명에 개시된 실시예들은 본 발명의 기술 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이고, 이러한 실시예에 의하여 본 발명의 기술 사상의 범위가 한정되는 것은 아니다. 본 발명의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 균등한 범위 내에 있는 모든 기술사상은 본 발명의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

[0195] 100: 이미지 생성 모델 학습 장치

110: 입력부

120: 출력부

130: 제어부

140: 통신부

150: 메모리

160: 이미지 생성 모델 학습 프로그램

131: 프레임 선택부

132: 이미지 정보 및 음성 정보 추출부

133: 이미지 특징 벡터 추출부

134: 음성 특징 벡터 추출부

135: 이미지 생성부

200: 이미지 생성 장치

210: 입력부

220: 출력부

230: 음성 특징 벡터 추출부

240: 이미지 생성기

300: 이미지 생성 장치

310: 입력부

320: 출력부

330: 제어부

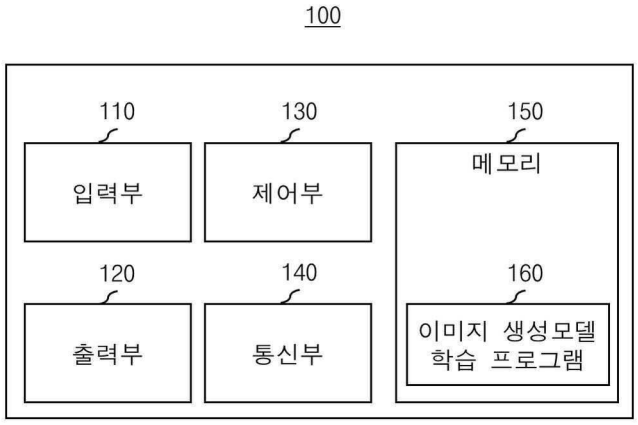
340: 통신부

350: 메모리

360: 이미지 생성 프로그램

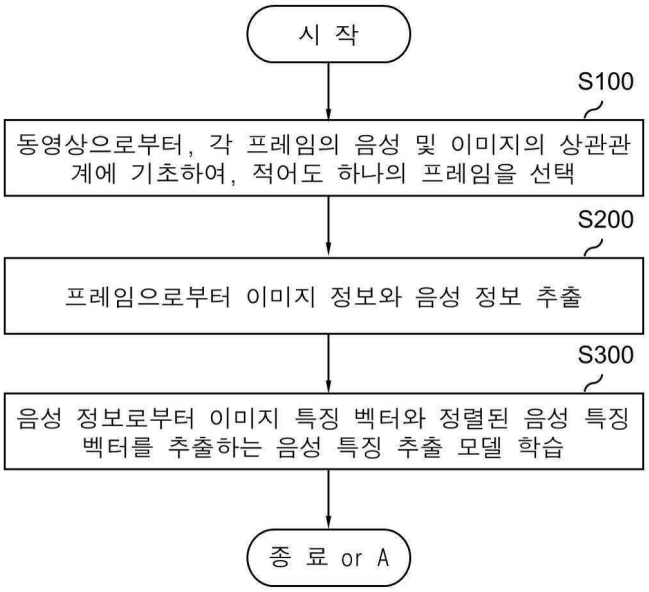
도면

도면1

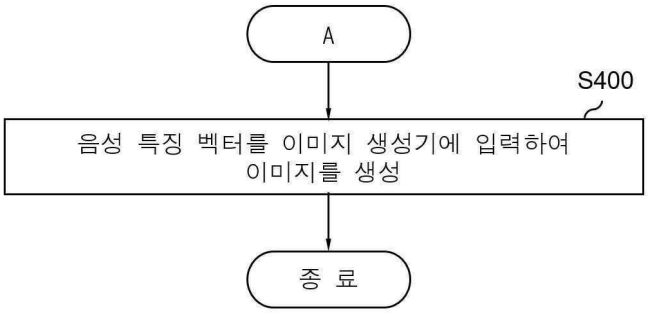


도면2

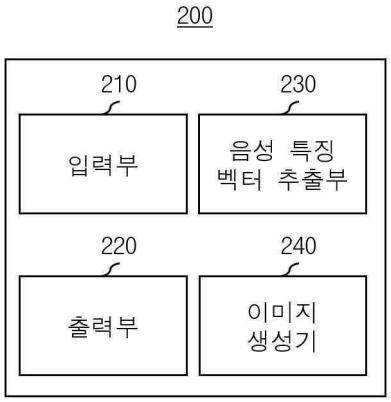
도면3



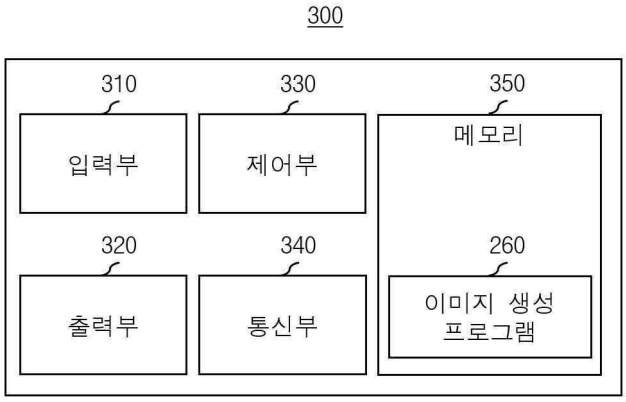
도면4


A
S400
음성 특징 벡터를 이미지 생성기에 입력하여
이미지를 생성
종 료

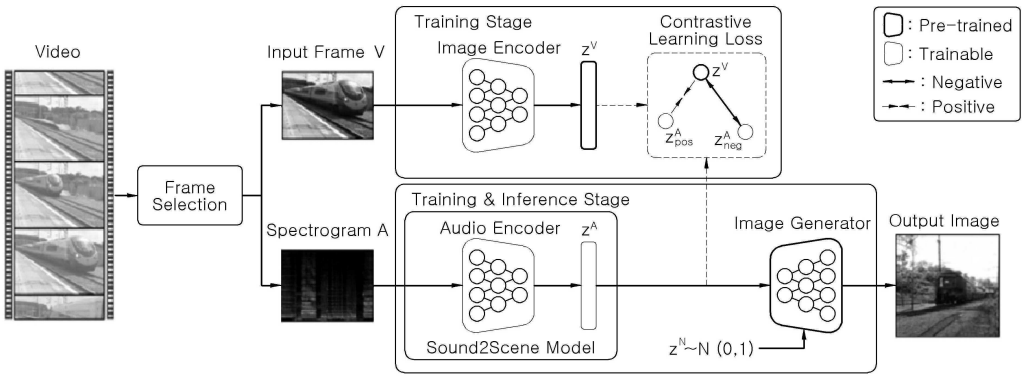
도면5


200
210
입력부
230
음성 특징
벡터 추출부
220
출력부
240
이미지
생성기

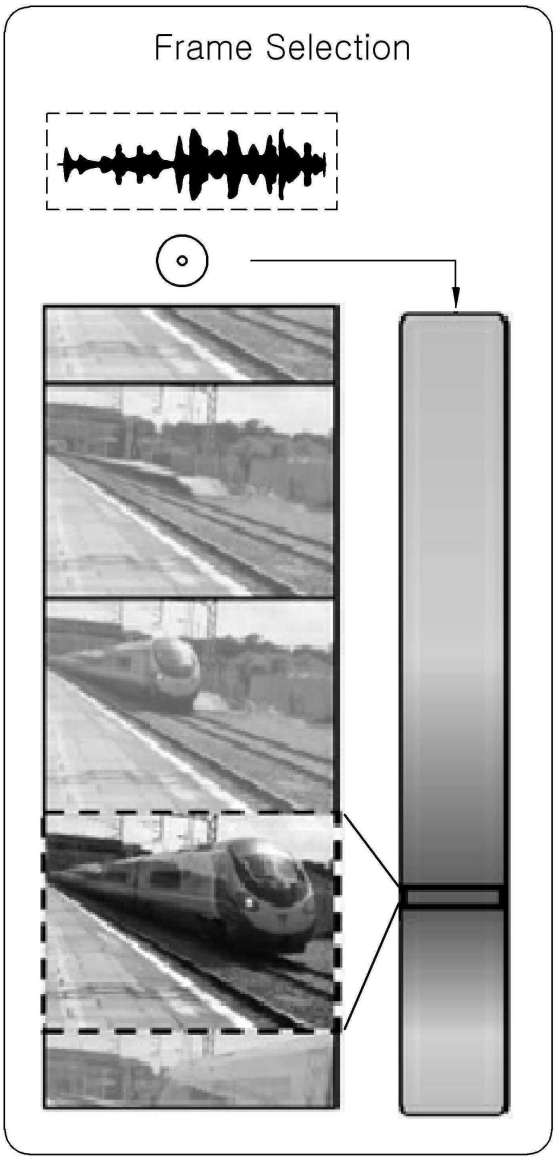
도면6


300
310
입력부
330
제어부
350
메모리
320
출력부
340
통신부
260
이미지 생성
프로그램

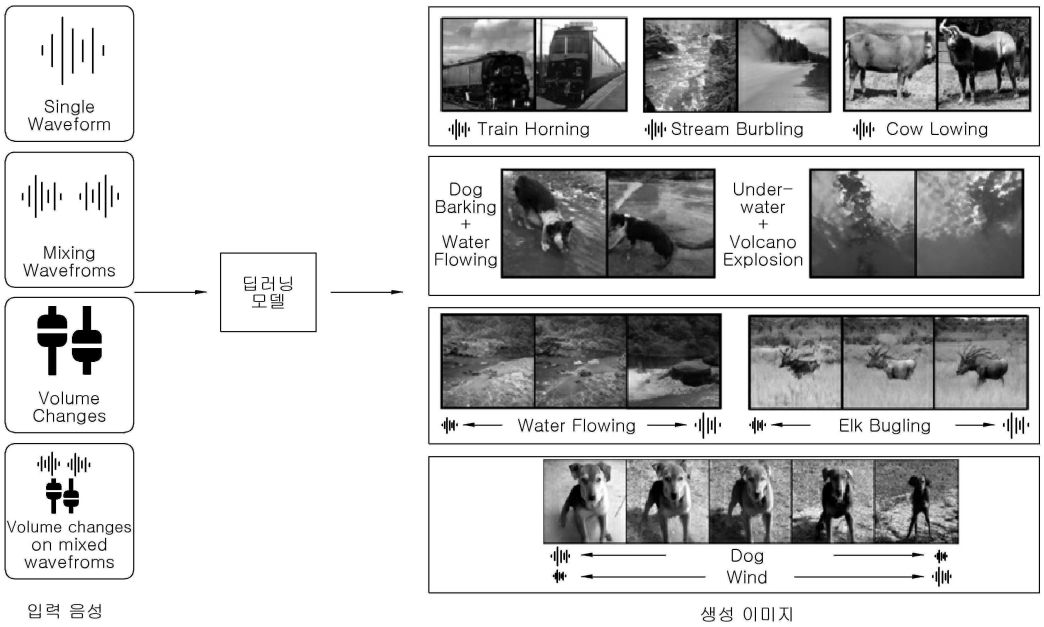
도면7


Video
Frame Selection
Input Frame V
Spectrogram A
Training Stage
Image Encoder
z^V
Contrastive Learning Loss
z^V
z^A_{pos}
z^A_{neg}
Training & Inference Stage
Audio Encoder
z^A
Sound2Scene Model
Image Generator
z^N~N (0,1)
Output Image
: Pre-trained
: Trainable
: Negative
: Positive

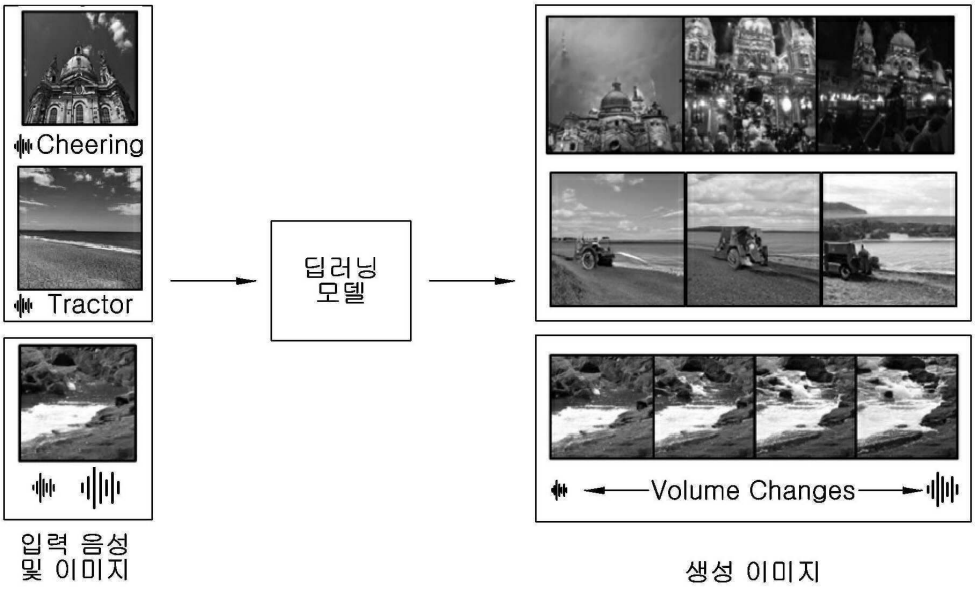
도면8


Frame Selection

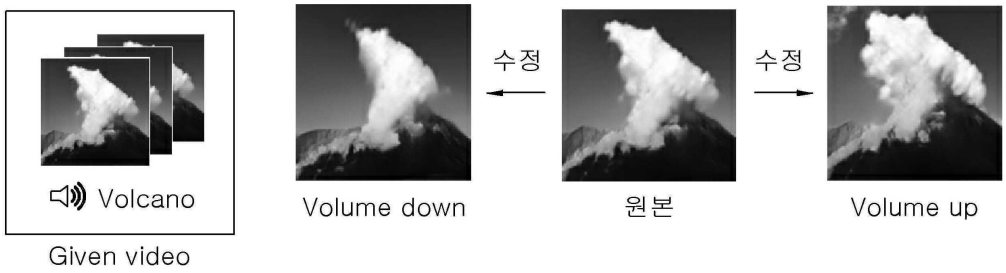
도면9



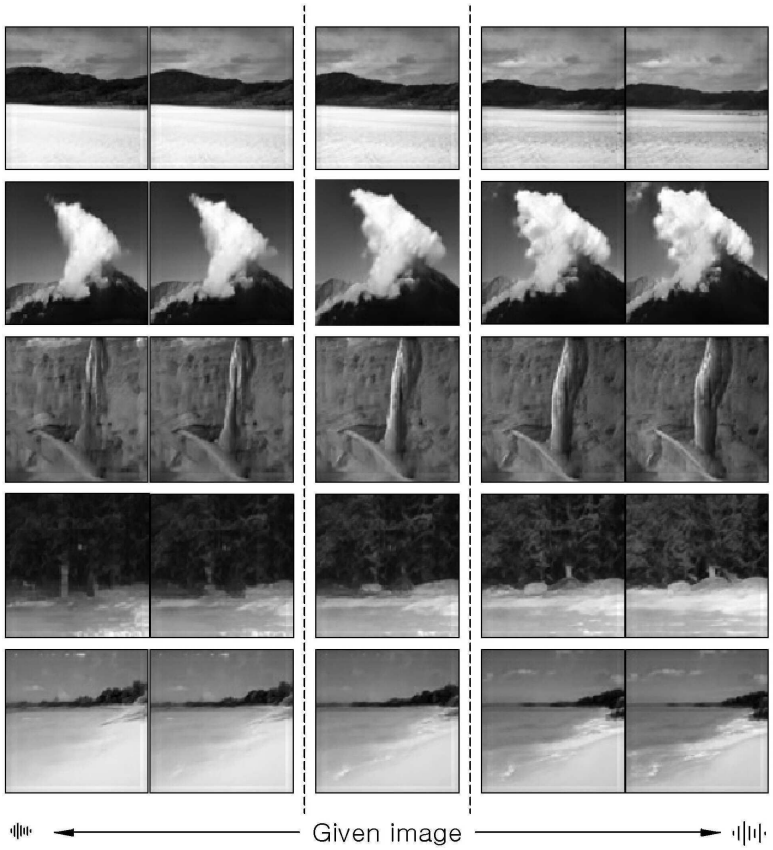
도면10



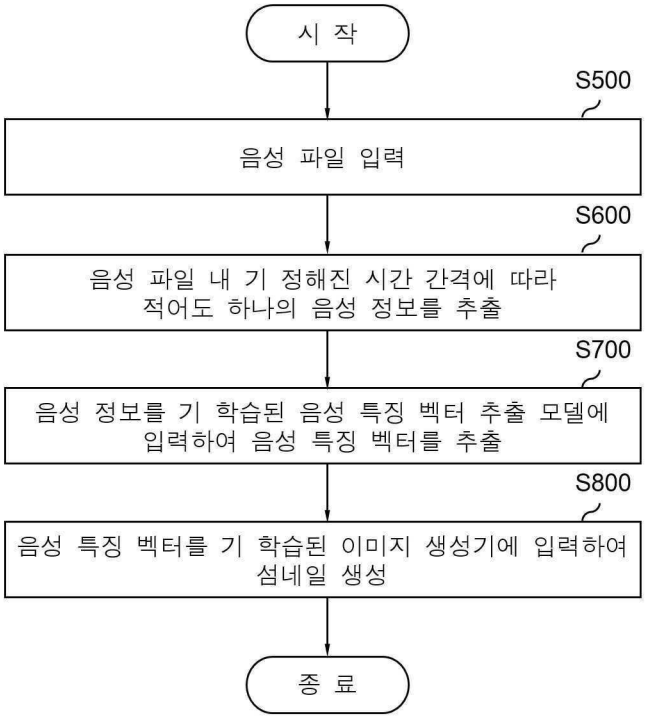
도면11



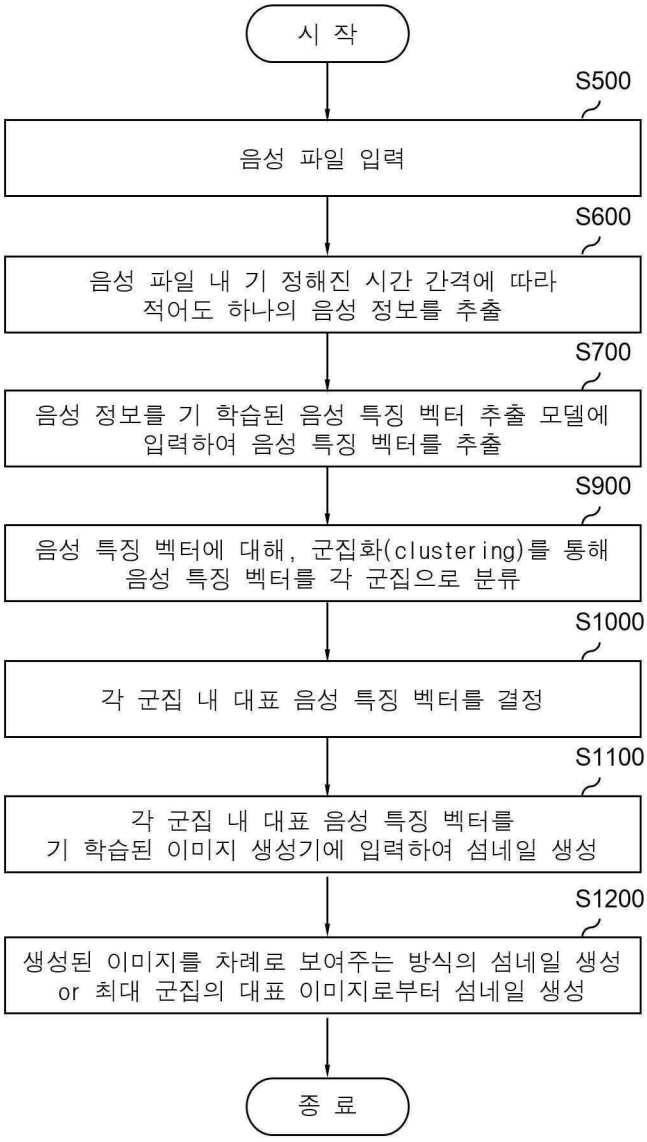
도면12


Given image

도면13


시 작
S500
음성 파일 입력
S600
음성 파일 내 기 정해진 시간 간격에 따라
적어도 하나의 음성 정보를 추출
S700
음성 정보를 기 학습된 음성 특징 벡터 추출 모델에
입력하여 음성 특징 벡터를 추출
S800
음성 특징 벡터를 기 학습된 이미지 생성기에 입력하여
섬네일 생성
종 료

도면14


시 작
S500
음성 파일 입력
S600
음성 파일 내 기 정해진 시간 간격에 따라
적어도 하나의 음성 정보를 추출
S700
음성 정보를 기 학습된 음성 특징 벡터 추출 모델에
입력하여 음성 특징 벡터를 추출
S900
음성 특징 벡터에 대해, 군집화(clustering)를 통해
음성 특징 벡터를 각 군집으로 분류
S1000
각 군집 내 대표 음성 특징 벡터를 결정
S1100
각 군집 내 대표 음성 특징 벡터를
기 학습된 이미지 생성기에 입력하여 섬네일 생성
S1200
생성된 이미지를 차례로 보여주는 방식의 섬네일 생성
or 최대 군집의 대표 이미지로부터 섬네일 생성
종 료

도면15