(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일
2024 년 10 월 10 일 (10.10.2024)

WIPO | PCT

(10) 국제공개번호
WO 2024/210337 A1

(51) 국제특허분류:
G06T 5/00 (2006.01) *G06V 10/62* (2022.01)
G06T 7/11 (2017.01) *G06V 10/52* (2022.01)
G06T 3/40 (2006.01) *G06V 20/40* (2022.01)
G06V 10/44 (2022.01)

(21) 국제출원번호: PCT/KR2024/002801

(22) 국제출원일: 2024 년 3 월 5 일 (05.03.2024)

(25) 출원언어: 한국어

(26) 공개언어: 한국어

(30) 우선권정보:
10-2023-0045040 2023 년 4 월 5 일 (05.04.2023) KR
10-2023-0113187 2023 년 8 월 28 일 (28.08.2023) KR

(71) 출원인: 삼성전자 주식회사 (**SAMSUNG ELECTRONICS CO., LTD.**) [KR/KR]; 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR).

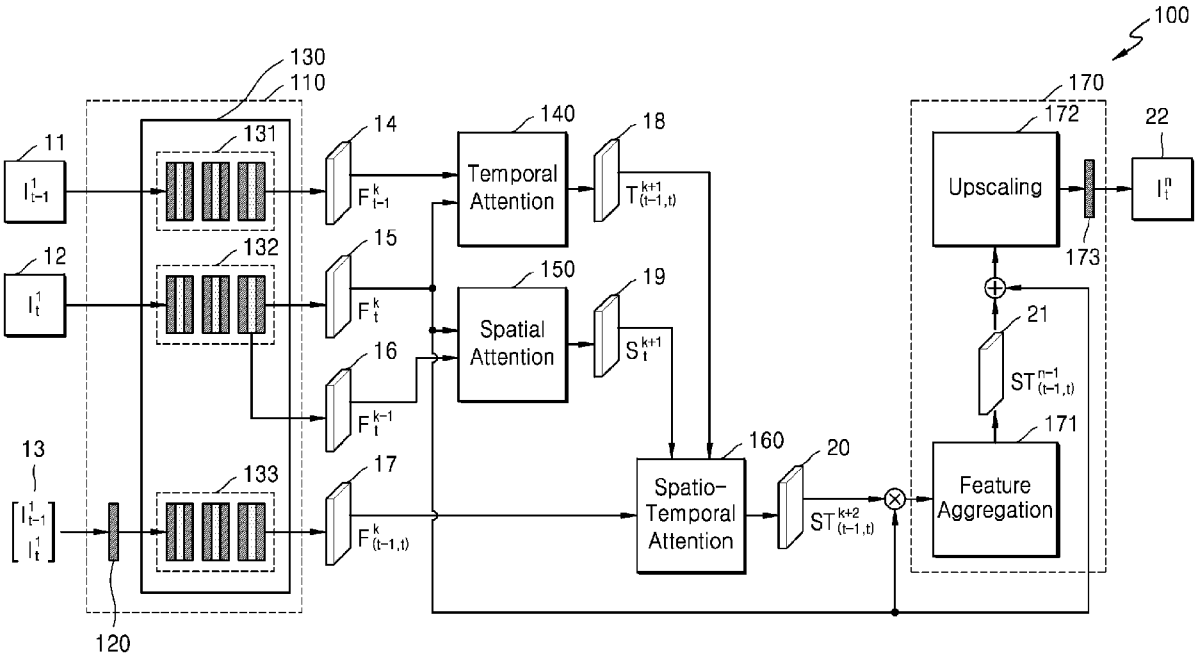
(72) 발명자: 박재연 (**PARK, Jaeyeon**); 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR). 안일준 (**AHN, Iljun**); 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR). 박관우 (**PARK, Kwanwoo**); 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR). 송영찬 (**SONG, Youngchan**); 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR).

(74) 대리인: 리앤목 특허법인 (**Y.P.LEE, MOCK & PARTNERS**); 06292 서울특별시 강남구 언주로30길 13 대림아크로텔 12층, Seoul (KR).

(81) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO,

(54) **Title:** METHOD AND DEVICE FOR PROCESSING VIDEO

(54) **발명의 명칭:** 동영상을 처리하는 방법 및 장치



(57) **Abstract:** Provided is a video processing method comprising the steps of: extracting a first image feature from a first input image included in a scene; extracting a second image feature from a second input image included in the scene, the second input image being a target frame; on the basis of the first image feature and the second image feature, generating a temporal feature associated with temporal change information between the first image feature and the second image feature; and generating an output image on the basis of the temporal feature.

(57) **요약서:** 장면(scene) 내에 포함된 제1 입력 이미지로부터 제1 이미지 특징을 추출하는 단계, 상기 장면 내에 포함된 제2 입력 이미지로부터 제2 이미지 특징을 추출하는 단계로서, 상기 제2 입력 이미지는 타겟 프레임이고, 상기 제1 이미지 특징 및 상기 제2 이미지 특징에 기초하여 상기 제1 이미지 특징과 상기 제2 이미지 특징 사이의 시간적 변화 정보와 연관된 시간적 특징(temporal feature)을 생성하는 단계, 및 상기 시간적 특징에 기초하여 출력 이미지를 생성하는 단계를 포함하는, 동영상 처리 방법이 제공된다.

[다음 쪽 계속]

WO 2024/210337 A1

NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) **지정국** (별도의 표시가 없는 한, 가능한 모든 종류의 역내 권리의 보호를 위하여): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 유라시아 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 유럽 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

공개:

— 국제조사보고서와 함께 (조약 제21조(3))

명세서

발명의 명칭: 동영상을 처리하는 방법 및 장치

기술분야

[1] 본 개시는 동영상을 처리하는 방법 및 이를 수행하는 장치에 관한 것이다.

배경기술

[2] 동영상에 포함된 각 프레임은 동영상 내 객체의 움직임에 의해 서로 다른 정보를 갖는다. 복수의 프레임(예: 시간적 정보)을 이용하여 동영상을 복원하는 경우, 동영상 내 객체의 움직임에 관한 정보(예: 광 흐름(optical flow) 등)가 계산될 수 있고 이미지 와핑(warping) 과정을 통해 프레임들이 정렬(align)될 수 있다.

[3] 예를 들어, 프레임의 정렬을 위해, 인접 프레임으로의 변환을 설명하는 모션 벡터를 결정하기 위한 모션 추정(motion estimation) 또는 인접 프레임을 입력 프레임으로 변환하기 위한 모션 보상(motion compensation) 등의 기술들이 사용될 수 있다.

[4] 그러나 열화된 입력 프레임, 휘도(luminance)의 급격한 변화, 객체의 급격한 동작, 또는 다른 객체에 의해 가려진 객체 등과 같은 다양한 원인으로 인해 광 흐름을 계산하는 데 오류가 발생하기 쉽다. 또한 입력 프레임의 해상도에 따라 광 흐름을 계산하고 이미지 와핑을 수행하는 데 높은 연산량이 필요할 수 있다. 광 흐름 계산 시 오류가 발생하면, 오류를 포함하는 광 흐름을 이용하여 복원된 동영상에서도 잡음이나 아티팩트(artifact)가 발생할 수 있다.

발명의 상세한 설명

과제 해결 수단

[5] 본 개시의 일 측면에 따르면, 동영상 처리 방법은 장면(scene) 내에 포함된 제1 입력 이미지로부터 제1 이미지 특징을 추출하는 단계; 상기 장면 내에 포함된 제2 입력 이미지로부터 제2 이미지 특징을 추출하는 단계로서, 상기 제2 입력 이미지는 타겟 프레임이고; 상기 제1 이미지 특징 및 상기 제2 이미지 특징에 기초하여 상기 제1 이미지 특징과 상기 제2 이미지 특징 사이의 시간적 변화 정보와 연관된 시간적 특징(temporal feature)을 생성하는 단계; 및 상기 시간적 특징에 기초하여 출력 이미지를 생성하는 단계;를 포함할 수 있고, 상기 시간적 특징을 생성하는 단계는, 상기 제1 이미지 특징에 대하여 제1 컨볼루션 연산을 수행하고, 상기 제2 이미지 특징에 대하여 제2 컨볼루션 연산을 수행하는 단계, 상기 제1 이미지 특징 내의 픽셀과 상기 제2 이미지 특징 내의 픽셀 사이의 이동량을 학습하도록 구성된 오프셋 네트워크를 이용하여, 상기 제1 컨볼루션 연산 결과 및 상기 제2 컨볼루션 연산 결과에 기초하여 오프셋 이미지 특징을 생성하는 단계, 상기 오프셋 이미지 특징에 대하여 제3 컨볼루션 연산을 수행하는 단계, 상기 오프셋 이미지 특징에 대하여 제4 컨볼루션 연산을 수행하는 단계, 및 상기 제2 컨볼루션 연산 결과를 제1 쿼리(query)로서 이용하고, 상기 제3 컨볼루션 연산 결과를 제1

키(key)로서 이용하고, 상기 제4 컨볼루션 연산 결과를 제1 벨류(value)로서 이용하는 제1 셀프-어텐션 연산을 수행하여 상기 시간적 특징을 생성하는 단계를 포함할 수 있다.

[6] 본 개시의 일 측면에 따르면, 컴퓨터 판독 가능한 기록 매체는 동영상을 처리하는 장치의 적어도 하나의 프로세서에 의해 실행될 때, 상기 장치가, 장면(scene) 내에 포함된 제1 입력 이미지로부터 제1 이미지 특징을 추출하고, 상기 장면 내에 포함된 제2 입력 이미지로부터 제2 이미지 특징을 추출하되, 상기 제2 입력 이미지는 타겟 프레임이고, 상기 제1 이미지 특징 및 상기 제2 이미지 특징에 기초하여 상기 제1 이미지 특징과 상기 제2 이미지 특징 사이의 시간적 변화 정보와 연관된 시간적 특징(temporal feature)을 생성하고, 상기 시간적 특징에 기초하여 출력 이미지를 생성하는 것을 포함하는 동작을 수행하게 할 수 있는 하나 이상의 인스트럭션을 저장하고, 상기 시간적 특징을 생성하는 것은, 상기 제1 이미지 특징에 대하여 제1 컨볼루션 연산을 수행하고, 상기 제2 이미지 특징에 대하여 제2 컨볼루션 연산을 수행하는 것, 상기 제1 이미지 특징 내의 픽셀과 상기 제2 이미지 특징 내의 픽셀 사이의 이동량을 학습하도록 구성된 오프셋 네트워크를 이용하여, 상기 제1 컨볼루션 연산 결과 및 상기 제2 컨볼루션 연산 결과에 기초하여 오프셋 이미지 특징을 생성하는 것, 상기 오프셋 이미지 특징에 대하여 제3 컨볼루션 연산을 수행하는 것, 상기 오프셋 이미지 특징에 대하여 제4 컨볼루션 연산을 수행하는 것, 및 상기 제2 컨볼루션 연산 결과를 제1 쿼리(query)로서 이용하고, 상기 제3 컨볼루션 연산 결과를 제1 키(key)로서 이용하고, 상기 제4 컨볼루션 연산 결과를 제1 벨류(value)로서 이용하는 제1 셀프-어텐션 연산을 수행하여 상기 시간적 특징을 생성하는 것을 포함할 수 있다.

[7] 본 개시의 일 측면에 따르면, 동영상을 처리하는 장치는, 적어도 하나의 프로세서; 및 하나 이상의 인스트럭션을 저장하도록 구성된 메모리를 포함할 수 있고, 상기 하나 이상의 인스트럭션은, 상기 적어도 하나의 프로세서에 의해 실행될 때, 상기 장치가, 장면(scene) 내에 포함된 제1 입력 이미지로부터 제1 이미지 특징을 추출하고, 상기 장면 내에 포함된 제2 입력 이미지로부터 제2 이미지 특징을 추출하되, 상기 제2 입력 이미지는 타겟 프레임이고, 상기 제1 이미지 특징 및 상기 제2 이미지 특징에 기초하여 상기 제1 이미지 특징과 상기 제2 이미지 특징 사이의 시간적 변화 정보와 연관된 시간적 특징(temporal feature)을 생성하고, 상기 시간적 특징에 기초하여 출력 이미지를 생성하는 것을 포함하는 동작을 수행하게 할 수 있고, 상기 시간적 특징을 생성하는 것은, 상기 제1 이미지 특징에 대하여 제1 컨볼루션 연산을 수행하고, 상기 제2 이미지 특징에 대하여 제2 컨볼루션 연산을 수행하는 것, 상기 제1 이미지 특징 내의 픽셀과 상기 제2 이미지 특징 내의 픽셀 사이의 이동량을 학습하도록 구성된 오프셋 네트워크를 이용하여, 상기 제1 컨볼루션 연산 결과 및 상기 제2 컨볼루션 연산 결과에 기초하여 오프셋 이미지 특징을 생성하는 것, 상기 오프셋 이미지 특징에 대하여 제3 컨볼루션 연산을 수행하는 것, 상기 오프셋 이미지 특징에 대하여 제4 컨볼루션 연산을 수

행하는 것, 및 상기 제2 컨볼루션 연산 결과를 제1 쿼리(query)로서 이용하고, 상기 제3 컨볼루션 연산 결과를 제1 키(key)로서 이용하고, 상기 제4 컨볼루션 연산 결과를 제1 밸류(value)로서 이용하는 제1 셀프-어텐션 연산을 수행하여 상기 시간적 특징을 생성하는 것을 포함할 수 있다.

도면의 간단한 설명

[8] 도 1은 본 개시의 일 실시예에 따른 동영상 처리 시스템의 전체적인 구조를 도시한다.

[9] 도 2는 본 개시의 일 실시예에 따른 시간적 어텐션 모듈 구조를 도시한다.

[10] 도 3은 본 개시의 일 실시예에 따른 오프셋 네트워크의 구조를 도시한다.

[11] 도 4는 본 개시의 일 실시예에 따른 공간적 어텐션 모듈의 구조를 도시한다.

[12] 도 5는 본 개시의 일 실시예에 따른 시공간적 어텐션 모듈의 구조를 도시한다.

[13] 도 6a는 본 개시의 일 실시예에 따른 동영상 처리 방법의 순서도이다.

[14] 도 6b는 본 개시의 일 실시예에 따른 동영상 처리 방법의 순서도이다.

[15] 도 7은 본 개시의 일 실시예에 따른 동영상 처리 장치의 블록도이다.

[16] 도 8은 본 개시의 일 실시예에 따른 동영상 처리 시스템이 텔레비전에 적용된 예시를 도시한다.

발명의 실시를 위한 형태

[17] 본 개시에서, "a, b 및 c 중 적어도 하나"의 표현은 "a", "b", "c", "a 및 b", "a 및 c", "b 및 c", "a, b 및 c 모두", 또는 그 변형들을 지칭할 수 있다.

[18] 본 개시에서 사용되는 용어는 실시예에서의 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 개시에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 개시의 전반에 걸친 내용을 토대로 이해되어야 한다.

[19] 제1, 제2 등의 용어는 다양한 구성요소들을 설명하는 데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 실시예의 권리범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다.

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

[21] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함할 수 있다. 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 용어들은 본 명세서에 기재된 기술분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가질 수 있다.

[22] 본 개시에서, "포함하다" 또는 "가지다" 등의 용어는 본 개시에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[23] 본 개시에서 '~부(유닛)', '모듈' 등으로 표현되는 구성요소는 2개 이상의 구성요소가 하나의 구성요소로 합쳐지거나 또는 하나의 구성요소가 보다 세분화된 기능별로 2개 이상으로 분화될 수도 있다. 또한 이하에서 설명할 구성요소 각각은 설명된 기능 이외에도 다른 구성요소가 담당하는 기능 중 일부 또는 전부의 기능을 추가적으로 수행할 수도 있으며, 구성요소 각각이 담당하는 주기능 중 일부 기능이 다른 구성요소에 의해 전담되어 수행될 수도 있음은 물론이다.

[24] 본 개시에서, 인공지능과 관련된 기능은 프로세서와 메모리를 통해 동작된다. 프로세서는 하나 또는 복수의 프로세서로 구성될 수 있다. 이때, 하나 또는 복수의 프로세서는 CPU, AP, DSP(Digital Signal Processor) 등과 같은 범용 프로세서, GPU, VPU(Vision Processing Unit)와 같은 그래픽 전용 프로세서 또는 NPU와 같은 인공지능 전용 프로세서일 수 있다. 하나 또는 복수의 프로세서는, 메모리에 저장된 기 정의된 동작 규칙 또는 인공지능 모델에 따라, 입력 데이터를 처리하도록 제어한다. 일 실시예에서, 하나 또는 복수의 프로세서가 인공지능 전용 프로세서인 경우, 인공지능 전용 프로세서는, 특정 인공지능 모델의 처리에 특화된 하드웨어 구조로 설계될 수 있다.

[25] 기 정의된 동작 규칙 또는 인공지능 모델은 학습을 통해 만들어진 것을 특징으로 한다. 여기서, 학습을 통해 만들어진다는 것은, 기본 인공지능 모델이 학습 알고리즘에 의하여 다수의 학습 데이터들을 이용하여 학습됨으로써, 원하는 특정 (또는, 목적)을 수행하도록 설정된 기 정의된 동작 규칙 또는 인공지능 모델이 만들어짐을 의미한다. 이러한 학습은 본 개시에 따른 인공지능이 수행되는 기기 자체에서 이루어질 수도 있고, 별도의 서버 및/또는 시스템을 통해 이루어질 수도 있다. 학습 알고리즘의 예로는, 지도형 학습(supervised learning), 비지도형 학습(unsupervised learning), 준지도형 학습(semi-supervised learning) 또는 강화 학습(reinforcement learning)이 있으나, 전술한 예에 한정되지 않는다.

[26] 인공지능 모델은, 복수의 신경망 레이어들로 구성될 수 있다. 복수의 신경망 레이어들 각각은 복수의 가중치들(weight values)을 갖고 있으며, 이전(previous) 레이어의 연산 결과와 복수의 가중치들 간의 연산을 통해 신경망 연산을 수행한다. 복수의 신경망 레이어들이 갖고 있는 복수의 가중치들은 인공지능 모델의 학습 결과에 의해 최적화될 수 있다. 예를 들어, 학습 과정 동안 인공지능 모

델에서 획득한 로스(loss) 값 또는 코스트(cost) 값이 감소 또는 최소화되도록 복수의 가중치들이 갱신될 수 있다. 인공 신경망은 심층 신경망(DNN: Deep Neural Network)을 포함할 수 있으며, 예를 들어, CNN(Convolutional Neural Network), DNN(Deep Neural Network), RNN(Recurrent Neural Network), RBM(Restricted Boltzmann Machine), DBN(Deep Belief Network), BRDNN(Bidirectional Recurrent Deep Neural Network) 또는 DQN(Deep Q-Networks) 등이 있으나, 전술한 예에 한정되지 않는다.

[27] 본 개시에서, 기기로 읽을 수 있는 저장매체는, 비일시적(non-transitory) 저장매체의 형태로 제공될 수 있다. 여기서, '비일시적 저장매체'는 실재(tangible)하는 장치이고, 신호(signal)(예: 전자기파)를 포함하지 않는다는 것을 의미할 뿐이며, 이 용어는 데이터가 저장매체에 반영구적으로 저장되는 경우와 임시적으로 저장되는 경우를 구분하지 않는다. 예로, '비일시적 저장매체'는 데이터가 임시적으로 저장되는 버퍼를 포함할 수 있다.

[28] 일 실시예에 따르면, 본 개시에 설명된 방법은 컴퓨터 프로그램 제품(computer program product)에 포함되어 제공될 수 있다. 컴퓨터 프로그램 제품은 상품으로서 판매자 및 구매자 간에 거래될 수 있다. 컴퓨터 프로그램 제품은 기기로 읽을 수 있는 저장 매체(예: compact disc read only memory (CD-ROM))의 형태로 배포되거나, 또는 어플리케이션 스토어를 통해 또는 두개의 사용자 장치들(예: 스마트폰들) 간에 직접, 온라인으로 배포(예: 다운로드 또는 업로드)될 수 있다. 온라인 배포의 경우에, 컴퓨터 프로그램 제품(예: 다운로더블 앱(downloadable app))의 적어도 일부는 제조사의 서버, 어플리케이션 스토어의 서버, 또는 중계 서버의 메모리와 같은 기기로 읽을 수 있는 저장 매체에 적어도 일시 저장되거나, 임시적으로 생성될 수 있다.

[29] 이하에서는 첨부한 도면을 참고하여 본 개시의 실시예에 대하여 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나 본 개시는 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다.

[30] 도 1은 본 개시의 일 실시예에 따른 동영상 처리 시스템(100)의 전체적인 구조를 도시한다.

[31] 본 개시의 일 실시예에 따른 동영상 처리 시스템(100)은 인공지능 모델을 이용하여 동영상을 복원할 수 있다. 예를 들어, 동영상의 복원은 두 프레임 사이에 새로운 프레임을 생성하여 삽입하는 프레임 보간(frame interpolation), 블러(blur) 등의 잡음을 제거하는 디노이징(denoising), 또는 저해상도(예: 1920x1080)의 동영상을 고해상도(예: 3840x2160)의 동영상으로 변환하는 초해상도(super resolution) 등을 포함할 수 있으나, 이에 제한되는 것은 아니다.

[32] 본 개시의 일 실시예에 따른 동영상 처리 시스템(100)은, I_{t-1}^{1}로 표시될 수 있는 제1 입력 이미지 (11), 및 I_{t}^{1}로 표시될 수 있는 제2 입력 이미지(12)를 입력 받아,

I_t^n로 표시될 수 있는 출력 이미지(22)를 생성할 수 있다. 제2 입력 이미지(12)는 복원의 대상인 타겟 프레임이다. 제1 입력 이미지(11)는 제2 입력 이미지(12)와 동일한 장면(scene) 내에 포함된 인접 프레임이다. 출력 이미지(22)는 제2 입력 이미지(12)로부터 복원된 이미지이다.

[33] 일 예로서, 동영상 처리 시스템(100)은 동영상의 메타 정보에 기초하여 제1 입력 이미지(11)와 제2 입력 이미지(12)가 동일한 장면 내에 포함되었는지 여부를 결정할 수 있다. 동영상의 메타 정보는 각각의 프레임이 속한 장면 정보 및 장면 전환이 일어나는 프레임 정보 중 적어도 하나를 포함할 수 있다. 다른 예로서, 동영상 처리 시스템(100)은 동영상의 프레임의 변화에 기초하여 제1 입력 이미지(11)와 제2 입력 이미지(12)가 동일한 장면 내에 포함되었는지 여부를 결정할 수 있다.

[34] 본 개시에서, 이미지 또는 특징을 나타내는 기호의 아래 첨자는 해당 이미지 또는 특징과 대응되는 프레임의 순서를 나타낸다. 예를 들어, I_{t-1}^1는 동영상의 t-1번째 프레임에 대응되는 이미지를 나타내고, I_t^1는 동영상의 t번째 프레임에 대응되는 이미지 나타낸다. 본 개시에서는 설명의 편의를 위해 t-1번째 프레임과 t번째 프레임을 이용하여 동영상 처리 시스템(100)의 동작을 설명하지만, 이에 제한되지 않으며, 동영상 처리 시스템(100)에 입력되는 이미지가 반드시 연속된 프레임일 필요는 없다. 다시 말하면, 제1 입력 이미지(11)와 제2 입력 이미지(12)가 동일한 장면 내에 포함되었다면, 제1 입력 이미지(11)는 t-2번째 프레임일 수도 있고, t+2번째 프레임일 수도 있다.

[35] 서로 다른 프레임에 해당하는 2개의 입력 이미지를 이용하여 동영상을 복원하기 위해서는 2개의 입력 이미지에 포함된 정보(예: 동영상 내 객체의 종류, 위치, 움직임 등) 간의 관련성이 이용될 수 있다. 관련성이 희박한 입력 이미지의 사용을 자제하는 것은 복원된 동영상의 품질 향상에 도움될 수 있다. 본 개시의 일 실시예에 따른 동영상 처리 시스템(100)은 동일한 장면 내에 포함된 입력 이미지들을 이용하여 동영상을 복원함으로써 시간적 정보를 효과적으로 활용할 수 있다.

[36] 본 개시의 일 실시예에 따른 동영상 처리 시스템(100)은 입력부(110), 시간적 어텐션(temporal attention) 모듈(140), 공간적 어텐션(spatial attention) 모듈(150), 시공간적 어텐션(spatio-temporal attention) 모듈(160) 및 출력부(170) 중 하나 이상을 포함할 수 있다. 다만, 동영상 처리 시스템(100)의 구성요소는 이에 제한되지 않으며, 동영상 처리 시스템(100)은 도 1에 도시된 구성요소 이외에 추가적인 구성요소를 포함할 수도 있고, 도 1에 도시된 구성 중 일부를 포함하지 않을 수도 있다. 일 예로서, 동영상 처리 시스템(100)은 공간적 어텐션 모듈(150) 및 시공간적 어텐션 모듈(160)을 포함하지 않은 형태로 구현될 수 있다. 다른 예로서, 동영상 처리 시스템(100)은 시공간적 어텐션 모듈(160)을 포함하지 않은 형태로 구현될 수 있다.

[37] 일 실시예에서, 입력부(110)는 입력 이미지로부터 이미지 특징을 추출하기 위한 구성으로서, 입력 레이어(120) 및 특징 추출 모듈(130)을 포함할 수 있다.

[38] 일 실시예에서, 특징 추출 모듈(130)은 제1 컨볼루션 신경망(131)을 이용하여, 제1 입력 이미지(11)로부터 $F_{t\text{-}1}^{k}$로 표시될 수 있는 제1 이미지 특징(14)을 추출할 수 있다. 제1 컨볼루션 신경망(131)은 복수의 컨볼루션 레이어 및 복수의 활성화 함수를 포함할 수 있다. 예를 들어, 제1 컨볼루션 신경망(131)의 활성화 함수는 시그모이드(sigmoid), ReLU(Rectified Linear Unit), leaky ReLU, GELU(Gaussian Error Linear Unit), Tanh 등을 포함할 수 있으나, 이에 제한되는 것은 아니다.

[39] 일 실시예에서, 특징 추출 모듈(130)은 제2 컨볼루션 신경망(132)을 이용하여, 제2 입력 이미지(12)로부터 F_{t}^{k}로 표시될 수 있는 제2 이미지 특징(15)을 추출할 수 있다. 제2 컨볼루션 신경망(132)은 복수의 컨볼루션 레이어 및 복수의 활성화 함수를 포함할 수 있다. 예를 들어, 제2 컨볼루션 신경망(132)의 활성화 함수는 시그모이드(sigmoid), ReLU(Rectified Linear Unit), leaky ReLU, GELU(Gaussian Error Linear Unit), Tanh 등을 포함할 수 있으나, 이에 제한되는 것은 아니다.

[40] 일 실시예에서, 특징 추출 모듈(130)은 입력 레이어(120) 및 제3 컨볼루션 신경망(133)을 이용하여, 제3 입력(13)으로부터 $F_{(t\text{-}1,\ t)}^{k}$로 표시될 수 있는 제4 이미지 특징(17)을 추출할 수 있다. 제3 입력(13)은 제1 입력 이미지(11)와 제2 입력 이미지(12)를 채널 방향으로 쌓은 행렬이다. 입력 레이어(120)는 2차원(2D) 컨볼루션 레이어일 수 있다. 제3 컨볼루션 신경망(133)은 복수의 컨볼루션 레이어 및 복수의 활성화 함수를 포함할 수 있다. 예를 들어, 제3 컨볼루션 신경망(133)의 활성화 함수는 시그모이드(sigmoid), ReLU(Rectified Linear Unit), leaky ReLU, GELU(Gaussian Error Linear Unit), Tanh 등을 포함할 수 있으나, 이에 제한되는 것은 아니다.

[41] 제1 컨볼루션 신경망(131), 제2 컨볼루션 신경망(132) 및 제3 컨볼루션 신경망(133)에 포함된 레이어의 개수는 하드웨어 사양, 원하는 화질 수준 등 다양한 요소를 고려하여 설계될 수 있으나, 본 개시에서는 설명의 편의를 위해 각 컨볼루션 신경망에 포함된 레이어의 개수가 k개인 예시가 설명된다.

[42] 이상에서, 특징 추출 모듈(130)이 3개의 컨볼루션 신경망을 포함하고, 각각의 컨볼루션 신경망이 별개의 동작을 수행하는 것으로 설명되었으나, 이에 제한되지 않으며, 특징 추출 모듈(130)의 구성은 다양하게 변형될 수 있다.

[43] 일 예로서, 특징 추출 모듈(130)은 하나의 컨볼루션 신경망을 이용하여, 제1 입력 이미지(11)로부터 제1 이미지 특징(14)을 추출하고, 제2 입력 이미지(12)로부터 제2 이미지 특징(15)을 추출하고, 제3 입력(13)으로부터 제4 이미지 특징(17)을 추출할 수 있다. 예를 들어, 제1 내지 제3 컨볼루션 신경망(131, 132, 133)은 하나의 컨볼루션 신경망에 포함될 수 있다.

[44] 다른 예로서, 특징 추출 모듈(130)은 하나의 컨볼루션 신경망을 이용하여, 제1 입력 이미지(11)로부터 제1 이미지 특징(14)을 추출하고, 제2 입력 이미지(12)로부터 제2 이미지 특징(15)을 추출할 수 있으며, 다른 하나의 컨볼루션 신경망을 이용하여 제3 입력(13)으로부터 제4 이미지 특징(17)을 추출할 수 있다. 예를 들어, 제1 및 제2 컨볼루션 신경망(131, 132)은 하나의 컨볼루션 신경망에 포함될 수 있다.

[45] 본 개시에서, 이미지 또는 특징을 나타내는 기호의 위 첨자는 해당 이미지 또는 특징이 통과한 레이어의 개수 또는 동영상 처리 시스템(100) 내에서의 해당 이미지 또는 특징의 상대적인 위치를 나타낸다. 예를 들어, F_t^k는 k번째 레이어로부터 출력된 특징을 나타내고, F_t^{k-1}는 k-1번째 레이어로부터 출력된 특징을 나타낸다.

[46] 일 실시예에서, 제2 컨볼루션 신경망(132)이 k개의 레이어를 포함하는 경우, k-1번째 레이어의 출력인, F_t^{k-1}로 표시될 수 있는, 제3 이미지 특징 (16)은 공간적 어텐션 모듈(150)의 입력으로 사용될 수 있다. 여기서, k-1번째 레이어의 출력은 제2 컨볼루션 신경망(132) 내의 k-1번째 컨볼루션 레이어의 출력, 또는 k-1번째 컨볼루션 레이어 다음의 활성화 함수의 출력 중 어느 하나를 의미할 수 있다. k-1번째 컨볼루션 레이어 다음에 활성화 함수가 존재하는지 여부는 설계 의도에 기초하여 결정될 수 있다. k-1번째 레이어의 출력은 k번째 컨볼루션 레이어의 입력으로 이해될 수 있다. 동영상 처리 시스템(100)이 제3 이미지 특징(16)을 이용하여 동영상을 복원하는 동작에 관하여는 공간적 어텐션 모듈(150)의 동작과 함께 후술하기로 한다.

[47] 일 실시예에서, 시간적 어텐션 모듈(140)은 제1 이미지 특징(14) 및 제2 이미지 특징(15)에 기초하여 제1 이미지 특징(14)과 제2 이미지 특징(15) 사이의 시간적 변화 정보와 연관된, $T_{(t-1,\,t)}^{k+1}$로 표시될 수 있는, 시간적 특징(temporal feature)(18)을 생성할 수 있다. 시간적 어텐션 모듈(140)은 시간적 특징(18)을 생성하기 위해 수정된 셀프-어텐션(self-attention)에 기초한 연산을 수행할 수 있다. 시간적 어텐션 모듈(140)의 구조 및 동작의 예시는 도 2 및 도 3을 참조하여 후술된다.

[48] 일 실시예에서, 공간적 어텐션 모듈(150)은 제2 이미지 특징(15) 및 제3 이미지 특징(16)에 기초하여 제2 입력 이미지(12)에 대한, S_t^{k+1}로 표시될 수 있는, 공간적 특징(spatial feature)(19)을 생성할 수 있다. 공간적 어텐션 모듈(150)은 공간적 특징(19)을 생성하기 위해 수정된 셀프-어텐션에 기초한 연산을 수행할 수 있다. 공간적 어텐션 모듈(150)의 구조 및 동작의 예시는 도 4를 참조하여 후술된다.

[49] 일 실시예에서, 시간적 어텐션 모듈(140)이 서로 다른 프레임으로부터 추출된 특징들을 이용하여 시간적 특징(18)을 생성할 수 있고, 공간적 어텐션 모듈(150)은 하나의 프레임으로부터 추출된 특징들을 이용하여 공간적 특징(19)을 생성할 수 있다. 이때, 공간적 어텐션 모듈(150)은 제2 컨볼루션 신경망(132)의 마지막 레

이어의 출력 및 그 이전 레이어의 출력을 입력으로 한다. 컨볼루션 신경망에서, 더 깊은 레이어는 더 큰 수용 영역(receptive field)을 제공할 수 있고, 각 레이어에서 출력되는 특징들은 동일한 입력으로부터 생성되었더라도 서로 다른 정보를 가질 수 있다. 따라서 제2 이미지 특징(15)과 제3 이미지 특징(16)은 모두 제2 입력 이미지(12)로부터 추출된 특징들이지만 각자가 다른 정보를 포함할 수 있으며, 공간적 어텐션 모듈(150)은 제2 이미지 특징(15)과 제3 이미지 특징(16)에 기초한 수정된 셀프-어텐션 연산을 통해 공간적 특징(19)을 생성할 수 있다.

[50] 일 실시예에서, 시공간적 어텐션 모듈(160)은 제4 이미지 특징(17), 시간적 특징(18) 및 공간적 특징(19)에 기초하여 제1 입력 이미지(11) 및 제2 입력 이미지(12)에 대한, $ST^{k+2}_{(t-1,\,t)}$ 로 표시될 수 있는, 시공간적 특징(spatio-temporal feature)(20)을 생성할 수 있다. 시공간적 어텐션 모듈(160)은 시공간적 특징(20)을 생성하기 위해 수정된 셀프-어텐션에 기초한 연산을 수행할 수 있다. 시공간적 어텐션 모듈(160)의 구조 및 동작의 예시는 도 5를 참조하여 후술된다.

[51] 일 실시예에서, 출력부(170)는 시공간적 특징(20)으로부터 출력 이미지(22)를 생성하기 위한 구성으로서, 특징 통합(feature aggregation) 모듈(171), 업스케일링(upscaling) 모듈(172) 및 출력 레이어(173)를 포함할 수 있다.

[52] 일 실시예에서, 특징 통합 모듈(171)은 시공간적 특징(20)과 제2 이미지 특징(15)을 행렬 곱한 입력에 대해 컨볼루션 연산을 수행할 수 있다. 특징 통합 모듈(171)은 하나 이상의 컨볼루션 레이어 및 하나 이상의 활성화 함수를 포함할 수 있다. 예를 들어, 특징 통합 모듈(171)의 활성화 함수는 시그모이드(sigmoid), ReLU(Rectified Linear Unit), leaky ReLU, GELU(Gaussian Error Linear Unit), Tanh 등을 포함할 수 있으나, 이에 제한되는 것은 아니다.

[53] 전술한 바와 같이, 일 실시예에서, 동영상 처리 시스템(100)은 공간적 어텐션 모듈(150) 및 시공간적 어텐션 모듈(160)을 포함하지 않은 형태로 구현될 수 있다. 이 경우, 특징 통합 모듈(171)은 시간적 특징(18)과 제2 이미지 특징(15)을 행렬 곱한 입력에 대해 컨볼루션 연산을 수행할 수 있다.

[54] 전술한 바와 같이, 일 실시예에서, 동영상 처리 시스템(100)은 시공간적 어텐션 모듈(160)을 포함하지 않은 형태로 구현될 수 있다. 이 경우, 특징 통합 모듈(171)은 중간 특징과 제2 이미지 특징(15)을 행렬 곱하여 획득된 입력에 대해 컨볼루션 연산을 수행할 수 있다. 중간 특징은 시간적 특징(18)과 공간적 특징(19)을 채널 방향으로 결합한 후 결합 결과를 채널 수 변경을 위한 컨볼루션 레이어를 통해 통과시킴으로써 획득될 수 있다.

[55] 일 실시예에서, 업스케일링 모듈(172)은 특징 통합 모듈(171)에서 출력된, $ST^{n-1}_{(t-1,\,t)}$ 로 표시될 수 있는, 특징(21)과 제2 이미지 특징(15)을 더한 입력에 대해 컨볼루션 연산 및 픽셀 셔플(pixel shuffle) 연산을 수행할 수 있다. 업스케일링 모듈(172)은 하나 이상의 컨볼루션 레이어, 하나 이상의 활성화 함수 및 하나 이상의 픽셀 셔플 레이어를 포함할 수 있다. 예를 들어, 업스케일링 모듈(172)의

활성화 함수는 시그모이드(sigmoid), ReLU(Rectified Linear Unit), leaky ReLU, GELU(Gaussian Error Linear Unit), Tanh 등을 포함할 수 있으나, 이에 제한되는 것은 아니다.

[56] 일 실시예에서, 출력부(170)는 출력 레이어(173)를 이용하여 업스케일링 모듈(172)의 출력에 대해 컨볼루션 연산을 수행하여 출력 이미지(22)를 생성할 수 있다. 출력 레이어(173)는 2차원(2D) 컨볼루션 레이어일 수 있다.

[57] 도 2는 본 개시의 일 실시예에 따른 시간적 어텐션 모듈(140)의 구조를 도시하고, 도 3은 본 개시의 일 실시예에 따른 제1 및 제2 오프셋 네트워크(250, 260)의 구조를 도시한다.

[58] 전술한 바와 같이, 일 실시예에서, 시간적 어텐션 모듈(140)은 제1 이미지 특징(14) 및 제2 이미지 특징(15)에 기초하여 제1 이미지 특징(14)과 제2 이미지 특징(15) 사이의 시간적 변화 정보와 연관된 시간적 특징(18)을 생성할 수 있다. 시간적 어텐션 모듈(140)은 시간적 특징(18)을 생성하기 위해 수정된 셀프-어텐션에 기초한 연산을 수행할 수 있다.

[59] 셀프-어텐션은 하나의 입력 특징으로부터 쿼리(query), 키(key), 밸류(value)라고 불리는 투영된 특징들을 생성하고, 쿼리(Q)와 키(K)를 행렬 곱하고, 행렬 곱 결과에 소프트맥스(softmax) 함수를 적용하여 가중치를 계산하고, 가중치와 밸류(V)를 행렬 곱하여 최종 출력을 계산하는 네트워크 또는 연산 과정으로 이해될 수 있다. 셀프-어텐션에 의해, 입력 특징의 각 요소(예: 픽셀)와 연관성이 많은 요소에 큰 가중치가 적용될 수 있으며, 셀프-어텐션 이후에 배치된 출력 레이어는 입력 특징의 많은 요소들 중 가중치가 더 큰 요소에 집중할 수 있다. 이하에서는, 쿼리(Q)와 키(K)를 행렬 곱한 후 소프트맥스(softmax) 함수를 적용하여 가중치를 계산하고, 가중치와 밸류(V)를 행렬 곱하는 연산을 "셀프-어텐션 연산"이라고 지칭한다. 다만, 이하에서 설명되는 바와 같이, 본 개시의 일 실시예에 따른 동영상 처리 시스템(100)은 시계열적 데이터(예: F_{t-1}^{k} 및 F_{t}^{k})뿐만 아니라 하나의 입력 이미지로부터 추출된 특징들(예: F_{t}^{k-1} 및 F_{t}^{k})을 처리하기 위하여서도 셀프-어텐션 연산을 수행함에 유의해야 한다.

[60] 본 개시의 일 실시예에 따른 동영상 처리 시스템(100)은, 이하에서 설명될 수정된 셀프-어텐션에 기초한 시간적 특징(18)을 생성하여 동영상을 복원함으로써, 오류가 발생하기 쉽고 연산량 부담이 큰 프레임 정렬(alignment) 과정(예: 광 흐름 추정, 이미지 와핑 등) 없이도 시간적 정보를 효과적으로 활용하여 동영상을 복원할 수 있다.

[61] 도 2를 참조하면, 일 실시예에서, 시간적 어텐션 모듈(140)은 컨볼루션 레이어(210)를 이용하여, 제1 이미지 특징(14)에 대하여, Q_1으로 표시될 수 있는, 제1 쿼리(202)를 생성하기 위한 컨볼루션 연산을 포함할 수 있다.

[62] 일 실시예에서, 시간적 어텐션 모듈(140)은 컨볼루션 레이어(240)를 이용하여, 제2 이미지 특징(15)에 대하여, Q_2로 표시될 수 있는, 제2 쿼리(205)를 생성하기 위한 컨볼루션 연산을 포함할 수 있다.

[63] 일 실시예에서, 시간적 어텐션 모듈(140)은 제1 오프셋 네트워크(250)를 이용하여, 제1 쿼리(202)로부터, $Offset_{t-1}$로 표시될 수 있는, 제1 오프셋(206)을 생성하는 연산을 포함할 수 있다.

[64] 일 실시예에서, 시간적 어텐션 모듈(140)은 제2 오프셋 네트워크(260)를 이용하여, 제2 쿼리(240)로부터, $Offset_t$로 표시될 수 있는, 제2 오프셋(208)을 생성하는 연산을 포함할 수 있다.

[65] 일 실시예에서, 시간적 어텐션 모듈(140)은 제1 이미지 특징에, 제2 오프셋(208)에서 제1 오프셋(206)을 감산하여 획득되는 제3 오프셋(Offset(t, t-1))(207)을 더하여, $F_{off\ t-1\to t}^{k}$로 표시될 수 있는, 오프셋이 적용된 이미지 특징(201)을 생성하는 연산을 포함할 수 있다. 일 실시예에서, 이미지 특징(201)은 오프셋 이미지 특징으로 지칭될 수 있다.

[66] 제1 오프셋 네트워크(250) 및 제2 오프셋 네트워크(260)는 제1 이미지 특징(14)과 제2 이미지 특징(15) 사이의 픽셀 간 이동량을 학습하는 뉴럴 네트워크이다. 예를 들어, 동영상 내에서 객체가 이동한 경우, 제1 이미지 특징(14)과 제2 이미지 특징(15)에서 해당 객체에 대응하는 픽셀의 위치(예: 인덱스)가 달라진다. 제3 오프셋(207)은 t-1번째 프레임과 연관된 제1 오프셋(206)과 t번째 프레임과 연관된 제2 오프셋(208)의 차이에 해당하는 것으로서, 두 프레임 간의 픽셀의 위치 차이만큼의 이동량에 관한 정보를 포함하게 된다. 예를 들어, 제3 오프셋(207)은 픽셀의 이동량에 해당하는 너비(width)와 높이(height)를 각각 x축 값과 y축 값으로 나타낼 수 있다. 시간적 어텐션 모듈(140)은 제1 이미지 특징(14)에 제3 오프셋(207)을 더함으로써 두 프레임 간의 픽셀의 이동량에 관한 정보가 반영된 오프셋이 적용된 이미지 특징(201)을 생성할 수 있다.

[67] 일 실시예에서, 제1 오프셋 네트워크(250)와 제2 오프셋 네트워크(260)는 각각 하나 이상의 컨볼루션 레이어와 하나 이상의 활성화 함수를 포함할 수 있다. 예를 들어, 도 3에 도시된 바와 같이, 제1 오프셋 네트워크(250)와 제2 오프셋 네트워크(260)는 각각 깊이 별(depthwise) 컨볼루션 레이어(310, 340), GELU(Gaussian Error Linear Unit) 활성화 함수(320, 350), 및 포인트 별(pointwise) 컨볼루션 레이어(330, 360)를 포함할 수 있다. 다만, 제1 오프셋 네트워크(250)와 제2 오프셋 네트워크(260)의 구조는 이에 제한되지 않으며, 예를 들어, 컨볼루션, 확장된(dilated) 컨볼루션, 전치된(transposed) 컨볼루션, 분리 가능한(separable) 컨볼루션, 깊이 별 분리 가능한(depthwise separable) 컨볼루션, 그룹(grouped) 컨볼루션 등 연산량을 줄일 수 있는 다양한 컨볼루션 기반의 네트워크가 사용될 수 있으며, 시그모이드(sigmoid), ReLU(Rectified Linear Unit), leaky ReLU, Tanh 등의 활성화 함수가 사용될 수 있다.

[68] 이상에서 제공된 예시들이 2개의 오프셋 네트워크를 포함하는 시간적 어텐션 모듈(140)과 연관되고 각각의 오프셋 네트워크가 별개의 동작을 수행하지만, 본 개시는 이에 제한되지 않으며, 시간적 어텐션 모듈(140)의 구성은 다양하게 변형될 수 있다. 예를 들어, 하나의 오프셋 네트워크를 이용하여 제1 및 제2 쿼리(202, 205)로부터 제3 오프셋(207)을 생성하는 연산을 포함할 수 있다. 예를 들어, 제1 및 제2 오프셋 네트워크(250, 260)는 하나의 뉴럴 네트워크에 포함될 수 있다.

[69] 다시 도 2를 참조하면, 일 실시예에서, 시간적 어텐션 모듈(140)은 컨볼루션 레이어(220)를 이용하여, 오프셋이 적용된 이미지 특징(201)에 대하여, V_1으로 표시될 수 있는, 제1 밸류(203)를 생성하기 위한 컨볼루션 연산을 포함할 수 있다.

[70] 일 실시예에서, 시간적 어텐션 모듈(140)은 컨볼루션 레이어(230)를 이용하여, 오프셋이 적용된 이미지 특징(201)에 대하여, K_1으로 표시될 수 있는, 제1 키(204)를 생성하기 위한 컨볼루션 연산을 포함할 수 있다.

[71] 일 실시예에서, 시간적 어텐션 모듈(140)은 제2 쿼리(205)와 제1 키(204)를 행렬 곱한 후 행렬 곱 결과에 소프트맥스(270) 함수를 적용하여 가중치를 계산하고, 가중치와 제1 밸류(203)를 행렬 곱하여 시간적 특징(18)을 생성할 수 있다.

[72] 전술한 바와 같이, 시간적 어텐션 모듈(140)이 수행하는 수정된 셀프-어텐션은 2개의 입력 특징(예: 제1 이미지 특징(14), 제2 이미지 특징(15))을 이용하여 2개의 쿼리(예: 제1 쿼리(202), 제2 쿼리(205))를 생성하고, 각각의 쿼리로부터 오프셋(207)을 계산하며, 오프셋이 적용된 특징(201)을 이용하여 키(204)와 밸류(203)를 생성하는 것을 특징으로 한다. 다시 말하면, 본 개시의 일 실시예에 따른 동영상 처리 시스템(100)은 이미지 레벨이 아닌, 특징 레벨에서 두 프레임에서의 픽셀 간 이동량을 고려함으로써 프레임 정렬(alignment) 과정 없이도 시간적 정보를 효과적으로 활용하여 동영상을 복원할 수 있다.

[73] 도 4는 본 개시의 일 실시예에 따른 공간적 어텐션 모듈(150)의 구조를 도시한다.

[74] 전술한 바와 같이, 공간적 어텐션 모듈(150)은 제2 이미지 특징(15) 및 제3 이미지 특징(16)에 기초하여 공간적 특징(spatial feature)(19)을 생성할 수 있다. 공간적 어텐션 모듈(150)은 공간적 특징(19)을 생성하기 위해 수정된 셀프-어텐션에 기초한 연산을 수행할 수 있다.

[75] 일 실시예에서, 공간적 어텐션 모듈(150)은 컨볼루션 레이어(410)를 이용하여, 제3 이미지 특징(16)에 대하여, V_2로 표시될 수 있는, 제2 밸류(401)를 생성하기 위한 컨볼루션 연산을 포함할 수 있다.

[76] 일 실시예에서, 공간적 어텐션 모듈(150)은 컨볼루션 레이어(420)를 이용하여, 제3 이미지 특징(16)에 대하여, K_2로 표시될 수 있는, 제2 키(402)를 생성하기 위한 컨볼루션 연산을 포함할 수 있다.

[77] 일 실시예에서, 공간적 어텐션 모듈(150)은 컨볼루션 레이어(430)를 이용하여, 제2 이미지 특징(15)에 대하여, Q_3로 표시될 수 있는, 제3 쿼리(403)를 생성하기 위한 컨볼루션 연산을 포함할 수 있다.

[78] 일 실시예에서, 공간적 어텐션 모듈(150)은 제3 쿼리(403)와 제2 키(402)를 행렬 곱한 후 행렬 곱 결과에 소프트맥스(440) 함수를 적용하여 가중치를 계산하고, 가중치와 제2 밸류(401)를 행렬 곱하여 공간적 특징(19)을 생성할 수 있다.

[79] 공간적 어텐션 모듈(150)이 수행하는 수정된 셀프-어텐션은 컨볼루션 신경망의 최종 레이어의 출력(예: 제2 이미지 특징(15))을 이용하여 쿼리(예: 제3 쿼리(403))를 생성하고, 최종 레이어의 이전 레이어의 출력(16)을 이용하여 키(예: 제2 키(402))와 밸류(예: 제2 밸류(401))를 생성하는 것을 특징으로 한다. 이에 따르면, 본 개시의 일 실시예에 따른 동영상 처리 시스템(100)은 시간적 방향뿐만 아니라 공간적 방향으로도 셀프-어텐션을 별도로 적용하여 입력 이미지에 포함된 공간적 정보를 효과적으로 활용할 수 있다.

[80] 도 5는 본 개시의 일 실시예에 따른 시공간적 어텐션 모듈(160)의 구조를 도시한다.

[81] 전술한 바와 같이, 시공간적 어텐션 모듈(160)은 제4 이미지 특징(17), 시간적 특징(18) 및 공간적 특징(19)에 기초하여 시공간적 특징(20)을 생성할 수 있다. 시공간적 어텐션 모듈(160)은 시공간적 특징(20)을 생성하기 위해 수정된 셀프-어텐션에 기초한 연산을 수행할 수 있다.

[82] 일 실시예에서, 시공간적 어텐션 모듈(160)은 컨볼루션 레이어(510)를 이용하여, 시간적 특징(18)에 대하여, V_3로 표시될 수 있는, 제3 밸류(501)를 생성하기 위한 컨볼루션 연산을 포함할 수 있다.

[83] 일 실시예에서, 시공간적 어텐션 모듈(160)은 컨볼루션 레이어(520)를 이용하여, 시간적 특징(18)에 대하여, K_3로 표시될 수 있는, 제3 키(502)를 생성하기 위한 컨볼루션 연산을 포함할 수 있다.

[84] 일 실시예에서, 시공간적 어텐션 모듈(160)은 컨볼루션 레이어(530)를 이용하여, 제4 이미지 특징(17)에 대하여, Q_4로 표시될 수 있는, 제4 쿼리(503)를 생성하기 위한 컨볼루션 연산을 포함할 수 있다.

[85] 일 실시예에서, 시공간적 어텐션 모듈(160)은 컨볼루션 레이어(540)를 이용하여, 공간적 특징(19)에 대하여, K_4로 표시될 수 있는, 제4 키(504)를 생성하기 위한 컨볼루션 연산을 포함할 수 있다.

[86] 일 실시예에서, 시공간적 어텐션 모듈(160)은 컨볼루션 레이어(550)를 이용하여, 공간적 특징(19)에 대하여, V_4로 표시될 수 있는, 제4 밸류(505)를 생성하기 위한 컨볼루션 연산을 포함할 수 있다.

[87] 일 실시예에서, 시공간적 어텐션 모듈(160)은 제4 쿼리(503)와 제3 키(502)를 행렬 곱한 후 행렬 곱 결과에 소프트맥스(560) 함수를 적용하여 시간적 특징(18)에 대한 가중치를 계산하고, 시간적 특징(18)에 대한 가중치와 제3 밸류(501)를 행렬

곱할 수 있다. 또한 시공간적 어텐션 모듈(160)은 제4 쿼리(503)와 제4 키(504)를 행렬 곱한 후 행렬 곱 결과에 소프트맥스(570) 함수를 적용하여 공간적 특징(19)에 대한 가중치를 계산하고, 공간적 특징(19)에 대한 가중치와 제4 밸류(505)를 행렬 곱할 수 있다. 시공간적 어텐션 모듈(160)은 시간적 특징(18)에 대한 가중치와 제3 밸류(501)를 행렬 곱한 결과와 공간적 특징(19)에 대한 가중치와 제4 밸류(505)를 행렬 곱한 결과를 더하여 시공간적 특징(20)을 생성할 수 있다.

[88] 시공간적 어텐션 모듈(160)이 수행하는 수정된 셀프-어텐션은 3개의 입력(예: 제4 이미지 특징(17), 시간적 특징(18), 공간적 특징(19))으로부터 1개의 쿼리(예: 제4 쿼리(503)), 2개의 키(예: 제1 키(502), 제2 키(504)), 2개의 밸류(예: 제1 밸류(501), 제2 밸류(505))를 생성하고, 2번의 셀프-어텐션 연산을 통해 시공간적 특징(20)을 생성하는 것을 특징으로 한다. 이에 따르면, 본 개시의 일 실시예에 따른 동영상 처리 시스템(100)은 시간적 정보와 공간적 정보를 효과적으로 결합하여 동영상을 처리할 수 있다.

[89] 도 6a는 본 개시의 일 실시예에 따른 동영상 처리 방법(600)의 순서도이다. 도 6b는 본 개시의 일 실시예에 따른 동영상 처리 방법(600)의 순서도이다. 동영상 처리 방법(600)은 동영상 처리 시스템(100)을 탑재한 동영상 처리 장치(예: 도 7에서 후술되는 동영상 처리 장치(700))에 의해 수행될 수 있다.

[90] 도 6a를 참조하면, 동작 610에서, 타겟 프레임인 제2 입력 이미지(12)와 동일한 장면(scene) 내에 포함된 제1 입력 이미지(11)로부터 제1 이미지 특징(14)을 추출할 수 있다. 동작 610은 특징 추출 모듈(130)의 동작에 대응될 수 있다. 일 실시예에서, 동작 610은 제1 컨볼루션 신경망(131)을 이용하여 수행될 수 있다.

[91] 동작 620에서, 제2 입력 이미지(12)로부터 제2 이미지 특징(15)을 추출할 수 있다. 동작 620은 특징 추출 모듈(130)의 동작에 대응될 수 있다. 일 실시예에서, 동작 620은 제2 컨볼루션 신경망(132)을 이용하여 수행될 수 있다.

[92] 동작 630에서, 제1 이미지 특징(14) 및 제2 이미지 특징(15)에 기초하여 제1 이미지 특징(14)과 제2 이미지 특징(15) 사이의 시간적 변화 정보와 연관된 시간적 특징(18)을 생성할 수 있다. 동작 630은 시간적 어텐션 모듈(140)의 동작에 대응될 수 있다.

[93] 예를 들어, 도 6b를 참조하면, 동작 631에서, 제1 이미지 특징(14)에 대하여 제1 컨볼루션 연산을 수행하고, 제2 이미지 특징(15)에 대하여 제2 컨볼루션 연산을 수행할 수 있다.

[94] 동작 632에서, 제1 이미지 특징(14)과 제2 이미지 특징(15) 사이의 픽셀 간 이동량을 학습하는 오프셋 네트워크(250, 260)를 이용하여, 제1 컨볼루션 연산 결과 및 제2 컨볼루션 연산 결과에 기초하여 이미지 특징(201)을 생성할 수 있다. 일 실시예에서, 동작 632는 제1 오프셋 네트워크(250)를 이용하여, 제1 컨볼루션 연산 결과로부터 제1 오프셋(206)을 생성하고, 제2 오프셋 네트워크(260)를 이용하여, 제2 컨볼루션 연산 결과로부터 제2 오프셋(208)을 생성하고, 제1 이미지 특징(14)에, 제2 오프셋(208)에서 제1 오프셋(206)을 감산하여 획득되는

제3 오프셋(207)을 더하여 오프셋이 적용된 이미지 특징(201)을 생성할 수 있다. 일 실시예에서, 제1 오프셋 네트워크(250)는 깊이 별(depthwise) 컨볼루션 레이어(310), GELU(Gaussian Error Linear Unit) 활성화 함수(320), 및 포인트 별(pointwise) 컨볼루션 레이어(330)를 포함할 수 있고, 제2 오프셋 네트워크(260)는 깊이 별(depthwise) 컨볼루션 레이어(340), GELU(Gaussian Error Linear Unit) 활성화 함수(350), 및 포인트 별(pointwise) 컨볼루션 레이어(360)를 포함할 수 있다.

[95] 동작 633에서, 오프셋이 적용된 이미지 특징(201)에 대하여 제3 컨볼루션 연산을 수행할 수 있다.

[96] 동작 634에서, 오프셋이 적용된 이미지 특징(201)에 대하여 제4 컨볼루션 연산을 수행할 수 있다.

[97] 동작 635에서, 제2 컨볼루션 연산 결과를 쿼리(query)로서 이용하고, 제3 컨볼루션 연산 결과를 키(key)로서 이용하고, 제4 컨볼루션 연산 결과를 밸류(value)로서 이용하는 셀프-어텐션 연산을 수행하여 시간적 특징(18)을 생성할 수 있다.

[98] 다시 도 6a를 참조하면, 동작 640에서, 시간적 특징(18)에 기초하여 출력 이미지(22)를 생성할 수 있다. 동작 640은 출력부(170)의 동작에 대응될 수 있다.

[99] 일 실시예에서, 동영상 처리 방법(600)은 제2 이미지 특징(15) 및 제3 이미지 특징(16)에 기초하여 제2 입력 이미지(12)에 대한 공간적 특징(19)을 생성하는 동작을 더 포함할 수 있다. 이때, 제3 이미지 특징(16)은 제2 컨볼루션 신경망(132) 내 k-1번째 레이어의 출력이고, k는 제2 컨볼루션 신경망(132)에 포함된 레이어의 개수이다. 일 실시예에서, 공간적 특징(19)을 생성하는 동작은, 제2 이미지 특징(15)에 대하여 제5 컨볼루션 연산을 수행하고, 제3 이미지 특징(16)에 대하여 제6 컨볼루션 연산을 수행하고, 제3 이미지 특징(16)에 대하여 제7 컨볼루션 연산을 수행하고, 제5 컨볼루션 연산 결과를 쿼리로서 이용하고, 제6 컨볼루션 연산 결과를 키로서 이용하고, 제7 컨볼루션 연산 결과를 밸류로서 이용하는 셀프-어텐션 연산을 수행하여 공간적 특징(19)을 생성하는 동작을 포함할 수 있다.

[100] 일 실시예에서, 동작 640은 공간적 특징(19)에 더 기초하여 출력 이미지(22)를 생성할 수 있다.

[101] 일 실시예에서, 동영상 처리 방법(600)은, 제1 입력 이미지(11)와 제2 입력 이미지(12)를 채널 방향으로 쌓은 행렬인 제3 입력(13)으로부터 제4 이미지 특징(17)을 추출하는 동작, 및 제4 이미지 특징(17), 시간적 특징(18) 및 공간적 특징(19)에 기초하여 제1 입력 이미지(11) 및 제2 입력 이미지(12)에 대한 시공간적 특징(20)을 생성하는 동작을 더 포함할 수 있다.

[102] 일 실시예에서, 동작 640은 시공간적 특징(20)에 더 기초하여 출력 이미지(22)를 생성할 수 있다.

[103] 일 실시예에서, 시공간적 특징(20)을 생성하는 동작은, 제4 이미지 특징(17)에 대하여 제8 컨볼루션 연산을 수행하는 동작, 시간적 특징(18)에 대하여 제9 컨볼루션 연산을 수행하는 동작, 시간적 특징(18)에 대하여 제10 컨볼루션 연산을 수행하는 동작, 공간적 특징(19)에 대하여 제11 컨볼루션 연산을 수행하는 동작, 공

간적 특징(19)에 대하여 제12 컨볼루션 연산을 수행하는 동작, 제8 컨볼루션 연산 결과를 쿼리로서 이용하고, 제9 컨볼루션 연산 결과를 키로서 이용하고, 제10 컨볼루션 연산 결과를 밸류로서 이용하는 셀프-어텐션 연산을 수행하여 제1 중간 결과를 생성하고, 제8 컨볼루션 연산 결과를 쿼리로서 이용하고, 제11 컨볼루션 연산 결과를 키로서 이용하고, 제12 컨볼루션 연산 결과를 밸류로서 이용하는 셀프-어텐션 연산을 수행하여 제2 중간 결과를 생성하고, 제1 중간 결과 및 제2 중간 결과를 더하여 시공간적 특징(20)을 생성하는 동작을 포함할 수 있다.

[104] 도 7은 본 개시의 일 실시예에 따른 동영상 처리 장치(700)의 블록도이다.

[105] 도 7에 도시된 동영상 처리 장치(700)는 전술한 동영상 처리 시스템(100)의 동작을 수행하여 동영상을 처리할 수 있다. 처리 대상인 동영상은 동영상 처리 장치(700)에 저장된 동영상, 및 동영상 처리 장치(700)가 외부 장치(예: 인터넷을 통해 동영상을 제공하는 OTT(over the top) 서비스 제공자의 서버 등)로부터 수신한 동영상 중 적어도 하나일 수도 있다.

[106] 일 실시예에서, 동영상 처리 장치(700)는 프로세서(710), 메모리(720) 및 통신 인터페이스(730) 중 하나 이상을 포함할 수 있다. 다만, 동영상 처리 장치(700)의 구성요소는 이에 제한되지 않으며, 도 7에 도시된 것보다 더 많은 구성요소를 포함할 수도 있고, 도 7에 도시된 구성요소들 중 일부를 포함하지 않을 수도 있다. 일 예로서, 동영상 처리 장치(700)는 동영상을 표시하기 위한 디스플레이 또는 사용자 입력을 수신하는 인터페이스 등을 더 포함하는 텔레비전 또는 모바일 장치(예: 스마트폰, 스마트워치, 태블릿PC 등)일 수도 있다. 다른 예로서, 동영상 처리 장치(700)는 PC(예: 데스크톱, 랩톱 등)에 포함되어 PC의 스토리지(예: SSD, HDD 등)에 저장된 동영상 또는 인터넷을 통해 수신되는 동영상을 처리하는 그래픽 카드이거나 그래픽 카드를 포함할 수도 있다. 일 실시예에서, 프로세서(710), 메모리(720) 및 통신 인터페이스(730) 중 일부 또는 전부는 하나의 칩(chip) 형태로 구현될 수도 있으며, 프로세서(710)는 하나 이상의 프로세서를 포함할 수 있다.

[107] 일 실시예에서, 프로세서(710)는 동영상 처리 장치(700)가 동작하도록 일련의 과정을 제어하는 구성으로서, 하나 또는 복수의 프로세서로 구성될 수 있다. 이 때, 하나 또는 복수의 프로세서는 CPU(Central Processing Unit), AP(Application Processor), DSP(Digital Signal Processor) 등과 같은 범용 프로세서, GPU(Graphic Processing Unit), VPU(Vision Processing Unit)와 같은 그래픽 전용 프로세서 또는 NPU(Neural Processing Unit)와 같은 인공지능 전용 프로세서일 수 있다. 예를 들어, 하나 또는 복수의 프로세서가 인공지능 전용 프로세서인 경우, 인공지능 전용 프로세서는, 특정 인공지능 모델의 처리에 특화된 하드웨어 구조로 설계될 수 있다.

[108] 일 실시예에서, 프로세서(710)는 메모리(720)에 데이터를 기록하거나, 메모리(720)에 저장된 데이터를 읽을 수 있으며, 특히 메모리(720)에 저장된 프로그램을 실행함으로써 미리 정의된 동작 규칙 또는 인공지능 모델에 따라 데이터를 처리

할 수 있다. 따라서, 프로세서(710)는 전술한 동영상 처리 시스템(100)의 동작들을 수행할 수 있다.

[109] 일 실시예에서, 메모리(720)는 다양한 프로그램이나 데이터를 저장하기 위한 구성으로서, 롬(ROM), 램(RAM), 하드디스크, CD-ROM 및 DVD 등과 같은 저장 매체 또는 저장 매체들의 조합으로 구성될 수 있다. 메모리(720)는 별도로 존재하지 않고 프로세서(710)에 포함되도록 구성될 수도 있다. 메모리(720)는 휘발성 메모리, 비휘발성 메모리 또는 휘발성 메모리와 비휘발성 메모리의 조합으로 구성될 수도 있다. 메모리(720)에는 전술한 동영상 처리 시스템(100)의 동작들을 수행하기 위한 프로그램이 저장될 수 있다. 메모리(720)는 프로세서(710)의 요청에 따라 저장된 데이터를 프로세서(710)에 제공할 수도 있다.

[110] 일 실시예에서, 통신 인터페이스(730)는 외부의 장치와 유선 또는 무선으로 신호(예: 제어 명령, 데이터 등)를 송수신하기 위한 구성으로서, 다양한 통신 프로토콜을 지원하는 통신 칩셋을 포함하도록 구성될 수 있다. 통신 인터페이스(730)는 외부로부터 신호를 수신하여 프로세서(710)로 출력하거나, 프로세서(710)로부터 출력된 신호를 외부로 전송할 수 있다. 일 예로서, 동영상 처리 장치(700)가 텔레비전 또는 모바일 장치인 경우, 통신 인터페이스(730)는 인터넷을 통해 동영상을 수신할 수 있는 모듈일 수 있다. 다른 예로서, 동영상 처리 장치(700)가 그래픽 카드인 경우, 메인보드를 통해 CPU, RAM, 또는 스토리지와 신호를 주고받을 수 있는 인터페이스(예: ISA(industry standard architecture bus), VESA Local Bus, NuBus, PCI(peripheral component interconnect bus), PCIE(PCI Express) 등)일 수 있다. 일 실시예에 따르면, 통신 인터페이스(730)는 외부 장치로부터 동영상을 수신할 수 있다.

[111] 도 8은 본 개시의 일 실시예에 따른 동영상 처리 시스템(100)이 텔레비전(800)에 적용된 예시를 도시한다. 일 예로서, 텔레비전(800)은 동영상 처리 장치(700)에 대응될 수 있다. 다른 예로서, 텔레비전(800)은 동영상 처리 시스템(100)을 구동하기 위한 전용 하드웨어(예: 칩)을 포함할 수 있으며, 이 경우 전용 하드웨어가 동영상 처리 장치(700)에 대응될 수 있다.

[112] 도 8에서, 텔레비전(800)은 네트워크(810)를 통해 다양한 외부 장치로부터 동영상을 수신할 수 있다. 예를 들어, 외부 장치는 IPTV 방송 제공자가 관리하는 IPTV 송신기(820), OTT 서비스 제공자의 OTT 서버(830), 동영상 스트리밍 서비스 제공자의 스트리밍 서버(840) 등 네트워크(810)를 통해 텔레비전(800)에 동영상을 제공할 수 있는 다양한 형태의 장치를 포함할 수 있다.

[113] 일 실시예에서, 외부 장치는 네트워크 전송 속도, 네트워크 사용 비용, 서버의 유지 비용 등 다양한 원인에 의해 동영상을 압축한 상태로 보관 또는 전송할 수 있다. 일 실시예에서, 외부 장치가 텔레비전(800)으로 고화질의 동영상을 전송하더라도 전송 과정에서의 손실 등에 의해 텔레비전(800)이 저화질의 동영상을 수신할 수도 있다. 다양한 원인에 의해 저화질의 동영상을 수신한 텔레비전(800)은 동영상 처리 시스템(100)을 이용하여 동영상의 화질을 개선한 후 재생할 수 있다.

전술한 바와 같이, 동영상 처리 시스템(100)은 동영상의 시간적 정보와 공간적 정보를 별도로 추출한 후 이들을 효과적으로 결합함으로써 프레임 정렬(alignment) 과정 없이도 동영상의 화질을 실시간으로 개선할 수 있으며, 텔레비전(800) 사용자의 경험을 향상시킬 수 있다.

[114] 본 개시의 일 측면에 따르면, 동영상 처리 방법은 장면(scene) 내에 포함된 제1 입력 이미지로부터 제1 이미지 특징을 추출하는 단계; 상기 장면 내에 포함된 제2 입력 이미지로부터 제2 이미지 특징을 추출하는 단계로서, 상기 제2 입력 이미지는 타겟 프레임이고; 상기 제1 이미지 특징 및 상기 제2 이미지 특징에 기초하여 상기 제1 이미지 특징과 상기 제2 이미지 특징 사이의 시간적 변화 정보와 연관된 시간적 특징(temporal feature)을 생성하는 단계; 및 상기 시간적 특징에 기초하여 출력 이미지를 생성하는 단계;를 포함할 수 있고, 상기 시간적 특징을 생성하는 단계는, 상기 제1 이미지 특징에 대하여 제1 컨볼루션 연산을 수행하고, 상기 제2 이미지 특징에 대하여 제2 컨볼루션 연산을 수행하는 단계, 상기 제1 이미지 특징 내의 픽셀과 상기 제2 이미지 특징 내의 픽셀 사이의 이동량을 학습하도록 구성된 오프셋 네트워크를 이용하여, 상기 제1 컨볼루션 연산 결과 및 상기 제2 컨볼루션 연산 결과에 기초하여 오프셋 이미지 특징을 생성하는 단계, 상기 오프셋 이미지 특징에 대하여 제3 컨볼루션 연산을 수행하는 단계, 상기 오프셋 이미지 특징에 대하여 제4 컨볼루션 연산을 수행하는 단계, 및 상기 제2 컨볼루션 연산 결과를 제1 쿼리(query)로서 이용하고, 상기 제3 컨볼루션 연산 결과를 제1 키(key)로서 이용하고, 상기 제4 컨볼루션 연산 결과를 제1 밸류(value)로서 이용하는 제1 셀프-어텐션 연산을 수행하여 상기 시간적 특징을 생성하는 단계를 포함할 수 있다.

[115] 일 실시예에서, 상기 오프셋 이미지 특징을 생성하는 단계는, 제1 오프셋 네트워크를 이용하여, 상기 제1 컨볼루션 연산 결과로부터 제1 오프셋을 생성하는 단계, 제2 오프셋 네트워크를 이용하여, 상기 제2 컨볼루션 연산 결과로부터 제2 오프셋을 생성하는 단계, 및 상기 제1 이미지 특징에, 상기 제2 오프셋에서 상기 제1 오프셋을 감산하여 획득되는 제3 오프셋을 더하여 상기 오프셋 이미지 특징을 생성하는 단계를 포함할 수 있다.

[116] 일 실시예에서, 상기 오프셋 네트워크는 깊이 별(depthwise) 컨볼루션 레이어, GELU(Gaussian Error Linear Unit) 활성화 함수, 및 포인트 별(pointwise) 컨볼루션 레이어를 포함할 수 있다.

[117] 일 실시예에서, 상기 제1 이미지 특징을 추출하는 단계는, 제1 컨볼루션 레이어를 이용하여 상기 제1 이미지 특징을 추출하는 단계를 포함할 수 있고, 상기 제2 이미지 특징을 추출하는 단계는, 제2 컨볼루션 레이어를 이용하여 상기 제2 이미지 특징을 추출하는 단계를 포함할 수 있다.

[118] 일 실시예에서, 상기 동영상 처리 방법은 상기 제2 이미지 특징 및 제3 이미지 특징에 기초하여 상기 제2 입력 이미지에 대한 공간적 특징(spatial feature)을 생성하는 단계를 더 포함할 수 있고, 상기 출력 이미지는 상기 공간적 특징에 더 기

초하여 생성될 수 있고, 상기 제3 이미지 특징은 상기 제2 컨볼루션 신경망 내 k-1 번째 레이어의 출력이고, k는 상기 제2 컨볼루션 신경망에 포함된 레이어의 개수이다.

[119] 일 실시예에서, 상기 공간적 특징을 생성하는 단계는, 상기 제2 이미지 특징에 대하여 제5 컨볼루션 연산을 수행하는 단계, 상기 제3 이미지 특징에 대하여 제6 컨볼루션 연산을 수행하는 단계, 상기 제3 이미지 특징에 대하여 제7 컨볼루션 연산을 수행하는 단계, 및 상기 제5 컨볼루션 연산 결과를 제2 쿼리로서 이용하고, 상기 제6 컨볼루션 연산 결과를 제2 키로서 이용하고, 상기 제7 컨볼루션 연산 결과를 제2 밸류로서 이용하는 제2 셀프-어텐션 연산을 수행하여 상기 공간적 특징을 생성하는 단계를 포함할 수 있다.

[120] 일 실시예에서, 상기 동영상 처리 방법은, 채널 방향으로 적층된 상기 제1 입력 이미지와 상기 제2 입력 이미지를 포함하는 행렬로부터 제4 이미지 특징을 추출하는 단계, 및 상기 제4 이미지 특징, 상기 시간적 특징 및 상기 공간적 특징에 기초하여 상기 제1 입력 이미지 및 상기 제2 입력 이미지에 대한 시공간적 특징(spatio-temporal feature)을 생성하는 단계를 더 포함할 수 있고, 상기 출력 이미지는 상기 시공간적 특징에 더 기초하여 생성될 수 있다.

[121] 일 실시예에서, 상기 시공간적 특징을 생성하는 단계는, 상기 제4 이미지 특징에 대하여 제8 컨볼루션 연산을 수행하는 단계, 상기 시간적 특징에 대하여 제9 컨볼루션 연산을 수행하는 단계, 상기 시간적 특징에 대하여 제10 컨볼루션 연산을 수행하는 단계, 상기 공간적 특징에 대하여 제11 컨볼루션 연산을 수행하는 단계, 상기 공간적 특징에 대하여 제12 컨볼루션 연산을 수행하는 단계, 상기 제8 컨볼루션 연산 결과를 제3 쿼리로서 이용하고, 상기 제9 컨볼루션 연산 결과를 제3 키로서 이용하고, 상기 제10 컨볼루션 연산 결과를 제3 밸류로서 이용하는 제3 셀프-어텐션 연산을 수행하여 제1 중간 결과를 생성하는 단계, 상기 제8 컨볼루션 연산 결과를 제4 쿼리로서 이용하고, 상기 제11 컨볼루션 연산 결과를 제4 키로서 이용하고, 상기 제12 컨볼루션 연산 결과를 제4 밸류로서 이용하는 제4 셀프-어텐션 연산을 수행하여 제2 중간 결과를 생성하는 단계, 및 상기 제1 중간 결과 및 상기 제2 중간 결과를 더하여 상기 시공간적 특징을 생성하는 단계를 포함할 수 있다.

[122] 일 실시예에서, 상기 동영상 처리 방법은 상기 동영상의 메타 정보 또는 상기 동영상의 프레임의 변화 중 적어도 하나에 기초하여, 상기 제1 입력 이미지와 상기 제2 입력 이미지가 상기 장면 내에 포함되었는지 여부를 결정하는 단계를 더 포함할 수 있다.

[123] 본 개시의 일 측면에 따르면, 컴퓨터 판독 가능한 기록 매체는 동영상을 처리하는 장치의 적어도 하나의 프로세서에 의해 실행될 때, 상기 장치가, 장면(scene) 내에 포함된 제1 입력 이미지로부터 제1 이미지 특징을 추출하고, 상기 장면 내에 포함된 제2 입력 이미지로부터 제2 이미지 특징을 추출하되, 상기 제2 입력 이미지는 타겟 프레임이고, 상기 제1 이미지 특징 및 상기 제2 이미지 특징에 기초

하여 상기 제1 이미지 특징과 상기 제2 이미지 특징 사이의 시간적 변화 정보와 연관된 시간적 특징(temporal feature)을 생성하고, 상기 시간적 특징에 기초하여 출력 이미지를 생성하는 것을 포함하는 동작을 수행하게 할 수 있는 하나 이상의 인스트럭션을 저장하고, 상기 시간적 특징을 생성하는 것은, 상기 제1 이미지 특징에 대하여 제1 컨볼루션 연산을 수행하고, 상기 제2 이미지 특징에 대하여 제2 컨볼루션 연산을 수행하는 것, 상기 제1 이미지 특징 내의 픽셀과 상기 제2 이미지 특징 내의 픽셀 사이의 이동량을 학습하도록 구성된 오프셋 네트워크를 이용하여, 상기 제1 컨볼루션 연산 결과 및 상기 제2 컨볼루션 연산 결과에 기초하여 오프셋 이미지 특징을 생성하는 것, 상기 오프셋 이미지 특징에 대하여 제3 컨볼루션 연산을 수행하는 것, 상기 오프셋 이미지 특징에 대하여 제4 컨볼루션 연산을 수행하는 것, 및 상기 제2 컨볼루션 연산 결과를 제1 쿼리(query)로서 이용하고, 상기 제3 컨볼루션 연산 결과를 제1 키(key)로서 이용하고, 상기 제4 컨볼루션 연산 결과를 제1 벨류(value)로서 이용하는 제1 셀프-어텐션 연산을 수행하여 상기 시간적 특징을 생성하는 것을 포함할 수 있다.

[124] 본 개시의 일 측면에 따르면, 동영상을 처리하는 장치(700)는, 적어도 하나의 프로세서(710); 및 하나 이상의 인스트럭션을 저장하도록 구성된 메모리(720)를 포함할 수 있고, 상기 하나 이상의 인스트럭션은, 상기 적어도 하나의 프로세서(710)에 의해 실행될 때, 상기 장치(700)가, 장면(scene) 내에 포함된 제1 입력 이미지로부터 제1 이미지 특징을 추출하고, 상기 장면 내에 포함된 제2 입력 이미지로부터 제2 이미지 특징을 추출하되, 상기 제2 입력 이미지는 타겟 프레임이고, 상기 제1 이미지 특징 및 상기 제2 이미지 특징에 기초하여 상기 제1 이미지 특징과 상기 제2 이미지 특징 사이의 시간적 변화 정보와 연관된 시간적 특징(temporal feature)을 생성하고, 상기 시간적 특징에 기초하여 출력 이미지를 생성하는 것을 포함하는 동작을 수행하게 할 수 있고, 상기 시간적 특징을 생성하는 것은, 상기 제1 이미지 특징에 대하여 제1 컨볼루션 연산을 수행하고, 상기 제2 이미지 특징에 대하여 제2 컨볼루션 연산을 수행하는 것, 상기 제1 이미지 특징 내의 픽셀과 상기 제2 이미지 특징 내의 픽셀 사이의 이동량을 학습하도록 구성된 오프셋 네트워크를 이용하여, 상기 제1 컨볼루션 연산 결과 및 상기 제2 컨볼루션 연산 결과에 기초하여 오프셋 이미지 특징을 생성하는 것, 상기 오프셋 이미지 특징에 대하여 제3 컨볼루션 연산을 수행하는 것, 상기 오프셋 이미지 특징에 대하여 제4 컨볼루션 연산을 수행하는 것, 및 상기 제2 컨볼루션 연산 결과를 제1 쿼리(query)로서 이용하고, 상기 제3 컨볼루션 연산 결과를 제1 키(key)로서 이용하고, 상기 제4 컨볼루션 연산 결과를 제1 벨류(value)로서 이용하는 제1 셀프-어텐션 연산을 수행하여 상기 시간적 특징을 생성하는 것을 포함할 수 있다.

[125] 일 실시예에서, 상기 오프셋 이미지 특징을 생성하는 것은, 제1 오프셋 네트워크를 이용하여, 상기 제1 컨볼루션 연산 결과로부터 제1 오프셋을 생성하는 것, 제2 오프셋 네트워크를 이용하여, 상기 제2 컨볼루션 연산 결과로부터 제2 오프셋을 생성하는 것, 및 상기 제1 이미지 특징에, 상기 제2 오프셋에서 상기 제1 오

프셋을 감산하여 획득되는 제3 오프셋을 더하여 상기 오프셋 이미지 특징을 생성하는 것을 포함할 수 있다.

[126] 일 실시예에서, 상기 오프셋 네트워크는 깊이 별(depthwise) 컨볼루션 레이어, GELU(Gaussian Error Linear Unit) 활성화 함수, 및 포인트 별(pointwise) 컨볼루션 레이어를 포함할 수 있다.

[127] 일 실시예에서, 상기 제1 이미지 특징을 추출하는 것은, 제1 컨볼루션 레이어를 이용하여 상기 제1 이미지 특징을 추출하는 것을 포함할 수 있고, 상기 제2 이미지 특징을 추출하는 것은, 제2 컨볼루션 레이어를 이용하여 상기 제2 이미지 특징을 추출하는 것을 포함할 수 있다.

[128] 일 실시예에서, 상기 동작은 상기 제2 이미지 특징 및 제3 이미지 특징에 기초하여 상기 제2 입력 이미지에 대한 공간적 특징(spatial feature)을 생성하는 것을 더 포함할 수 있고, 상기 출력 이미지는 상기 공간적 특징에 더 기초하여 생성될 수 있고, 상기 제3 이미지 특징은 상기 제2 컨볼루션 신경망 내 k-1번째 레이어의 출력이고, k는 상기 제2 컨볼루션 신경망에 포함된 레이어의 개수이다.

[129] 일 실시예에서, 상기 공간적 특징을 생성하는 것은, 상기 제2 이미지 특징에 대하여 제5 컨볼루션 연산을 수행하는 것, 상기 제3 이미지 특징에 대하여 제6 컨볼루션 연산을 수행하는 것, 상기 제3 이미지 특징에 대하여 제7 컨볼루션 연산을 수행하는 것, 및 상기 제5 컨볼루션 연산 결과를 제2 쿼리로서 이용하고, 상기 제6 컨볼루션 연산 결과를 제2 키로서 이용하고, 상기 제7 컨볼루션 연산 결과를 제2 밸류로서 이용하는 제2 셀프-어텐션 연산을 수행하여 상기 공간적 특징을 생성하는 것을 포함할 수 있다.

[130] 일 실시예에서, 상기 동작은, 채널 방향으로 적층된 상기 제1 입력 이미지와 상기 제2 입력 이미지를 포함하는 행렬로부터 제4 이미지 특징을 추출하는 것, 및 상기 제4 이미지 특징, 상기 시간적 특징 및 상기 공간적 특징에 기초하여 상기 제1 입력 이미지 및 상기 제2 입력 이미지에 대한 시공간적 특징(spatio-temporal feature)을 생성하는 것을 더 포함할 수 있고, 상기 출력 이미지는 상기 시공간적 특징에 더 기초하여 생성될 수 있다.

[131] 일 실시예에서, 상기 시공간적 특징을 생성하는 것은, 상기 제4 이미지 특징에 대하여 제8 컨볼루션 연산을 수행하는 것, 상기 시간적 특징에 대하여 제9 컨볼루션 연산을 수행하는 것, 상기 시간적 특징에 대하여 제10 컨볼루션 연산을 수행하는 것, 상기 공간적 특징에 대하여 제11 컨볼루션 연산을 수행하는 것, 상기 공간적 특징에 대하여 제12 컨볼루션 연산을 수행하는 것, 상기 제8 컨볼루션 연산 결과를 제3 쿼리로서 이용하고, 상기 제9 컨볼루션 연산 결과를 제3 키로서 이용하고, 상기 제10 컨볼루션 연산 결과를 제3 밸류로서 이용하는 제3 셀프-어텐션 연산을 수행하여 제1 중간 결과를 생성하는 것, 상기 제8 컨볼루션 연산 결과를 제4 쿼리로서 이용하고, 상기 제11 컨볼루션 연산 결과를 제4 키로서 이용하고, 상기 제12 컨볼루션 연산 결과를 제4 밸류로서 이용하는 제4 셀프-어텐션 연

산을 수행하여 제2 중간 결과를 생성하는 것, 및 상기 제1 중간 결과 및 상기 제2 중간 결과를 더하여 상기 시공간적 특징을 생성하는 것을 포함할 수 있다.

[132] 일 실시예에서, 상기 동작은, 상기 동영상의 메타 정보 또는 상기 동영상의 프레임의 변화 중 적어도 하나에 기초하여 상기 제1 입력 이미지와 상기 제2 입력 이미지가 상기 장면 내에 포함되었는지 여부를 결정하는 것을 더 포함할 수 있다.

청구범위

[청구항 1] 장면(scene) 내에 포함된 제1 입력 이미지로부터 제1 이미지 특징을 추출하는 단계;
상기 장면 내에 포함된 제2 입력 이미지로부터 제2 이미지 특징을 추출하는 단계로서, 상기 제2 입력 이미지는 타겟 프레임이고;
상기 제1 이미지 특징 및 상기 제2 이미지 특징에 기초하여 상기 제1 이미지 특징과 상기 제2 이미지 특징 사이의 시간적 변화 정보와 연관된 시간적 특징(temporal feature)을 생성하는 단계; 및
상기 시간적 특징에 기초하여 출력 이미지를 생성하는 단계;
를 포함하고,
상기 시간적 특징을 생성하는 단계는,
상기 제1 이미지 특징에 대하여 제1 컨볼루션 연산을 수행하고, 상기 제2 이미지 특징에 대하여 제2 컨볼루션 연산을 수행하는 단계,
상기 제1 이미지 특징 내의 픽셀과 상기 제2 이미지 특징 내의 픽셀 사이의 이동량을 학습하도록 구성된 오프셋 네트워크를 이용하여, 상기 제1 컨볼루션 연산 결과 및 상기 제2 컨볼루션 연산 결과에 기초하여 오프셋 이미지 특징을 생성하는 단계,
상기 오프셋 이미지 특징에 대하여 제3 컨볼루션 연산을 수행하는 단계,
상기 오프셋 이미지 특징에 대하여 제4 컨볼루션 연산을 수행하는 단계, 및
상기 제2 컨볼루션 연산 결과를 제1 쿼리(query)로서 이용하고, 상기 제3 컨볼루션 연산 결과를 제1 키(key)로서 이용하고, 상기 제4 컨볼루션 연산 결과를 제1 벨류(value)로서 이용하는 제1 셀프-어텐션 연산을 수행하여 상기 시간적 특징을 생성하는 단계
를 포함하는, 동영상 처리 방법.

[청구항 2] 제1항에 있어서,
상기 오프셋 이미지 특징을 생성하는 단계는,
제1 오프셋 네트워크를 이용하여, 상기 제1 컨볼루션 연산 결과로부터 제1 오프셋을 생성하는 단계,
제2 오프셋 네트워크를 이용하여, 상기 제2 컨볼루션 연산 결과로부터 제2 오프셋을 생성하는 단계, 및
상기 제1 이미지 특징에, 상기 제2 오프셋에서 상기 제1 오프셋을 감산하여 획득되는 제3 오프셋을 더하여 상기 오프셋 이미지 특징을 생성하는 단계를 포함하는, 동영상 처리 방법.

[청구항 3] 제1항 또는 제2항에 있어서,
상기 오프셋 네트워크는
깊이 별(depthwise) 컨볼루션 레이어,

GELU(Gaussian Error Linear Unit) 활성화 함수, 및
포인트 별(pointwise) 컨볼루션 레이어를 포함하는, 동영상 처리 방법.

[청구항 4] 제1항 내지 제3항 중 어느 한 항에 있어서,
상기 제1 이미지 특징을 추출하는 단계는, 제1 컨볼루션 레이어를 이용하여 상기 제1 이미지 특징을 추출하는 단계를 포함하고,
상기 제2 이미지 특징을 추출하는 단계는, 제2 컨볼루션 레이어를 이용하여 상기 제2 이미지 특징을 추출하는 단계를 포함하는, 동영상 처리 방법.

[청구항 5] 제1항 내지 제4항 중 어느 한 항에 있어서,
상기 동영상 처리 방법은 상기 제2 이미지 특징 및 제3 이미지 특징에 기초하여 상기 제2 입력 이미지에 대한 공간적 특징(spatial feature)을 생성하는 단계를 더 포함하고,
상기 출력 이미지는 상기 공간적 특징에 더 기초하여 생성되고,
상기 제3 이미지 특징은 상기 제2 컨볼루션 신경망 내 k-1번째 레이어의 출력이고, k는 상기 제2 컨볼루션 신경망에 포함된 레이어의 개수인, 동영상 처리 방법.

[청구항 6] 제1항 내지 제5항 중 어느 한 항에 있어서,
상기 공간적 특징을 생성하는 단계는,
상기 제2 이미지 특징에 대하여 제5 컨볼루션 연산을 수행하는 단계,
상기 제3 이미지 특징에 대하여 제6 컨볼루션 연산을 수행하는 단계,
상기 제3 이미지 특징에 대하여 제7 컨볼루션 연산을 수행하는 단계, 및
상기 제5 컨볼루션 연산 결과를 제2 쿼리로서 이용하고, 상기 제6 컨볼루션 연산 결과를 제2 키로서 이용하고, 상기 제7 컨볼루션 연산 결과를 제2 밸류로서 이용하는 제2 셀프-어텐션 연산을 수행하여 상기 공간적 특징을 생성하는 단계를 포함하는, 동영상 처리 방법.

[청구항 7] 제1항 내지 제6항 중 어느 한 항에 있어서,
상기 동영상 처리 방법은,
채널 방향으로 적층된 상기 제1 입력 이미지와 상기 제2 입력 이미지를 포함하는 행렬로부터 제4 이미지 특징을 추출하는 단계, 및
상기 제4 이미지 특징, 상기 시간적 특징 및 상기 공간적 특징에 기초하여 상기 제1 입력 이미지 및 상기 제2 입력 이미지에 대한 시공간적 특징(spatio-temporal feature)을 생성하는 단계를 더 포함하고,
상기 출력 이미지는 상기 시공간적 특징에 더 기초하여 생성되는, 동영상 처리 방법.

[청구항 8] 제1항 내지 제7항 중 어느 한 항에 있어서,
상기 시공간적 특징을 생성하는 단계는,
상기 제4 이미지 특징에 대하여 제8 컨볼루션 연산을 수행하는 단계,
상기 시간적 특징에 대하여 제9 컨볼루션 연산을 수행하는 단계,
상기 시간적 특징에 대하여 제10 컨볼루션 연산을 수행하는 단계,

상기 공간적 특징에 대하여 제11 컨볼루션 연산을 수행하는 단계,
상기 공간적 특징에 대하여 제12 컨볼루션 연산을 수행하는 단계,
상기 제8 컨볼루션 연산 결과를 제3 쿼리로서 이용하고, 상기 제9 컨볼루션 연산 결과를 제3 키로서 이용하고, 상기 제10 컨볼루션 연산 결과를 제3 밸류로서 이용하는 제3 셀프-어텐션 연산을 수행하여 제1 중간 결과를 생성하는 단계
상기 제8 컨볼루션 연산 결과를 제4 쿼리로서 이용하고, 상기 제11 컨볼루션 연산 결과를 제4 키로서 이용하고, 상기 제12 컨볼루션 연산 결과를 제4 밸류로서 이용하는 제4 셀프-어텐션 연산을 수행하여 제2 중간 결과를 생성하는 단계, 및
상기 제1 중간 결과 및 상기 제2 중간 결과를 더하여 상기 시공간적 특징을 생성하는 단계를 포함하는, 동영상 처리 방법.

[청구항 9] 제1항 내지 제8항 중 어느 한 항에 있어서,
상기 동영상의 메타 정보 또는 상기 동영상의 프레임의 변화 중 적어도 하나에 기초하여, 상기 제1 입력 이미지와 상기 제2 입력 이미지가 상기 장면 내에 포함되었는지 여부를 결정하는 단계를 더 포함하는, 동영상 처리 방법.

[청구항 10] 제1항 내지 제9항 중 어느 한 항의 방법을 수행하기 위한 컴퓨터 프로그램을 저장한, 컴퓨터 판독 가능한 기록 매체.

[청구항 11] 동영상을 처리하는 장치(700)에 있어서,
적어도 하나의 프로세서(710); 및
하나 이상의 인스트럭션을 저장하도록 구성된 메모리(720)를 포함하고,
상기 하나 이상의 인스트럭션은, 상기 적어도 하나의 프로세서(710)에 의해 실행될 때, 상기 장치(700)가,
장면(scene) 내에 포함된 제1 입력 이미지로부터 제1 이미지 특징을 추출하고,
상기 장면 내에 포함된 제2 입력 이미지로부터 제2 이미지 특징을 추출하되, 상기 제2 입력 이미지는 타겟 프레임이고,
상기 제1 이미지 특징 및 상기 제2 이미지 특징에 기초하여 상기 제1 이미지 특징과 상기 제2 이미지 특징 사이의 시간적 변화 정보와 연관된 시간적 특징(temporal feature)을 생성하고,
상기 시간적 특징에 기초하여 출력 이미지를 생성하는 것을 포함하는 동작을 수행하게 하되,
상기 시간적 특징을 생성하는 것은,
상기 제1 이미지 특징에 대하여 제1 컨볼루션 연산을 수행하고, 상기 제2 이미지 특징에 대하여 제2 컨볼루션 연산을 수행하는 것,
상기 제1 이미지 특징 내의 픽셀과 상기 제2 이미지 특징 내의 픽셀 사이의 이동량을 학습하도록 구성된 오프셋 네트워크를 이용하여, 상기 제1

컨볼루션 연산 결과 및 상기 제2 컨볼루션 연산 결과에 기초하여 오프셋 이미지 특징을 생성하는 것,
상기 오프셋 이미지 특징에 대하여 제3 컨볼루션 연산을 수행하는 것,
상기 오프셋 이미지 특징에 대하여 제4 컨볼루션 연산을 수행하는 것, 및
상기 제2 컨볼루션 연산 결과를 제1 쿼리(query)로서 이용하고, 상기 제3 컨볼루션 연산 결과를 제1 키(key)로서 이용하고, 상기 제4 컨볼루션 연산 결과를 제1 밸류(value)로서 이용하는 제1 셀프-어텐션 연산을 수행하여 상기 시간적 특징을 생성하는 것
을 포함하는, 장치.

[청구항 12] 제11항에 있어서,
상기 오프셋 이미지 특징을 생성하는 것은,
제1 오프셋 네트워크를 이용하여, 상기 제1 컨볼루션 연산 결과로부터 제1 오프셋을 생성하는 것,
제2 오프셋 네트워크를 이용하여, 상기 제2 컨볼루션 연산 결과로부터 제2 오프셋을 생성하는 것, 및
상기 제1 이미지 특징에, 상기 제2 오프셋에서 상기 제1 오프셋을 감산하여 획득되는 제3 오프셋을 더하여 상기 오프셋 이미지 특징을 생성하는 것을 포함하는, 장치.

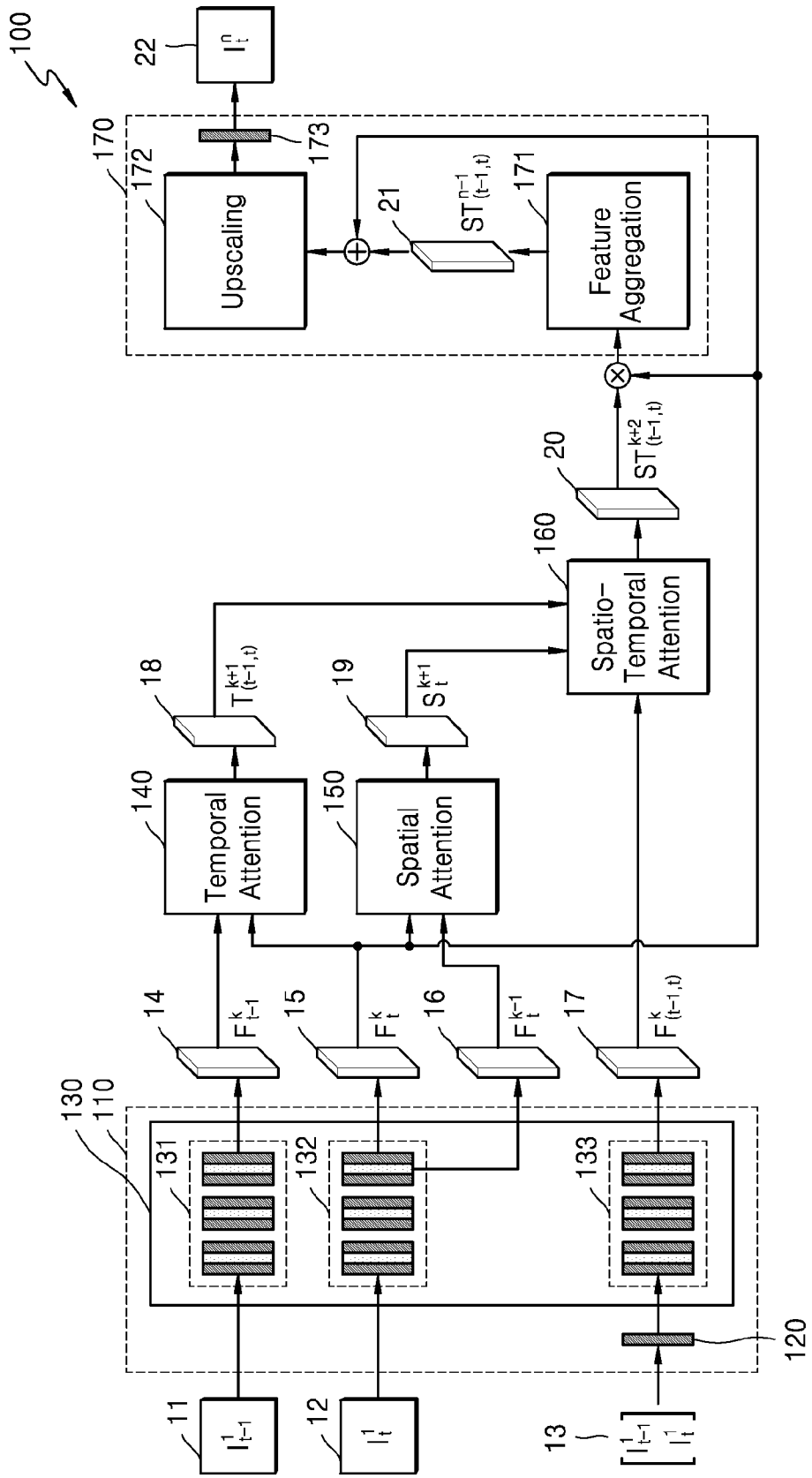
[청구항 13] 제11항 내지 제12항 중 어느 한 항에 있어서,
상기 제1 이미지 특징을 추출하는 것은, 제1 컨볼루션 레이어를 이용하여 상기 제1 이미지 특징을 추출하는 것을 포함하고,
상기 제2 이미지 특징을 추출하는 것은, 제2 컨볼루션 레이어를 이용하여 상기 제2 이미지 특징을 추출하는 것을 포함하는, 장치.

[청구항 14] 제11항 내지 제13항 중 어느 한 항에 있어서,
상기 동작은 상기 제2 이미지 특징 및 제3 이미지 특징에 기초하여 상기 제2 입력 이미지에 대한 공간적 특징(spatial feature)을 생성하는 것을 더 포함하고,
상기 출력 이미지는 상기 공간적 특징에 더 기초하여 생성되고,
상기 제3 이미지 특징은 상기 제2 컨볼루션 신경망 내 k-1번째 레이어의 출력이고, k는 상기 제2 컨볼루션 신경망에 포함된 레이어의 개수인, 장치.

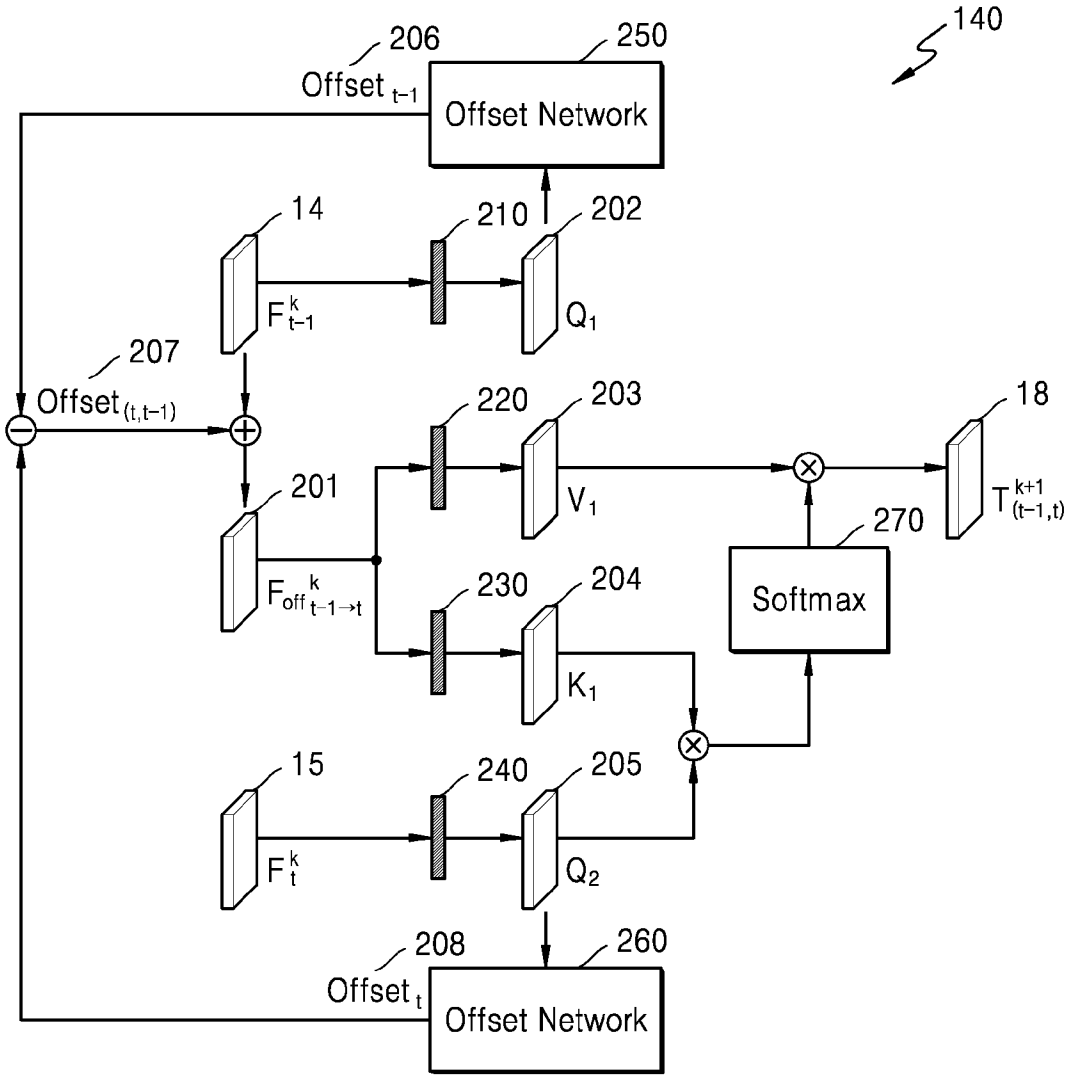
[청구항 15] 제11항 내지 제14항 중 어느 한 항에 있어서,
상기 동작은,
채널 방향으로 적층된 상기 제1 입력 이미지와 상기 제2 입력 이미지를 포함하는 행렬로부터 제4 이미지 특징을 추출하는 것, 및
상기 제4 이미지 특징, 상기 시간적 특징 및 상기 공간적 특징에 기초하여 상기 제1 입력 이미지 및 상기 제2 입력 이미지에 대한 시공간적 특징(spatio-temporal feature)을 생성하는 것을 더 포함하고,

상기 출력 이미지는 상기 시공간적 특징에 더 기초하여 생성되는, 장치.

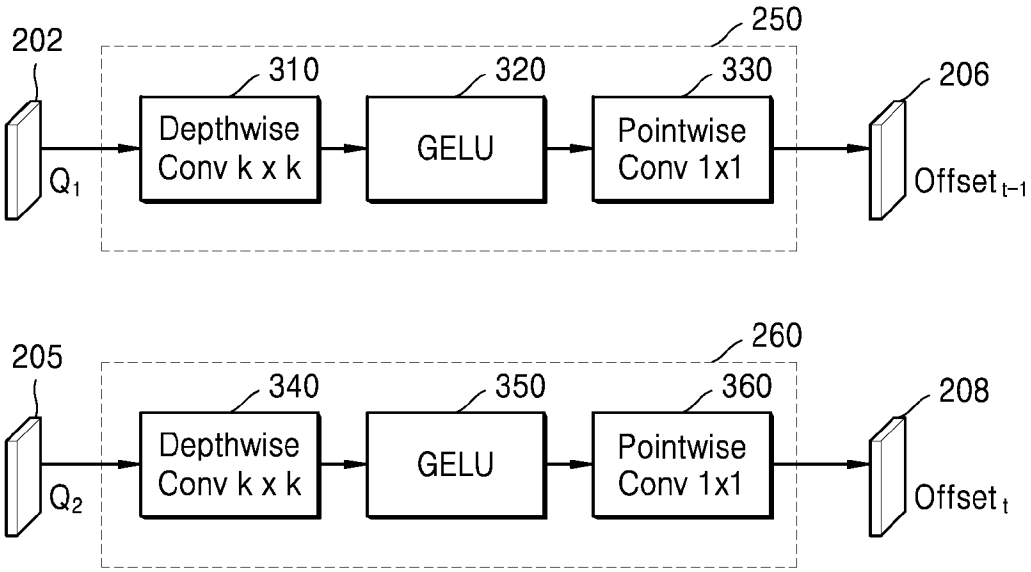
[도1]


100
22
I_t^n
170
172
Upscaling
173
21
$ST_{(t-1,t)}^{n-1}$
171
Feature Aggregation
20
$ST_{(t-1,t)}^{k+2}$
160
Spatio-Temporal Attention
18
$T_{(t-1,t)}^{k+1}$
19
S_t^{k+1}
140
Temporal Attention
150
Spatial Attention
14
F_{t-1}^k
15
F_t^k
16
F_t^{k-1}
17
$F_{(t-1,t)}^k$
130
110
131
132
133
120
11
I_{t-1}^1
12
I_t^1
13
$[I_{t-1}^1 \; I_t^1]$

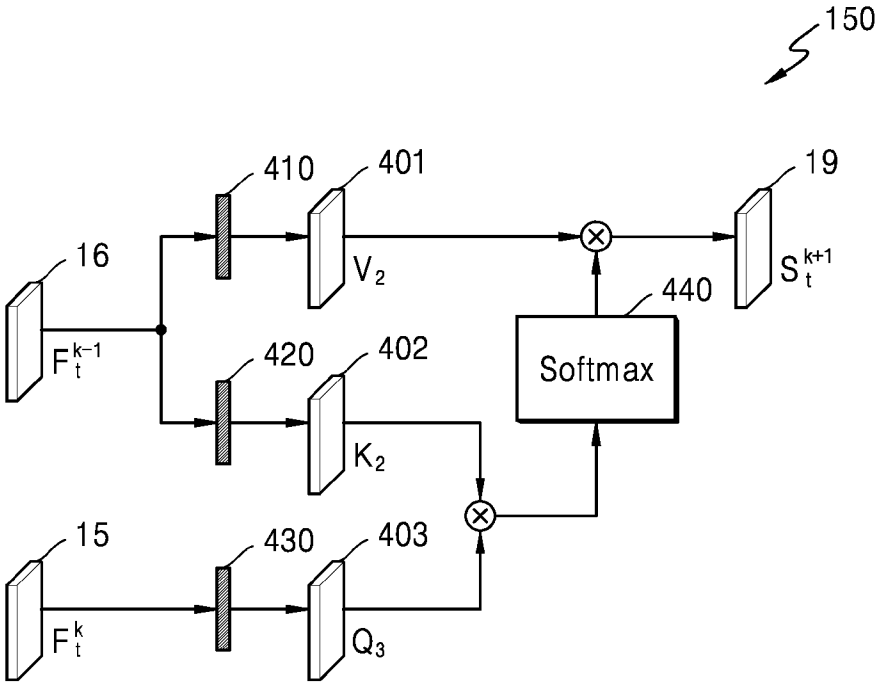
[도2]


140
206
Offset t-1
250
Offset Network
14
210
202
F k t-1
Q1
207
Offset (t,t-1)
220
203
18
V1
T k+1 (t-1,t)
201
270
Softmax
F off k t-1→t
230
204
K1
15
240
205
F k t
Q2
208
260
Offset t
Offset Network

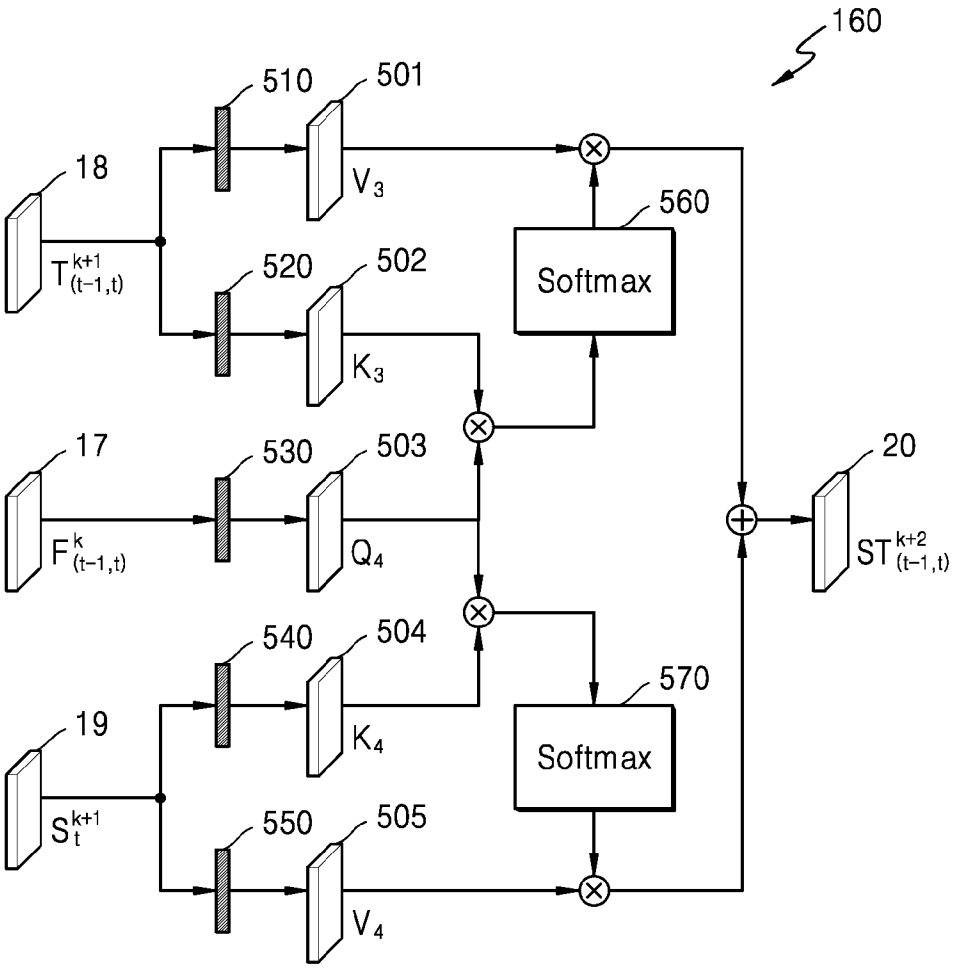
[도3]


202
250
310
320
330
206
Depthwise Conv k x k
GELU
Pointwise Conv 1x1
Q1
Offset t-1
205
260
340
350
360
208
Depthwise Conv k x k
GELU
Pointwise Conv 1x1
Q2
Offset t

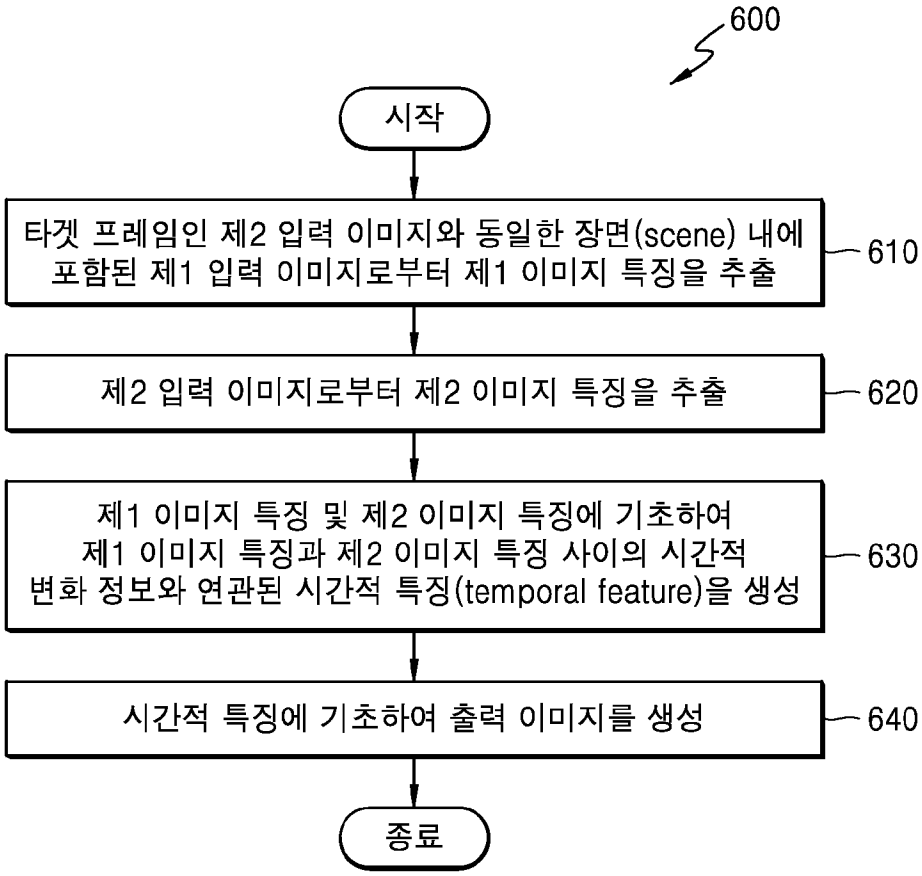
[도4]


150
410
401
19
16
V2
St k+1
440
Softmax
420
402
Ft k-1
K2
15
430
403
Ft k
Q3

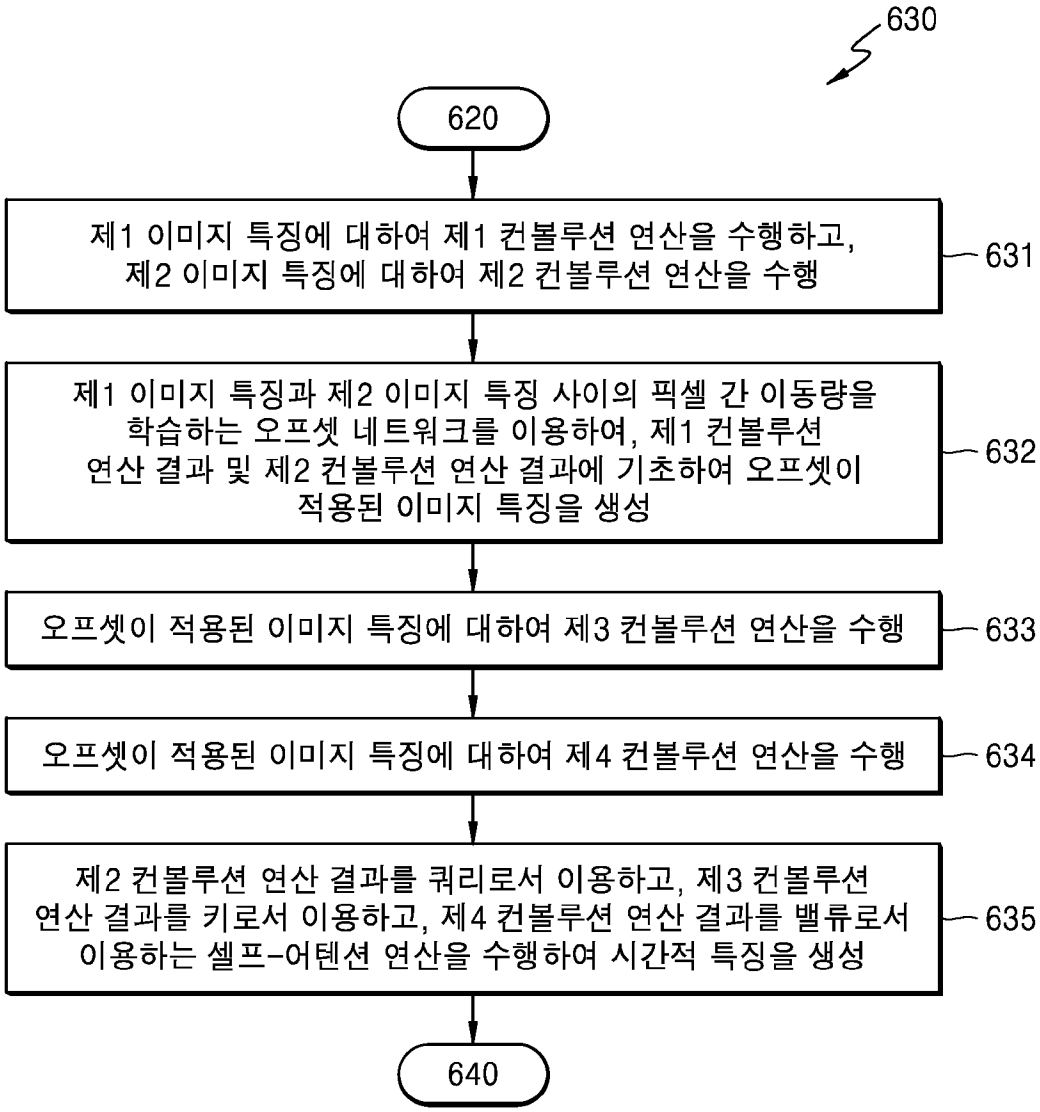
[도5]


160
510
501
18
V3
560
Softmax
520
502
Tk+1 (t-1,t)
K3
17
530
503
20
Fk (t-1,t)
Q4
STk+2 (t-1,t)
540
504
570
19
K4
Softmax
St k+1
550
505
V4

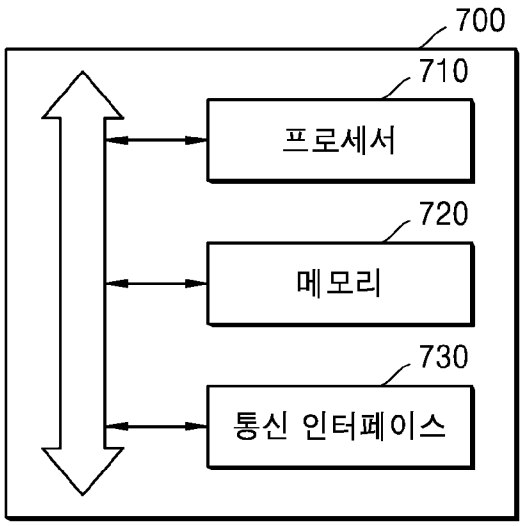
[도6a]


600
시작
타겟 프레임인 제2 입력 이미지와 동일한 장면(scene) 내에 포함된 제1 입력 이미지로부터 제1 이미지 특징을 추출
610
제2 입력 이미지로부터 제2 이미지 특징을 추출
620
제1 이미지 특징 및 제2 이미지 특징에 기초하여 제1 이미지 특징과 제2 이미지 특징 사이의 시간적 변화 정보와 연관된 시간적 특징(temporal feature)을 생성
630
시간적 특징에 기초하여 출력 이미지를 생성
640
종료

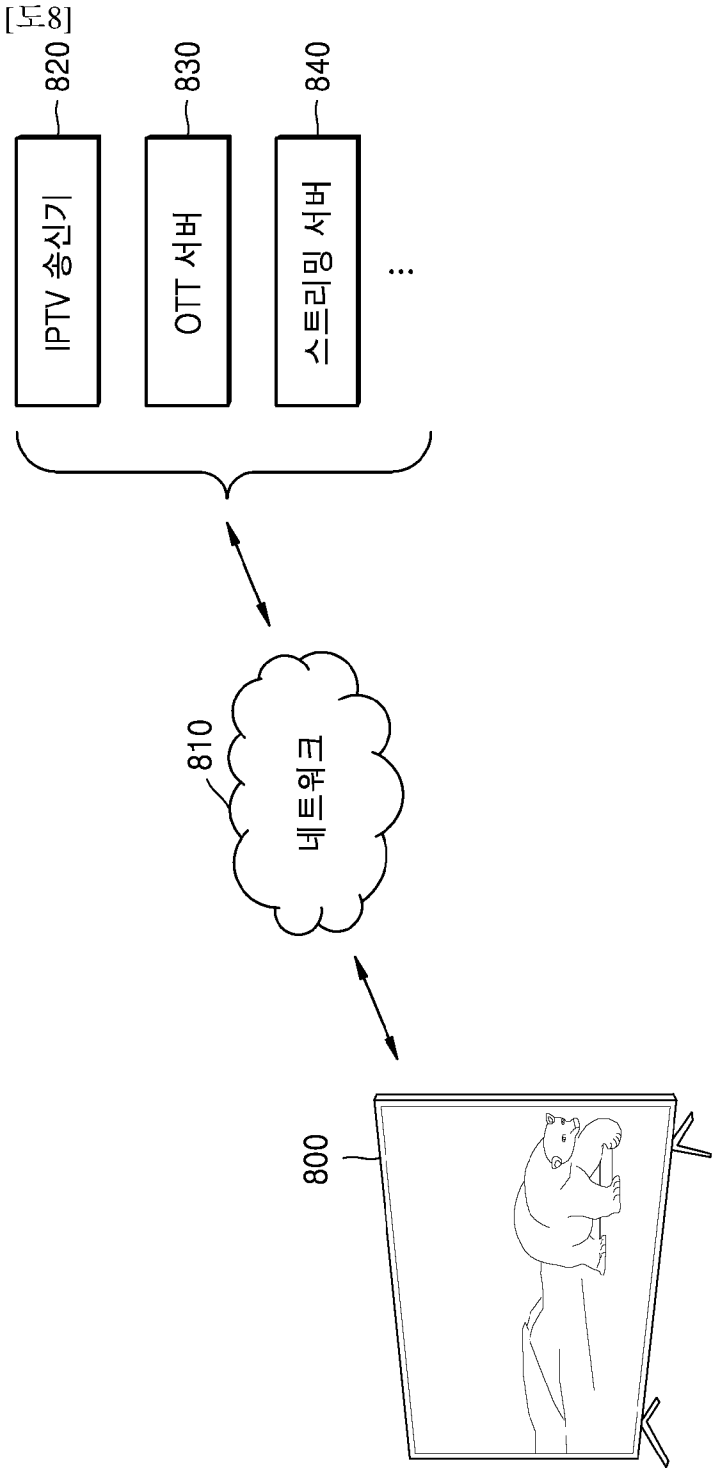
[도6b]


630
620
제1 이미지 특징에 대하여 제1 컨볼루션 연산을 수행하고, 제2 이미지 특징에 대하여 제2 컨볼루션 연산을 수행
631
제1 이미지 특징과 제2 이미지 특징 사이의 픽셀 간 이동량을 학습하는 오프셋 네트워크를 이용하여, 제1 컨볼루션 연산 결과 및 제2 컨볼루션 연산 결과에 기초하여 오프셋이 적용된 이미지 특징을 생성
632
오프셋이 적용된 이미지 특징에 대하여 제3 컨볼루션 연산을 수행
633
오프셋이 적용된 이미지 특징에 대하여 제4 컨볼루션 연산을 수행
634
제2 컨볼루션 연산 결과를 쿼리로서 이용하고, 제3 컨볼루션 연산 결과를 키로서 이용하고, 제4 컨볼루션 연산 결과를 밸류로서 이용하는 셀프-어텐션 연산을 수행하여 시간적 특징을 생성
635
640

[도7]


700
710
프로세서
720
메모리
730
통신 인터페이스

[도8]



INTERNATIONAL SEARCH REPORT

International application No.
PCT/KR2024/002801

A. CLASSIFICATION OF SUBJECT MATTER

G06T 5/00(2006.01)i; **G06T 7/11**(2017.01)i; **G06T 3/40**(2006.01)i; **G06V 10/44**(2022.01)i; **G06V 10/62**(2022.01)i; **G06V 10/52**(2022.01)i; **G06V 20/40**(2022.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06T 5/00(2006.01); G06F 18/00(2023.01); G06K 9/62(2006.01); G06N 3/08(2006.01); G06T 17/00(2006.01); G06T 3/40(2006.01); G06T 7/00(2006.01); G06T 7/254(2017.01)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models: IPC as above
Japanese utility models and applications for utility models: IPC as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS (KIPO internal) & keywords: 장면(scene), 이미지(image), 타겟(target), 프레임(frame), 시간적 특징(temporal feature), 컨볼루션(convolution), 픽셀(pixel), 오프셋(offset), 쿼리(query), 키(key), 벨류(value), 셀프-어텐션(Self-Attention)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	KR 10-2019-0123946 A (INDUSTRY-ACADEMIC COOPERATION FOUNDATION, YONSEI UNIVERSITY) 04 November 2019 (2019-11-04) See paragraphs [0030]-[0033]; and claim 9.	1-15
A	KR 10-2023-0044830 A (KAKAO ENTERPRISE CORP.) 04 April 2023 (2023-04-04) See claims 1-15.	1-15
A	JP 2017-191608 A (RICOH CO., LTD.) 19 October 2017 (2017-10-19) See claims 1-13.	1-15
A	CN 114897726 A (UNIVERSITY ZHONGSHAN) 12 August 2022 (2022-08-12) See claims 1-10.	1-15
A	CN 113362230 A (UNIVERSITY KUNMING SCIENCE & TECHNOLOGY) 07 September 2021 (2021-09-07) See claims 1-6.	1-15

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
04 June 2024	**04 June 2024**

Name and mailing address of the ISA/KR	Authorized officer
Korean Intellectual Property Office **Government Complex-Daejeon Building 4, 189 Cheongsa-ro, Seo-gu, Daejeon 35208**	
Facsimile No. **+82-42-481-8578**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/KR2024/002801

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
KR 10-2019-0123946 A	04 November 2019	KR 10-2044626 B1	13 November 2019
KR 10-2023-0044830 A	04 April 2023	None	
JP 2017-191608 A	19 October 2017	CN 107305635 A	31 October 2017
		EP 3232371 A1	18 October 2017
CN 114897726 A	12 August 2022	None	
CN 113362230 A	07 September 2021	None	

국 제 조 사 보 고 서

국제출원번호
PCT/KR2024/002801

A. 발명이 속하는 기술분류(국제특허분류(IPC))

G06T 5/00(2006.01)i; **G06T 7/11**(2017.01)i; **G06T 3/40**(2006.01)i; **G06V 10/44**(2022.01)i; **G06V 10/62**(2022.01)i; **G06V 10/52**(2022.01)i; **G06V 20/40**(2022.01)i

B. 조사된 분야

조사된 최소문헌(국제특허분류를 기재)

G06T 5/00(2006.01); G06F 18/00(2023.01); G06K 9/62(2006.01); G06N 3/08(2006.01); G06T 17/00(2006.01); G06T 3/40(2006.01); G06T 7/00(2006.01); G06T 7/254(2017.01)

조사된 기술분야에 속하는 최소문헌 이외의 문헌

한국등록실용신안공보 및 한국공개실용신안공보: 조사된 최소문헌란에 기재된 IPC
일본등록실용신안공보 및 일본공개실용신안공보: 조사된 최소문헌란에 기재된 IPC

국제조사에 이용된 전산 데이터베이스(데이터베이스의 명칭 및 검색어(해당하는 경우))

eKOMPASS(특허청 내부 검색시스템) & 키워드: 장면(scene), 이미지(image), 타겟(target), 프레임(frame), 시간적 특징(temporal feature), 컨볼루션(convolution), 픽셀(pixel), 오프셋(offset), 쿼리(query), 키(key), 벨류(value), 셀프-어텐션(Self-Attention)

C. 관련 문헌

카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
A	KR 10-2019-0123946 A (연세대학교 산학협력단) 2019.11.04 단락 [0030]-[0033]; 및 청구항 9	1-15
A	KR 10-2023-0044830 A (주식회사 카카오엔터프라이즈) 2023.04.04 청구항 1-15	1-15
A	JP 2017-191608 A (RICOH CO., LTD.) 2017.10.19 청구항 1-13	1-15
A	CN 114897726 A (UNIVERSITY ZHONGSHAN) 2022.08.12 청구항 1-10	1-15
A	CN 113362230 A (UNIVERSITY KUNMING SCIENCE & TECHNOLOGY) 2021.09.07 청구항 1-6	1-15

☐ 추가 문헌이 C(계속)에 기재되어 있습니다.

☑ 대응특허에 관한 별지를 참조하십시오.

* 인용된 문헌의 특별 카테고리:

"A" 특별히 관련이 없는 것으로 보이는 일반적인 기술수준을 정의한 문헌
"D" 본 국제출원에서 출원인이 인용한 문헌
"E" 국제출원일보다 빠른 출원일 또는 우선일을 가지나 국제출원일 이후에 공개된 선출원 또는 특허 문헌
"L" 우선권 주장에 의문을 제기하는 문헌 또는 다른 인용문헌의 공개일 또는 다른 특별한 이유(이유를 명시)를 밝히기 위하여 인용된 문헌
"O" 구두 개시, 사용, 전시 또는 기타 수단을 언급하고 있는 문헌
"P" 우선일 이후에 공개되었으나 국제출원일 이전에 공개된 문헌
"T" 국제출원일 또는 우선일 후에 공개된 문헌으로, 출원과 상충하지 않으며 발명의 기초가 되는 원리나 이론을 이해하기 위해 인용된 문헌
"X" 특별한 관련이 있는 문헌. 해당 문헌 하나만으로 청구된 발명의 신규성 또는 진보성이 없는 것으로 본다.
"Y" 특별한 관련이 있는 문헌. 해당 문헌이 하나 이상의 다른 문헌과조합하는 경우로 그 조합이 당업자에게 자명한 경우 청구된 발명은 진보성이 없는 것으로 본다.
"&" 동일한 대응특허문헌에 속하는 문헌

국제조사의 실제 완료일	국제조사보고서 발송일
2024년06월04일(04.06.2024)	**2024년06월04일(04.06.2024)**

ISA/KR의 명칭 및 우편주소	심사관
대한민국 특허청 **(35208) 대전광역시 서구 청사로 189, 4동 (둔산동, 정부대전청사)** 팩스 번호 **+82-42-481-8578**	**양정록** 전화번호 **+82-42-481-5709**

국 제 조 사 보 고 서
대응특허에 관한 정보

국제출원번호
PCT/KR2024/002801

국제조사보고서에서 인용된 특허문헌	공개일	대응특허문헌	공개일
KR 10-2019-0123946 A	2019/11/04	KR 10-2044626 B1	2019/11/13
KR 10-2023-0044830 A	2023/04/04	없음	
JP 2017-191608 A	2017/10/19	CN 107305635 A	2017/10/31
		EP 3232371 A1	2017/10/18
CN 114897726 A	2022/08/12	없음	
CN 113362230 A	2021/09/07	없음	