(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년01월30일
(11) 등록번호 10-2493980
(24) 등록일자 2023년01월26일

(51) 국제특허분류(Int. Cl.)
G06N 20/00 (2019.01) *G06T 3/40* (2006.01)
G06T 5/00 (2019.01) *G06T 7/20* (2017.01)
(52) CPC특허분류
G06N 20/00 (2021.08)
G06T 3/4046 (2013.01)
(21) 출원번호 10-2020-0180046
(22) 출원일자 2020년12월21일
심사청구일자 2020년12월21일
(65) 공개번호 10-2022-0089431
(43) 공개일자 2022년06월28일
(56) 선행기술조사문헌
KR1020190114340 A
"합성곱 신경망을 적용한 동영상의 프레임율과 해상도 상향변환 시스템", 서강대학교 대학원 컴퓨터공학과, 2016.
KR1020190062129 A
KR1020190103916 A

(73) 특허권자
포항공과대학교 산학협력단
경상북도 포항시 남구 청암로 77 (지곡동)
(72) 발명자
이승용
경상북도 포항시 남구 지곡로 155, 8동 803호
이준용
경상북도 포항시 북구 양덕로50번길 33, 101동 1602호
(뒷면에 계속)
(74) 대리인
특허법인이상

전체 청구항 수 : 총 16 항 심사관 : 박승철

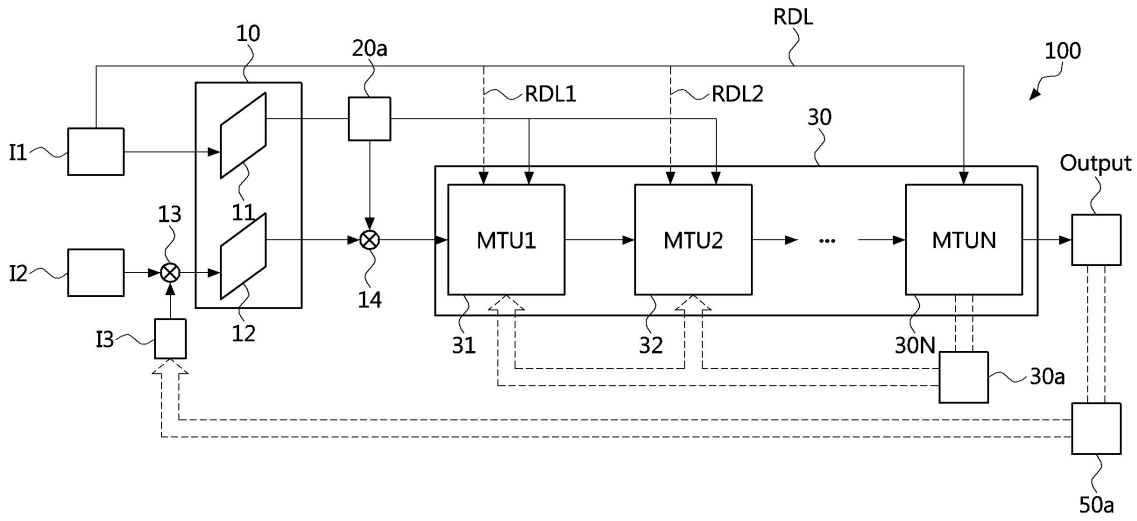
(54) 발명의 명칭 **머신러닝 기반의 비디오 화질개선 방법 및 장치**

(57) 요 약

구조-디테일 분리 기반으로 비디오 화질 개선을 수행하는 머신러닝 기반의 비디오 화질개선 방법 및 장치가 개시된다. 비디오 화질개선 방법은, 입력 프레임 중 하나인 현재 타겟 프레임으로부터 제1 컨볼루션 레이어에 의해 변환된 구조 피처맵을 복수의 다중 태스크 유닛들 중 제1 다중 태스크 유닛과 제1 다중 태스크 유닛의 출력측에

(뒷면에 계속)

대 표 도



연결되는 제2 다중 태스크 유닛에 입력하는 단계와, 입력 프레임 중 다른 하나인 이전 타겟 프레임 및 상기 입력 프레임 중 또 다른 하나인 이전 프레임의 보정 프레임을 채널 차원으로 연결한 것으로부터 제2 컨볼루션 레이어에 의해 변환된 특징 공간에 상기 구조 피처맵을 추가한 메인 입력을 제1 다중 태스크 유닛에 입력하는 단계와, 제2 다중 태스크 유닛의 출력측의 마지막에 연결되는 제N 다중 태스크 유닛에 상기 현재 타겟 프레임을 입력하는 단계를 포함한다. 제N 다중 태스크 유닛에서 출력되는 현재 타겟 프레임에 대한 화질개선 프레임인 보정 프레임을 통해 계산된 목적함수를 이용해 비디오 화질개선 모델의 기계학습을 수행한다.

(52) CPC특허분류
G06T 5/001 (2013.01)
G06T 7/20 (2013.01)
G06T 2207/20081 (2013.01)

(72) 발명자
손형석
경상북도 포항시
조성현
경상북도 포항시 남구 지곡로 278, 207동 102호

이 발명을 지원한 국가연구개발사업

과제고유번호	1711103034
과제번호	2015-0-00174-006
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	SW컴퓨팅산업원천기술개발
연구과제명	(SW 스타랩) 빅 비주얼 데이터 기반의 고품질 사진 메이크업 SW 개발
기 여 율	70/100
과제수행기관명	포항공과대학교 산학협력단
연구기간	2020.01.01 ~ 2020.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711097558
과제번호	2017M3C4A7066317
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	차세대정보·컴퓨팅기술개발(R&D)
연구과제명	초실감 원격가상 인터랙션을 위한 teleportation 기술 개발
기 여 율	30/100
과제수행기관명	포항공과대학교
연구기간	2020.04.01 ~ 2020.12.31

명 세 서

청구범위

청구항 1

비디오 화질개선 장치에서 복수의 다중 태스크 유닛들을 쌓은 컨볼루션 신경망 기반의 비디오 화질개선 모델을 기계학습하는 비디오 화질개선 방법으로서,

입력 프레임 중 하나인 현재 타겟 프레임으로부터 제1 컨볼루션 레이어에 의해 변환된 구조 피처맵(structure feature map)을 상기 복수의 다중 태스크 유닛들 중 제1 다중 태스크 유닛과 상기 제1 다중 태스크 유닛의 출력측에 연결되는 제2 다중 태스크 유닛에 입력하는 단계;

상기 입력 프레임 중 다른 하나인 이전 타겟 프레임 및 상기 입력 프레임 중 또 다른 하나인 이전 프레임의 보정 프레임을 채널 차원으로 연결한 것으로부터 제2 컨볼루션 레이어에 의해 변환된 특징 공간에 상기 구조 피처맵을 추가한 메인 입력을 상기 제1 다중 태스크 유닛에 입력하는 단계; 및

상기 제2 다중 태스크 유닛의 출력측의 마지막에 연결되는 제N 다중 태스크 유닛에 상기 현재 타겟 프레임을 입력하는 단계;를 포함하며,

상기 제N 다중 태스크 유닛은 상기 현재 타겟 프레임의 보정 프레임을 출력하고,

상기 현재 타겟 프레임의 보정 프레임을 통해 계산된 목적함수를 이용해 비디오 화질개선 모델의 기계학습을 수행하는, 머신러닝 기반의 비디오 화질개선 방법.

청구항 2

청구항 1에 있어서,

상기 제N 다중 태스크 유닛에서 생성된 상기 이전 타겟 프레임의 제N 화질개선 피처맵(N^{th} deblurred feature map)을 상기 제1 다중 태스크 유닛과 상기 제2 다중 태스크 유닛에 입력하는 단계를 더 포함하는 머신러닝 기반의 비디오 화질개선 방법.

청구항 3

청구항 2에 있어서,

상기 제1 다중 태스크 유닛의 제1 디테일 피처맵 네트워크에 의해 상기 메인 입력을 받고 제1 디테일 피처맵을 출력하는 단계; 및

상기 제1 다중 태스크 유닛의 정합 네트워크의 모션 레이어에 의해 상기 제1 디테일 피처맵에 상기 구조 피처맵을 추가한 구조 주입 피처맵(structure-injected feature map)을 현재 프레임 피처맵으로 변환하는 단계;

를 더 포함하는 머신러닝 기반의 비디오 화질개선 방법.

청구항 4

청구항 3에 있어서,

상기 정합 네트워크의 모션 보상 모듈에 의해 상기 현재 프레임 피처맵과 이전 타겟 프레임의 이전 프레임 피처맵과의 모션을 추정하고 추정된 모션에 기초하여 상기 이전 타겟 프레임의 화질개선 피처맵을 상기 현재 타겟 프레임에 대해 정렬하는 단계;를 더 포함하는 머신러닝 기반의 비디오 화질개선 방법.

청구항 5

청구항 4에 있어서,

상기 정합 네트워크의 연결 네트워크를 통해 상기 이전 타겟 프레임의 정렬된 화질개선 피처맵과 상기 제1 디테일 피처맵을 채널 차원으로 연결한 제1 화질개선 피처맵을 상기 제2 다중 태스크 유닛의 입력측으로 출력하는

단계;를 더 포함하는 머신러닝 기반의 비디오 화질개선 방법.

청구항 6

청구항 5에 있어서,

상기 컨볼루션 신경망의 훈련 시, 상기 제1 다중 태스크 유닛의 제1 디블러링 네트워크의 디블러 레이어에 의해 상기 제1 디테일 피처맵을 출력 잔차 이미지로 변환하고 상기 디블러 레이어의 출력측에 연결되는 스킵 커넥션을 통해 상기 현재 타겟 프레임을 더한 제1 보정 프레임을 출력하는 단계를 더 포함하며,

상기 제1 보정 프레임은 상기 제1 다중 태스크 유닛의 가중치를 업데이트하는데 이용되는 머신러닝 기반의 비디오 화질개선 방법.

청구항 7

청구항 1에 있어서,

상기 기계학습을 실행할 때에 상기 현재 타겟 프레임의 보정 프레임과 선명한 교사 프레임의 평균 제곱근 편차를 구하여 목적함수를 최소화하는 단계를 더 포함하는 머신러닝 기반의 비디오 화질개선 방법.

청구항 8

청구항 7에 있어서,

상기 기계학습을 실행할 때에 상기 이전 타겟 프레임의 교사 프레임과 상기 현재 타겟 프레임의 교사 프레임으로부터 생성한 교사 옵티컬플로우를 이용하여 상기 제1 다중 태스크 유닛 및 상기 제2 다중 태스크 유닛의 정합 네트워크들에서 각각 생성한 상관관계 매트릭스(correlation matrix)의 각 픽셀별 크로스 엔트로피(cross entropy)를 구하여 목적함수를 최소화하는 단계를 더 포함하는 머신러닝 기반의 비디오 화질개선 방법.

청구항 9

입력 프레임 중 하나인 현재 타겟 프레임을 변환하여 구조 피처맵(structure feature map)을 생성하는 제1 컨볼루션 레이어;

상기 입력 프레임 중 다른 하나인 이전 타겟 프레임 및 상기 입력 프레임 중 또 다른 하나인 이전 타겟 프레임의 보정 프레임을 채널 차원으로 연결한 것을 변환하여 특징 공간을 생성하는 제2 컨볼루션 레이어;

상기 특징 공간에 상기 구조 피처맵을 추가한 메인 입력을 생성하는 연결 네트워크;

상기 메인 입력과 상기 구조 피처맵을 입력받고 제1 디테일 피처맵 네트워크와 제1 정합 네트워크를 통해 상기 현재 타겟 프레임의 제1 화질개선 피처맵을 생성하는 컨볼루션 신경망 기반의 제1 다중 태스크 유닛; 및

상기 제1 다중 태스크 유닛의 출력측의 마지막에 연결되고 상기 현재 타겟 프레임을 입력받고 제N 디테일 피처맵 네트워크와 제N 디블러링 네트워크를 통해 상기 현재 타겟 프레임의 보정 프레임을 출력하는 제N 다중 태스크 유닛;

을 포함하며,

상기 현재 타겟 프레임의 보정 프레임을 통해 계산된 목적함수를 이용해 비디오 화질개선 모델의 기계학습을 수행하는, 머신러닝 기반의 비디오 화질개선 장치.

청구항 10

청구항 9에 있어서,

상기 제1 다중 태스크 유닛과 상기 제N 다중 태스크 유닛 사이에 제2 다중 태스크 유닛을 포함한 하나 이상의 다중 태스크 유닛을 더 포함하며,

상기 제2 다중 태스크 유닛은 상기 제1 다중 태스크 유닛의 출력측에 연결되어 상기 제1 화질개선 피처맵을 입력받고 상기 구조 피처맵을 입력받으며 제2 디테일 피처맵 네트워크와 제2 정합 네트워크를 통해 상기 현재 타겟 프레임의 제2 화질개선 피처맵을 생성하는, 머신러닝 기반의 비디오 화질개선 장치.

청구항 11

청구항 10에 있어서,

상기 제N 다중 태스크 유닛에서 생성된 상기 이전 타겟 프레임의 제N 화질개선 피처맵(N^{th} deblurred feature map)은 상기 제1 다중 태스크 유닛과 상기 제2 다중 태스크 유닛을 포함한 하나 이상의 다중 태스크 유닛에 입력되는, 머신러닝 기반의 비디오 화질개선 장치.

청구항 12

청구항 9에 있어서,

상기 제1 다중 태스크 유닛은,

상기 메인 입력을 받고 제1 디테일 피처맵을 생성하는 제1 디테일 피처맵 네트워크; 및

모션 레이어에 의해 상기 제1 디테일 피처맵에 상기 구조 피처맵을 추가한 구조-주입 피처맵(structure-injected feature map)을 현재 프레임 피처맵으로 변환하는 제1 정합 네트워크;를 포함하는 머신러닝 기반의 비디오 화질개선 장치.

청구항 13

청구항 12에 있어서,

상기 정합 네트워크는,

상기 현재 프레임 피처맵과 이전 프레임 피처맵과의 모션을 추정하고 추정된 모션에 기초하여 상기 이전 프레임의 화질개선 피처맵을 상기 현재 타겟 프레임에 대해 정렬하는 모션 보상 모듈; 및

상기 모션 보상 모듈의 출력측에 연결되며 상기 이전 타겟 프레임의 정렬된 화질개선 피처맵과 상기 제1 디테일 피처맵을 채널 차원으로 연결한 제1 화질개선 피처맵을 신호 흐름상 상기 제1 다중 태스크 유닛의 후단에 위치하는 제2 다중 태스크 유닛의 입력측으로 출력하는 연결 네트워크;를 구비하는 머신러닝 기반의 비디오 화질개선 장치.

청구항 14

청구항 13에 있어서,

상기 제1 다중 태스크 유닛은, 디블러 레이어에 의해 상기 제1 디테일 피처맵을 출력 잔차 이미지로 변환하고 상기 디블러 레이어의 출력측에 연결되는 스킵 커넥션을 통해 상기 현재 타겟 프레임을 더한 제1 보정 프레임을 출력하는 제1 디블러링 네트워크를 더 포함하며,

상기 제1 보정 프레임은 상기 제1 다중 태스크 유닛의 가중치를 업데이트하는데 이용되는 머신러닝 기반의 비디오 화질개선 장치.

청구항 15

청구항 9에 있어서,

상기 현재 타겟 프레임의 보정 프레임과 선명한 교사 프레임의 평균 제곱근 편차를 구하여 목적함수를 최소화하는 최적화 유닛을 더 포함하는 머신러닝 기반의 비디오 화질개선 장치.

청구항 16

청구항 15에 있어서,

상기 최적화 유닛은 상기 이전 타겟 프레임의 교사 프레임과 상기 현재 타겟 프레임의 교사 프레임으로부터 생성한 교사 옵티컬플로우를 이용하여 상기 제1 다중 태스크 유닛의 정합 네트워크에서 생성한 상관관계 매트릭스(correlation matrix)의 각 픽셀별 크로스 엔트로피(cross entropy)를 구하여 목적함수를 추가로 최소화하는 머신러닝 기반의 비디오 화질개선 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 비디오 화질개선 기술에 관한 것으로, 보다 구체적으로는 구조-디테일 분리 기반으로 블러(blur)를 포함하는 비디오의 화질 개선을 수행하는 머신러닝 기반의 비디오 화질개선 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 대부분의 기존 비디오 화질 개선 기법들은 비디오의 프레임들을 화질 개선 하기 위해 각 프레임의 주변 프레임들의 정보를 이용한다. 또한, 주변 프레임들을 화질 개선하기 위한 프레임에 정합(align)하여 화질 개선 성능의 향상을 노린다.

[0003] 하지만, 기존 기법들에서는 블러를 포함하는 프레임들을 정확히 정합하는 것은 쉽지 않고, 오히려 잘못 정합된 프레임들은 화질 개선 성능 저하의 원인이 된다. 이를 해결하기 위해 종래 문헌(Sunghyun Cho, Jue Wang, and Seungyong Lee. Video deblurring for hand-held cameras using patch-based synthesis. ACM Transactions on Graphics, 31(4):64:1-64:9, 2012)과 같이 화질 개선과 정합을 반복적으로 수행하는 점진적 화질 개선-정합 기법이 제안되었지만, 정합에 필요한 높은 연산량 때문에, 실제 사용에 제약이 있다.

[0004] 이와 같이 블러를 포함하는 비디오의 화질개선을 위한 새로운 방안이 요구되고 있는 실정이다.

발명의 내용

해결하려는 과제

[0005] 본 발명은 전술한 종래 기술의 요구에 부응하기 위해 도출된 것으로, 본 발명의 목적은 다중 태스크 유닛들을 쌓은 비디오 화질개선 네트워크에 구조-디테일 분리 기반의 학습 구조를 채용함으로써 모션 블러를 포함하는 비디오의 화질 개선 과정에서 다중 태스크 유닛들이 서로 상충하는 특성의 피처맵을 요구하는 화질 개선 작업과 정합 작업을 조화롭게 학습할 수 있는, 머신러닝 기반의 비디오 화질개선 방법 및 장치를 제공하는데 있다.

[0006] 본 발명의 다른 목적은, 기존 대비 상대적으로 매우 낮은 가벼운 연산량을 가진 네트워크 모듈로 다중 태스크 유닛을 구현할 수 있고, 복수의 다중 태스크 유닛들을 쌓아 점진적 자동 화질개선 및 정합 유닛들로 작동시킴으로써 비디오 화질 개선에서 주변 프레임들을 이용하는 것, 주변 프레임들은 정합하는 것, 그리고 화질 개선과 정합을 점진적으로 수행하는 것을 효과적으로 통합하여 디블러링 성능을 현저하게 향상시킬 수 있는, 머신러닝 기반의 비디오 화질개선 방법 및 장치를 제공하는데 있다.

과제의 해결 수단

[0007] 상기 기술적 과제를 해결하기 위한 본 발명의 일 측면에 따른 머신러닝 기반의 비디오 화질개선 방법은, 비디오 화질개선 장치에서 복수의 다중 태스크 유닛들을 쌓은 컨볼루션 신경망 기반의 비디오 화질개선 모델을 기계학습하는 비디오 화질개선 방법으로서, 입력 프레임 중 하나인 현재 타겟 프레임으로부터 제1 컨볼루션 레이어에 의해 변환된 구조 피처맵(structure feature map)을 상기 복수의 다중 태스크 유닛들 중 제1 다중 태스크 유닛과 상기 제1 다중 태스크 유닛의 출력측에 연결되는 제2 다중 태스크 유닛에 입력하는 단계; 상기 입력 프레임 중 다른 하나인 이전 타겟 프레임 및 상기 입력 프레임 중 또 다른 하나인 이전 프레임의 보정 프레임을 채널 차원으로 연결한 것으로부터 제2 컨볼루션 레이어에 의해 변환된 특징 공간에 상기 구조 피처맵을 추가한 메인 입력을 상기 제1 다중 태스크 유닛에 입력하는 단계; 및 상기 제2 다중 태스크 유닛의 출력측의 마지막에 연결되는 제N 다중 태스크 유닛에 상기 현재 타겟 프레임을 입력하는 단계;를 포함한다. 제N 다중 태스크 유닛은 상기 현재 타겟 프레임에 대한 화질개선 프레임인 보정 프레임을 출력한다. 머신러닝 기반의 비디오 화질개선 방법은 현재 타겟 프레임의 보정 프레임을 통해 계산된 목적함수를 이용해 비디오 화질개선 모델의 기계학습을 수행할 수 있다.

[0008] 일실시예에서, 비디오 화질개선 방법은, 상기 제N 다중 태스크 유닛에서 생성된 상기 이전 타겟 프레임의 제N 화질개선 피처맵(N^{th} deblurred feature map)을 상기 제1 다중 태스크 유닛과 상기 제2 다중 태스크 유닛에 입력하는 단계를 더 포함한다.

[0009] 일실시예에서, 비디오 화질개선 방법은, 상기 제1 다중 태스크 유닛의 제1 디테일 피처맵 네트워크에 의해 상기

메인 입력을 받고 제1 디테일 피처맵을 출력하는 단계; 및 상기 제1 다중 태스크 유닛의 정합 네트워크의 모션 레이어에 의해 상기 제1 디테일 피처맵에 상기 구조 피처맵을 추가한 구조 주입 피처맵(structure-injected feature map)을 현재 프레임 피처맵으로 변환하는 단계를 더 포함한다.

[0010] 일실시예에서, 비디오 화질개선 방법은, 상기 정합 네트워크의 모션 보상 모듈에 의해 상기 현재 프레임 피처맵과 이전 타겟 프레임의 이전 프레임 피처맵과의 모션을 추정하고 추정된 모션에 기초하여 상기 이전 타겟 프레임의 화질개선 피처맵을 상기 현재 타겟 프레임에 대해 정렬하는 단계를 더 포함한다.

[0011] 일실시예에서, 비디오 화질개선 방법은, 상기 정합 네트워크의 연결 네트워크를 통해 상기 이전 타겟 프레임의 정렬된 화질개선 피처맵과 상기 제1 디테일 피처맵을 채널 차원으로 연결한 제1 화질개선 피처맵을 상기 제2 다중 태스크 유닛의 입력측으로 출력하는 단계를 더 포함한다.

[0012] 일실시예에서, 비디오 화질개선 방법은, 상기 컨볼루션 신경망의 훈련 시, 상기 제1 다중 태스크 유닛의 제1 디블러링 네트워크의 디블러 레이어에 의해 상기 제1 디테일 피처맵을 출력 잔차 이미지로 변환하고 상기 디블러 레이어의 출력측에 연결되는 스킵 커넥션을 통해 상기 현재 타겟 프레임을 더한 제1 보정 프레임을 출력하는 단계를 더 포함한다. 상기 제1 보정 프레임은 상기 제1 다중 태스크 유닛의 가중치를 업데이트하는데 이용된다.

[0013] 일실시예에서, 비디오 화질개선 방법은, 상기 기계학습을 실행할 때에 상기 현재 타겟 프레임의 보정 프레임과 선명한 교사 프레임의 평균 제곱근 편차를 구하여 목적함수를 최소화하는 단계를 더 포함한다.

[0014] 일실시예에서, 비디오 화질개선 방법은, 상기 기계학습을 실행할 때에 상기 이전 타겟 프레임의 교사 프레임과 상기 현재 타겟 프레임의 교사 프레임으로부터 생성한 교사 옵티컬플로우를 이용하여 상기 제1 다중 태스크 유닛 및 상기 제2 다중 태스크 유닛의 정합 네트워크들에서 각각 생성한 상관관계 매트릭스(correlation matrix)의 각 픽셀별 크로스 엔트로피(cross entropy)를 구하여 목적함수를 최소화하는 단계를 더 포함할 수 있다.

[0015] 상기 기술적 과제를 해결하기 위한 본 발명의 일 측면에 따른 머신러닝 기반의 비디오 화질개선 장치는, 입력 프레임 중 하나인 현재 타겟 프레임을 변환하여 구조 피처맵(structure feature map)을 생성하는 제1 컨볼루션 레이어; 상기 입력 프레임 중 다른 하나인 이전 타겟 프레임 및 상기 입력 프레임 중 또 다른 하나인 이전 타겟 프레임의 보정 프레임을 채널 차원으로 연결한 것을 변환하여 특징 공간을 생성하는 제2 컨볼루션 레이어; 상기 특징 공간에 상기 구조 피처맵을 추가한 메인 입력을 생성하는 연결 네트워크; 상기 메인 입력과 상기 구조 피처맵을 입력받고 제1 디테일 피처맵 네트워크와 제1 정합 네트워크를 통해 상기 현재 타겟 프레임의 제1 화질개선 피처맵을 생성하는 컨볼루션 신경망 기반의 제1 다중 태스크 유닛; 및 상기 제1 다중 태스크 유닛의 출력측에 연결되고 상기 현재 타겟 프레임을 입력받고 제N 디테일 피처맵 네트워크와 제N 디블러링 네트워크를 통해 상기 현재 타겟 프레임의 보정 프레임을 출력하는 제N 다중 태스크 유닛을 포함한다. 비디오 화질개선 장치는 현재 타겟 프레임에 대한 화질개선 프레인인 보정 프레임을 통해 계산된 목적함수를 이용해 비디오 화질개선 모델의 기계학습을 수행할 수 있다.

[0016] 일실시예에서, 상기 제1 다중 태스크 유닛과 상기 제N 다중 태스크 유닛 사이에 제2 다중 태스크 유닛을 포함한 하나 이상의 다중 태스크 유닛을 더 포함하며, 상기 제2 다중 태스크 유닛은 상기 제1 다중 태스크 유닛의 출력측에 연결되어 상기 제1 화질개선 피처맵을 입력받고 상기 구조 피처맵을 입력받으며 제2 디테일 피처맵 네트워크와 제2 정합 네트워크를 통해 상기 현재 타겟 프레임의 제2 화질개선 피처맵을 생성한다.

[0017] 일실시예에서, 상기 제N 다중 태스크 유닛에서 생성된 상기 이전 타겟 프레임의 제N 화질개선 피처맵(N^{th} deblurred feature map)은 상기 제1 다중 태스크 유닛과 상기 제2 다중 태스크 유닛을 포함한 하나 이상의 다중 태스크 유닛에 입력된다.

[0018] 일실시예에서, 상기 제1 다중 태스크 유닛은, 상기 메인 입력을 받고 제1 디테일 피처맵을 생성하는 제1 디테일 피처맵 네트워크; 및 모션 레이어에 의해 상기 제1 디테일 피처맵에 상기 구조 피처맵을 추가한 구조-주입 피처맵(structure-injected feature map)을 현재 프레임 피처맵으로 변환하는 제1 정합 네트워크를 포함한다.

[0019] 일실시예에서, 상기 정합 네트워크는, 상기 현재 프레임 피처맵과 이전 프레임 피처맵과의 모션을 추정하고 추정된 모션에 기초하여 상기 이전 프레임의 화질개선 피처맵을 상기 현재 타겟 프레임에 대해 정렬하는 모션 보상 모듈; 및 상기 모션 보상 모듈의 출력측에 연결되며 상기 이전 타겟 프레임의 정렬된 화질개선 피처맵과 상기 제1 디테일 피처맵을 채널 차원으로 연결한 제1 화질개선 피처맵을 제2 다중 태스크 유닛의 입력측으로 출력하는 연결 네트워크를 구비한다. 제2 다중 태스크 유닛은 신호 흐름상 제1 다중 태스크 유닛의 후단에 직접 연결되도록 위치한다.

[0020] 일실시예에서, 상기 제1 다중 태스크 유닛은, 디블러 레이어에 의해 상기 제1 디테일 피처맵을 출력 잔차 이미지로 변환하고 상기 디블러 레이어의 출력측에 연결되는 스킵 커넥션을 통해 상기 현재 타겟 프레임을 더한 제1 보정 프레임을 출력하는 제1 디블러링 네트워크를 더 포함한다. 상기 제1 보정 프레임은 상기 제1 다중 태스크 유닛의 가중치를 업데이트하는데 이용될 수 있다.

[0021] 일실시예에서, 비디오 화질개선 장치는, 상기 현재 타겟 프레임의 보정 프레임과 선명한 교사 프레임의 평균 제곱근 편차를 구하여 목적함수를 최소화하는 최적화 유닛을 더 포함할 수 있다.

[0022] 일실시예에서, 상기 최소화 유닛은 상기 이전 타겟 프레임의 교사 프레임과 상기 현재 타겟 프레임의 교사 프레임으로부터 생성한 교사 옵티컬플로우를 이용하여 상기 제1 다중 태스크 유닛의 정합 네트워크에서 생성한 상관관계 매트릭스(correlation matrix)의 각 픽셀별 크로스 엔트로피(cross entropy)를 구하여 목적함수를 추가로 최소화할 수 있다.

발명의 효과

[0023] 전술한 머신러닝 기반의 비디오 화질개선 방법 및 장치를 사용하는 경우에는 다중 태스크 유닛들을 쌓은 비디오 화질개선 네트워크에 구조-디테일 분리 기반의 학습 구조를 채용함으로서 다중 태스크 유닛들이 서로 상충하는 특성의 피처맵을 요구하는 화질 개선 작업과 정합 작업을 조화롭게 학습할 수 있다.

[0024] 또한, 기존 대비 상대적으로 가벼운 연산량이면서 매우 낮은 연산량을 가진 네트워크 모듈로 다중 태스크 유닛을 구현하고 복수의 다중 태스크 유닛들을 쌓아 점진적 자동 화질개선 및 정합 유닛들로 동작하도록 구현함으로써 비디오 화질 개선에서 주변 프레임들을 이용하는 것, 주변 프레임들은 정합하는 것, 그리고 화질 개선과 정합을 점진적으로 수행하는 것을 통해 디블러링 성능을 현저하게 향상시킬 수 있다.

[0025] 또한, 점진적 비디오 화질개선을 위한 구조-디테일 분리 기반의 기계학습 방법 및 장치를 제공할 수 있고, 이를 통해 비디오 화질개선 모델의 기계학습을 효과적으로 수행하는데 기여할 수 있다.

[0026] 또한, 본 발명에 의하면, 다중 태스크 유닛이 가벼운 연산량을 가질 수 있도록 하는 구조-디테일 분리 기반의 학습 방법을 통해 다중 태스크 유닛들이 점진적으로 화질 개선과 정합을 수행함에 있어 종래 기술에서 낭비되었던 정합에 필요한 높은 연산량을 낮추면서 효율적으로 화질 개선에 도움을 줄 수 있는 효과가 있다.

도면의 간단한 설명

[0027] 도 1은 본 발명의 일실시예에 따른 머신러닝 기반의 비디오 화질개선 장치에 대한 블록도이다.

도 2는 도 1의 비디오 화질개선 장치의 전체 네트워크 아키텍처를 나타낸 블록도이다.

도 3은 도 2의 비디오 화질개선 장치의 다중 태스크 유닛의 기본 구성을 나타낸 블록도이다.

도 4는 도 1의 비디오 화질개선 장치의 변형예를 설명하기 위한 블록도이다.

도 5는 도 1의 비디오 화질개선 장치에 의한 점진적 디블러링-정합에 대한 정성적 결과를 설명하기 위한 도면이다.

도 6은 도 1의 비디오 화질개선 장치와 비교예들에 대한 정성적 결과를 비교하여 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0028] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세하게 설명하고자 한다. 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0029] 제1, 제2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

[0030] 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 이

해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.

[0031] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0032] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가진 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0033] 이하, 첨부한 도면들을 참조하여, 본 발명의 바람직한 실시예를 보다 상세하게 설명하고자 한다. 본 발명을 설명함에 있어 전체적인 이해를 용이하게 하기 위하여 도면상의 동일한 구성요소에 대해서는 동일한 참조부호를 사용하고 동일한 구성요소에 대해서 중복된 설명은 생략한다.

[0034] 먼저 도면을 참조하여 본 발명의 실시예를 상세히 설명하기에 앞서 본 발명은 효과적인 비디오 화질 개선을 위해 비디오의 각 프레임을 화질 개선할 때 화질 개선과 정합을 점진적으로 수행하도록 구성되는 것임을 밝힌다. 이를 위해, 본 발명에서는 화질 개선과 정합을 수행할 수 있는 다중 태스크 유닛을 여러 개 쌓아 이를 지도 학습한다. 효율적인 다중 태스크 유닛의 점진적 사용을 위해선 각 다중 태스크 유닛이 낮은 연산량을 가지고 있어야 하는데, 이를 위해 본 실시예에서는 구조-디테일 분리 기반의 학습 방법을 구현한다. 이하의 설명에서는 비디오 화질 개선의 예시로 본 발명을 비디오 디블러링에 적용한 학습 방법 및 결과를 중심으로 설명하기로 한다.

[0035] 그리고 구조-디테일 분리 기반의 학습에 대한 필요성을 간략히 설명하면 다음과 같다.

표 1

[0036]

Components	w/o $\mathcal{M}$		w/ $\mathcal{M}$			
	Deblurring		Motion compensation		Deblurring	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
w/o a long skip-connection	29.44	0.909	**26.73**	**0.897**	29.83	0.914
w/ a long skip-connection	**29.78**	**0.913**	24.14	0.832	**29.94**	**0.916**

[0037] [표 1]은 디블러링 및 정합에 필요한 피처맵의 특성에 대한 실험 결과이다. 본 실험에서는 정합(M)이 없는 경우와 있는 경우, 긴 스킵 커넥션(long skip-connection)을 사용하는 경우와 사용하지 않는 경우에 대한 디블러링(Deblurring) 성능과 모션 정합(Motion compensation) 성능을 PSNR(Peak signal-to-Noise Ration)와 SSIM(Structrual similarity)을 기준으로 평가하였다.

[0038] 효율적인 점진적 디블러링-정합을 위해서는 이를 수행하는 다중 태스크 유닛의 연산량이 낮아야 하며, 다중 태스크 유닛이 추출하는 피처맵의 특성이 디블러링과 정합을 수행함에 있어 상충하지 않아야 한다. 하지만, 각 다중 태스크 유닛의 피처맵이 입력 프레임의 디테일 정보를 포함해야 효과적인 디블러링을 수행할 수 있으며 피처맵이 입력 프레임의 디테일 정보뿐만 아니라 구조 정보를 함께 포함해야 효과적인 프레임 정합을 수행할 수 있는 점을 고려하여, 본 실시예에서는 [표 1]의 실험 결과를 참조하여 비디오 화질개선을 위한 신경망이 디블러링을 수행할 때, 피처맵이 디테일 정보를 포함하도록 하는 긴 스킵 커넥션을 사용한다. 그 경우, 디블러링의 성능이 증가함을 볼 수 있다([표 1]의 2-3번째 열).

[0039] 한편, 전체 네트워크가 디블러링과 프레임 정합(M)을 함께 수행할 때에는 긴 스킵 커넥션을 통해 피처맵이 구조 정보를 추가로 포함해야 정합의 성능을 향상시킬 수 있음을 확인할 수 있다([표 1]의 4-5번째 열). 이를 통해 디블러링과 프레임 정합이 필요로 하는 피처맵의 특성이 서로 다름을 알 수 있다. 즉 디블러링, 디테일과 프레임 정합, 그리고 디테일과 구조의 조합에서 피처맵의 특성은 서로 다르다.

[0040] 또한, 디블러링과 프레임 정합을 함께 수행하는 경우, 긴 스킵 커넥션을 사용하면 정합의 성능은 하락하지만,

디블러링의 성능이 프레임 정합을 하지 않았을 때보다 더 높다는 것 또한 확인할 수 있다([표 1]의 6-7번째 열).

[0041] 이와 같이 프레임 정합이 디블러링 성능에 큰 영향을 끼친다는 결론을 내릴 수 있으며, 이에 본 발명에서는 프레임 정합 성능을 더 향상시켜 더 높은 디블러링 성능을 도출하고자 한다.

[0042] 위의 실험을 통해 추론한 사실을 기반으로, 본 발명에서는 하나의 다중 태스크 유닛이 디블러링-정합을 수행함에 있어 서로 상충하는 피처맵의 특성을 조화롭게 하여 효과적이고 효율적인 점진적 디블러링-정합을 수행할 수 있도록 하는 구조-디테일 분리 기반의 학습 방법을 사용한다.

[0043] 구조-디테일 분리 기반의 학습 방법은 하나의 다중 태스크 유닛이 2개의 분리된 피처맵 줄기(stream)를 유지하도록 구현된다. 디블러링을 수행하기 위한 메인 줄기(main stream)는, 디테일 정보를 포함하는 디테일 피처맵(detail feature map)이 전체 네트워크에 흐르도록 배치하고, 정합을 수행하기 위한 보조 줄기는 미리 변환하여 준비한 구조 정보를 가진 구조 피처맵(structural feature map)을 유지하다가 디테일 정보가 흐르는 메인 줄기에 주기적으로 합류시켜 각 다중 태스크 유닛이 정합을 수행할 때 추가적으로 필요한 구조 정보를 점진적으로 포함시키도록 구현된다.

[0044] 이러한 구조-디테일 분리 기반의 학습 방법은 전체 네트워크가 다중 작업(디블러링-정합)을 수행할 때, 각 작업이 필요로 하는 특성의 피처맵을 적재적소에 제공하도록 학습돼 다중 태스크 유닛의 낮은 연산량을 보장하고, 여러 개의 다중 태스크 유닛을 쌓아도 2개의 기본 줄기들을 계속 유지할 수 있어 효율적인 점진적 디블러링-정합 수행을 가능케한다.

[0045] 전술한 구조-디테일 분리 기반의 학습 방법은 전체 네트워크 구조를 나타내면 도 1과 같다.

[0046] 도 1은 본 발명의 일실시예에 따른 머신러닝 기반의 비디오 화질개선 장치에 대한 블록도이다.

[0047] 도 1을 참조하면, 본 실시예에 따른 비디오 화질개선 장치(100)는 입력 변환 유닛(10)과, 복수의 다중 태스크 유닛들을 포함한 다중 유닛 스택(30)를 포함한다. 입력 변환 유닛(10)은 제1 컨볼루션 레이어(11)와 제2 컨볼루션 레이어(12)를 구비하고, 다중 유닛 스택(30)은 제1 다중 태스크 유닛(MTU1, 31), 제1 다중 태스크 유닛(31)의 출력측에 연결되는 제2 다중 태스크 유닛(MTU2, 32), 및 제2 다중 태스크 유닛(32)의 출력측에 연결되는 제N 다중 태스크 유닛(MTU N, 30N)을 구비한다.

[0048] 여기서, N은 2 이상의 임의의 자연수이다. N이 2인 경우, 제N 다중 태스크 유닛(30N)은 제2 다중 태스크 유닛이 되고 상기의 제2 다중 태스크 유닛(32)은 생략될 수 있다. 또한, N이 3인 경우, 제2 다중 태스크 유닛(32)과 제N 다중 태스크 유닛(30N) 사이에 추가되는 다른 다중 태스크 유닛은 생략될 수 있다. 아울러, N이 3보다 큰 경우, 제2 다중 태스크 유닛(32)과 제N 다중 태스크 유닛(30N) 사이에는 적어도 하나 이상의 다른 다중 태스크 유닛이 쌓아지고, 이때 제N 다중 태스크 유닛(30N)은 하나 이상의 다른 다중 태스크 유닛들 중 제2 다중 태스크 유닛(32)에서 신호 흐름 상 가장 먼 위치인 마지막에 위치한 다중 태스크 유닛의 출력측에 연결된다.

[0049] 입력 변환 유닛(10)은 3개의 입력 프레임을 받는다. 3개의 입력 프레임은 현재 타겟 프레임(I1), 이전 타겟 프레임(I2) 및 이전 타겟 프레임의 보정 프레임(I3; 50a)을 포함한다. 또한 입력 변환 유닛(10)은 제1 컨볼루션 레이어(11)을 통해 현재 타겟 프레임(I1)을 변환하여 구조 피처맵(structure feature map, 20a)을 생성한다. 그리고 제2 컨볼루션 레이어(12)를 통해 이전 타겟 프레임(I2) 및 이전 타겟 프레임의 보정 프레임(I3)을 연결 연산자나 연결 네트워크(13)를 통해 채널 차원으로 연결한 것을 변환하여 특징 공간에 대응하는 입력 피처맵을 생성한다. 입력 피처맵의 특징 공간에 또 다른 연결 네트워크(14)를 통해 구조 피처맵(20a)을 추가한 메인 입력은 제1 다중 태스크 유닛(31)에 입력된다.

[0050] 제1 다중 태스크 유닛(31)은 상기의 메인 입력과 구조 피처맵(20a)을 입력받고 제1 디테일 피처맵 네트워크와 제1 정합 네트워크를 통해 현재 타겟 프레임의 제1 화질개선 피처맵을 생성하고, 학습 시, 제1 디테일 피처맵 네트워크와 제1 디블러링 네트워크를 통해 현재 타겟 프레임의 화질개선 프레임인 제1 보정 프레임을 생성한다. 제1 디블러링 네트워크는 잔차 디테일 학습 경로(RDL)의 제1 가지 경로(RDL1)를 통해 현재 타겟 프레임(I1)을 입력받을 수 있다. 제1 보정 프레임은 제1 다중 태스크 유닛(31)의 가중치를 업데이트하는데 이용될 수 있다.

[0051] 제2 다중 태스크 유닛(32)은 제1 다중 태스크 유닛(31)의 출력측에 연결된 제1 입력단에서 제1 화질개선 피처맵을 입력받고 제2 입력단에서 구조 피처맵(20a)을 입력받으며 제2 디테일 피처맵 네트워크와 제2 정합 네트워크를 통해 현재 타겟 프레임의 제2 화질개선 피처맵을 생성한다.

[0052] 또한, 제2 다중 태스크 유닛(32)은, 학습 시, 제2 디테일 피처맵 네트워크와 제2 디블러링 네트워크를 통해 현

재 타겟 프레임의 화질개선 프레임인 제2 보정 프레임을 생성한다. 이때, 제2 다중 태스크 유닛은 잔차 디테일 학습 경로(RDL)의 제2 가지 경로(RDL2)를 통해 현재 타겟 프레임(I1)을 입력받을 수 있다. 제2 보정 프레임은 제2 다중 태스크 유닛(32)의 가중치를 업데이트하는데 이용될 수 있다.

[0053] 제N 다중 태스크 유닛(30N)은 제2 다중 태스크 유닛(32)의 출력측의 마지막에 연결되고 잔차 디테일 학습 경로(RDL)를 통해 현재 타겟 프레임(I1)을 입력받고 제N 디테일 피처맵 네트워크와 제N 디블러링 네트워크를 통해 현재 타겟 프레임의 보정 프레임(최종 보정 프레임)을 출력한다. 비디오 화질개선 장치는 최종 보정 프레임을 통해 계산된 목적함수를 이용해 비디오 화질개선 모델의 기계학습을 수행한다.

[0054] 여기에서, 제N 다중 태스크 유닛(30N)에서 생성된 이전 타겟 프레임(I2)의 제N 화질개선 피처맵(N^{th} deblurred feature map)(30a)은 제1 다중 태스크 유닛(31)과 제2 다중 태스크 유닛(32)를 포함한 다른 모든 다중 태크스 유닛들에 제공된다.

[0055] 또한, 비디오 화질개선 장치(100)는 현재 타겟 프레임의 최종 보정 프레임과 선명한 교사 프레임의 평균 제곱근 편차를 구하여 목적함수를 최소화하는 최적화 유닛을 더 포함할 수 있다. 최소화 유닛은 이전 타겟 프레임의 교사 프레임과 현재 타겟 프레임의 교사 프레임으로부터 생성한 교사 옵티컬플로우를 이용하여 제1 다중 태스크 유닛(31) 및 제2 다중 태스크 유닛(32)의 정합 네트워크들에서 각각 생성한 상관관계 매트릭스(correlation matrix)의 각 픽셀별 크로스 엔트로피(cross entropy)를 구하여 목적함수를 추가로 최소화할 수 있다.

[0056] 손실함수에 대응하는 목적함수의 최소화된 값은 기존의 옵티마이저(Optimizer) 등과 동일하거나 유사한 구성이나 기능을 갖춘 최적화 유닛을 통해 컨볼루션 레이어의 가중치를 업데이트하는데 이용될 수 있다.

[0057] 본 실시예에 의하면, 다중 태스크 유닛을 컨볼루션 신경망에 기반한 디테일 피처맵 네트워크, 디블러링 네트워크, 정합 네트워크로 구성하고, 디블러링 네트워크와 정합 네트워크가 디테일 피처맵 네트워크에서 추출된 디테일 피처맵을 공유함으로써 다중 태스크 유닛들에서 각각 화질 개선과 정합을 효과적으로 수행하여 학습할 수 있다.

[0058] 그 경우, 디블러링 네트워크는 디블러링 네트워크의 출력과 현재 타겟 프레임을 합하는 스킵 커넥션을 통해 디테일 피처맵 네트워크를 통해 생성된 디테일 피처맵이 입력 프레임 중 하나인 현재 타겟 프레임의 디테일 정보를 포함하도록 유도 학습될 수 있다.

[0059] 그리고, 정합 네트워크는 미리 계산된 구조 정보를 포함한 구조 피처맵을 각 다중 태스크 유닛에 제공하도록 설계함으로서 디테일 피처맵 네트워크를 통해 생성된 디테일 피처맵이 구조 정보가 아닌 디테일 정보만 포함하도록 유도 학습될 수 있다. 이를 위해, 정합 네트워크는 한 개의 컨볼루션 레이어를 통해 미리 계산한 입력 프레임의 구조 정보를 포함하는 구조 피처맵을 디테일 피처맵 네트워크가 생성한 디테일 피처맵에 합하고, 한 개의 컨볼루션 레이어로 구성된 모션 레이어를 통해 앞서 합해진 피처맵을 합성하도록 설계될 수 있다.

[0060] 도 2는 도 1의 비디오 화질개선 장치의 전체 네트워크 아키텍처를 나타낸 블록도로서, 구조-디테일 분리 기반의 점진적 디블러링-정합 학습을 위한 디블러링 네트워크 아키텍처를 나타낸다.

[0061] 도 2를 참조하면, 본 실시예에 따른 비디오 화질개선 장치는, 디블러링을 위한 메인 네트워크로서 N개의 다중 태스크 유닛들이 쌓인 구조를 가진다. 각 다중 태스크 유닛은 기본적으로 디테일 피처맵 네트워크(detail feature map network) $\mathcal{F}^n$, 디블러링 네트워크(deblurring network) $\mathcal{D}^n$ 그리고 정합 네트워크(motion compensation network) $\mathcal{M}^n$로 구성된다(도 3 참조). 여기서, n은 1에서 N까지의 임의의 자연수이다.

[0062] 여기서 메인 네트워크는 구조-디테일 분리 기반의 학습을 위해 현재 타겟 프레임 I_t^b와 디블러링 네트워크 $\mathcal{D}^N$를 스킵 커넥션으로 연결시켜 잔차 디테일 학습(residual detail learning)을 유도하고 디블러링 네트워크를 통과하는 보정 프레임의 피처맵들이 디테일 정보를 포함하도록 한다(메인 줄기). 이 메인 줄기는 학습시에만 동작한다.

[0063] 또한, 비디오 화질개선 장치의 네트워크 시작 부분에서 미리 뽑은 구조 피처맵(structural feature map) $\hat{f}_t$는 각 다중 태스크 유닛의 정합 네트워크에 합류되어(제1 보조줄기) 메인 줄기의 디테일 피처맵에 구조 정보를 추가하여 보다 더 정확한 정합을 가능케 한다.

[0064] 좀더 구체적으로, 비디오 화질개선 장치는 t번째 모션블러 프레임인 현재 타겟 프레임 I_t^b, t-1번째 모션블러 프

레임인 이전 타겟 프레임 I_{t-1}^{b}, 디블러링된 t-1번째 프레임 I_{t-1}^{r}을 입력으로 받는다.

[0065] 비디오 화질개선 장치는 입력 프레임들이 첫 번째 다중 태스크 유닛에 들어가기 전 프레임들을 피처맵으로 변환한다.

[0066] 현재 타겟 프레임 I_{t}^{b}는 하나의 컨볼루션 레이어(convolution layer, 제1 컨볼루션 레이어)를 거쳐 구조 정보를 포함하는 구조 피처맵 $\acute{f}_{t}$ 로 변환된다. 구조 피처맵 $\acute{f}_{t}$의 경우, 입력 프레임 I_{t}^{b}이 하나의 제1 컨볼루션 레이어를 거치기 때문에 프레임의 디테일 정보보다 구조 정보를 더 포함하게 된다.

[0067] 이전 타겟 프레임 I_{t-1}^{b}과 이전 타켓 프레임의 보정 프레임 I_{t-1}^{r}은 채널 차원으로 연결(concatenate)된 뒤 다른 하나의 컨볼루션 레이어(제2 컨볼루션 레이어)를 거쳐 피처맵으로 변환된다. 변환된 피처맵은 특징 공간으로서 구조 피처맵과 채널 차원으로 연결돼 첫 번째 다중 태스크 유닛(제1 다중 태스크 유닛)의 입력으로 들어간다.

[0068] 도 3은 도 2의 비디오 화질개선 장치의 다중 태스크 유닛의 기본 구성을 나타낸 블록도이다.

[0069] 도 3을 참조하면, 본 실시예에 따른 비디오 화질개선 장치의 다중 태스크 유닛은 기본적으로 디테일 피처맵 네트워크 $\mathcal{F}^{n}$, 디블러링 네트워크 $\mathcal{D}^{n}$ 및 정합 네트워크 $\mathcal{M}^{n}$를 구비한다.

[0070] 여기서, 다중 태스크 유닛은 제1 다중 태스크 유닛, 제2 다중 태스크 유닛 및 제N 다중 태스크 유닛 중 어느 하나에 대응될 수 있다. 제2 다중 태스크 유닛의 입력들 중 하나는 제1 다중 태스크 유닛의 출력측에 연결되고, 제N 다중 태스크 유닛의 입력들 중 하나는 제2 다중 태스크 유닛의 출력측에 연결되거나 제2 다중 태스크 유닛과 제N 다중 태스크 유닛 사이에 배치되는 하나 이상의 다른 다중 태스크 유닛들 중 마지막 다중 태스크 유닛의 출력측에 연결된다. N은 3 이상의 임의의 자연수이다. 다만, 다중 태스크 유닛이 제N 다중 태스크 유닛인 경우, 다중 태스크 유닛은 정합 네트워크 $\mathcal{M}^{n}$를 생략할 수 있다.

[0071] 디테일 피처맵 네트워크는 구조-디테일 분리 기반의 학습에 있어 디테일 정보를 포함하는 피처맵이 디블러링 네트워크에 걸쳐 유지되도록 하는 메인 줄기를 유지하는 뼈대와 같은 역할을 한다. 디테일 피처맵 네트워크는 인코더 디코더 구조의 컨볼루션 레이어들로 구성되며 n-1개의 이전 프레임들의 다중 태스크 유닛의 출력을 입력으로 받아 디테일 피처맵 f_{t}^{n}을 생성한다.

[0072] 디테일 피처맵 네트워크 $\mathcal{F}^{n}$는 디블러링 네트워크 $\mathcal{D}^{n}$과 정합 네트워크 $\mathcal{M}^{n}$에 의해 공유된다. 디블러링 네트워크 $\mathcal{D}^{n}$는 디테일 피처맵 f_{t}^{n}을 한 개의 컨볼루션 레이어로 이루어진 디블러 레이어(deblur layer)를 통해 디블러링에 필요한 디테일 정보를 포함하는 잔차(residual) 프레임 $I_{t}^{res,n}$으로 변환하고 잔차 프레임 $I_{t}^{res,n}$을 스킵 커넥션(skip connection)으로 연결된 현재 타겟 프레임 I_{t}^{b}과 더해 디블러된(deblurred) 프레임 $I_{t}^{r,n}$(이하 제1 보정 프레임, 제2 보정 프레임 또는 보정 프레임이라 한다)을 생성한다. 스킵 커넥션은 학습 시 더 정확한 디블러링을 위해, 잔차 프레임 $I_{t}^{res,n}$이 더 정확한 디테일 정보를 포함하도록 디테일 피처맵 네트워크 $\mathcal{F}^{n}$으로 하여금 지속적으로 디테일 피처맵 네트워크에 디테일 정보를 포함하도록 유도한다.

[0073] 한편, 정합을 수행하는 정합 네트워크 $\mathcal{M}^{n}$은 먼저 앞서 뽑은 구조 피처맵 $\acute{f}_{t}$를 f_{t}^{n}에 더해 구조 정보를 추가하고(structure injection) 이를 한 개의 컨볼루션 레이어로 이루어진 모션 레이어(motion layer)를 통해 합성시켜 t번째 프레임의 n번째 다중 태스크 유닛의 디테일-구조 피처맵 $\hat{f}_{t}^{n}$을 생성한다.

[0074] 그 다음, 정합 네트워크에 구비된 모션 보상 모듈(motion compensation module)은 $\hat{f}_{t-1}^{n}$(t-1번째 프레임의 n번째 다중 태스크 유닛의 디테일-구조 피처맵)과 $\hat{f}_{t}^{n}$ 사이의 상관관계(correlation)를 기반으로 픽셀별 매칭을 수행하고, 이 매칭을 기반으로 f_{t-1}^{N}(t-1번째 프레임의 N번째 다중 태스크 유닛의 디테일 피처맵)을 정합하여 화질개선 피처맵 $\tilde{f}_{t-1}^{N,n}$을 생성한다.

[0075] 또한 모션 보상 모듈은 t-1번째 프레임(이전 프레임 또는 이전 타겟 프레임)의 n번째 다중 태스크 유닛의 디테

일-구조 피처맵 $\hat{f}_{t-1}^{n}$과 t번째 프레임(현재 프레임 또는 현재 타겟 프레임)의 n번째 다중 태스크 유닛의 디테일-구조 피처맵 $\hat{f}_{t}^{n}$ 사이의 픽셀 별 상관관계(ccorrelation)를 구하고, 각 픽셀마다 상관관계 값이 가장 큰 매칭 픽셀의 좌표의 옵셋(offset)을 기반으로 옵티컬플로우(optical flow)를 추출한다.

[0076] 마지막으로 정합 네트워크는 f_{t}^{n} 과 $\tilde{f}_{t-1}^{N,n}$ 을 채널 차원으로 연결하여 n+1번째 다중 태스크 유닛의 입력으로 전달한다.

[0077] 본 실시예의 비디오 화질개선 장치에서, 1번째 다중 태스크 유닛의 경우 구조 피처맵 $\hat{f}_{t}$와 이전 타겟 프레임 I_{t-1}^{b} 및 디블러링된 t-1번째 프레임인 이전 타겟 프레임의 보정 프레임 I_{t-1}^{r}이 채널 차원으로 연결된 피처맵을 입력으로 받으며, N번째 다중 태스크 유닛의 경우 정합 네트워크는 생략되고, 리블러링 네트워크 $\mathcal{D}^{N}$가 최종 디블러링 결과인 보정 프레임 I_{t}^{r}을 출력한다(도 2 참조).

[0078] 본 실시예에 의하면, 구조-디테일 분리 기반으로 설계된 다중 태스크 유닛을 여러 개 쌓아 설계한 비디오 화질개선 모델이 점진적으로 화질 개선과 정합을 수행하여 현재 타겟 프레임(I_{t}^{b})으로부터 보정 프레임(I_{t}^{r})을 효과적으로 생성하도록 한다. 즉, 구조-디테일 분리 기반으로 설계된 다중 태스크 유닛의 디테일 피처맵 네트워크에서 출력된 디테일 피처맵과 이전 타겟 프레임의 마지막 다중 태스크 유닛에서 생성된 디테일 피처맵이 정합 네트워크를 통해 정합된 잔차 프레임의 피처맵이 채널 차원으로 연결(concatenate)되어 다음의 다중 태스크 유닛의 입력되도록 함으로서 복수의 다중 태스크 유닛들에 의한 화질 개선과 정합을 점진적으로 효과적으로 수행할 수 있다. 여기에서, 마지막 다중 태스크 유닛은 정합 네트워크를 생략하고 디블러링 네트워크의 스킵 커넥션의 통해 최종 보정프레임(I_{t}^{r})을 출력한다.

[0079] 전술한 복수의 다중 태스크 유닛들을 쌓은 컨볼루션 신경망의 디블러링 학습을 위해 사용한 목적함수(L_{deblur})는 다음의 [수학식 1]과 같다.

수학식 1

$$L_{deblur} = \sum_{n=1}^{N} \lambda_n MSE(I_t^{r,n}, I_t^{GT})$$

[0080]

[0081] [수학식 1]에서 MSE는 평균 제곱 오차(mean squared error)를 의미한다. 또한 [수학식 1]에서 $n \in \{1,\ldots,N-1\}$ 일 때 $\lambda_n = 0.1$을 사용하고, $n = N$일 때 $\lambda_n = 1$을 사용할 수 있다. I_t^{GT}는 t번째 선명한 교사(ground-truth) 프레임을 의미한다.

[0082] 또한, 정합 학습을 위해 사용한 목적함수(L_{motion})는 다음의 [수학식 2]와 같다.

수학식 2

$$L_{motion} = \sum_{n=1}^{N-1} \sum_{x=1}^{M} \sum_{i=1}^{D^2} C_t^{GT}(x,i) log(softmax(C_t^n(x,i)))$$

[0083]

[0084] [수학식 2]에서 C_t^{GT}는 I_{t-1}^{GT}와 I_t^{GT}의 상관관계 매트릭스를, C_t^n은 $\hat{f}_{t-1}^{n}$와 $\hat{f}_{t}^{n}$의 상관관계 매트릭스를 의미한다. x는 상관관계 매트릭스의 픽셀 별 위치, M은 매트릭스의 마지막 위치, i는 매트릭스의 채널 위치, D^2는 매트릭스의 전체 채널을 의미한다. 그리고 softmax()는 네트워크의 활성화 함수를 나타낸다.

[0085] 최종적으로, 디블러링을 위한 전체 네트워크의 학습을 위한 목적함수(L_{total})는 다음의 [수학식 3]과 같다.

수학식 3

[0086] $$L_{total} = L_{deblur} + \alpha L_{motion}$$

[0087] [수학식 3]에서 목적함수 가중치(α)는 0.1을 사용할 수 있다.

[0088] 또한, 추가적인 디블러링 네트워크 학습을 위해 핸드 헬드 카메라로 캡쳐한 비디오로서 대부분의 프레임에서 서로 다른 블러를 포함하는 디블러링 데이터셋을 사용할 수 있다.

[0089] 예를 들어, 데이터셋은 71쌍의 모션블러 비디오와 선명한 비디오(교사 비디오)를 포함하고 총 6,708쌍의 1280×720 해상도의 모션블러 프레임과 선명한 프레임으로 구성할 수 있다. 이들 중, 10쌍의 비디오를 테스트셋으로 사용하고 나머지는 네트워크 학습을 위해 사용할 수 있다.

[0090] 학습 시, 한 배치마다 13장의 연속된 비디오 프레임을 모션블러 비디오와 선명한 비디오에서 무작위로 추출할 수 있다. 그 다음, 임의의 위치를 정해 256×256 해상도의 패치를 추출한 비디오 프레임들에서 크랍(crop)할 수 있다.

[0091] 크랍된 모션블러 프레임을 $I_t^b; t \in [1, \ldots, 13]$, 선명한 교사 프레임을 $I_t^{GT}; t \in [1, \ldots, 13]$ 표기할 수 있다(수학식 1 및 수학식 2 참조). 배치 크기는 8로 설정하고 $\beta_1 = 0.9$, $\beta_2 = 0.999$로 설정한 아담 옵티마이저(Adam Optimizer)를 사용할 수 있다. 네트워크 학습 시, 첫 400,000 반복(iteration)에는 1.0×10^{-4}의 학습율(learning rate)을 사용할 수 있고, 나머지 200,000 반복에는 2.5×10^{-5}의 학습율을 사용할 수 있다.

[0092] 도 4는 도 1의 비디오 화질개선 장치의 변형예를 설명하기 위한 블록도이다.

[0093] 도 4를 참조하면, 본 실시예에 따른 비디오 화질개선 장치는 프로세서(200)와 메모리(300)를 포함한다. 프로세서(200)에는 도 1이나 도 2 및 도 3을 참조하여 앞서 설명한 비디오 화질개선 장치(100)가 탑재된다. 이 경우, 비디오 화질개선 장치(100)는 적어도 하나 이상의 소프트웨어 모듈 형태로 프로세서(200)에 탑재될 수 있다. 이와 같이, 본 실시예에 따른 비디오 화질개선 장치는 넓은 의미에서 프로세서(200)를 포함할 수 있다.

[0094] 또한, 비디오 화질개선 장치는 머신러닝 기반의 비디오 화질개선 장치의 각 구성요소에 대응하는 기능을 수행하는 복수의 소프트웨어 모듈들이나 프로그램을 저장한 메모리 등의 컴퓨터로 읽을 수 있는 기록매체의 형태로 구현될 수 있다.

[0095] 전술한 경우, 메모리에 저장되는 머신러닝 기반의 비디오 화질개선 방법을 구현하는 프로그램은, 입력 프레임 중 하나인 현재 타겟 프레임으로부터 제1 컨볼루션 레이어에 의해 변환된 구조 피처맵(structure feature map)을 복수의 다중 태스크 유닛들 중 제1 다중 태스크 유닛과 제1 다중 태스크 유닛의 출력측에 연결되는 제2 다중 태스크 유닛에 입력하고; 입력 프레임 중 다른 하나인 이전 타겟 프레임 및 입력 프레임 중 또 다른 하나인 이전 프레임의 보정 프레임을 채널 차원으로 연결한 것으로부터 제2 컨볼루션 레이어에 의해 변환된 특징 공간에 구조 피처맵을 추가한 메인 입력을 제1 다중 태스크 유닛에 입력하고; 그리고 제2 다중 태스크 유닛의 출력측의 마지막에 연결되는 제N 다중 태스크 유닛에 현재 타겟 프레임을 입력하기 위한 일련의 과정들을 포함하도록 구현될 수 있다.

[0096] 또한, 상기의 프로그램은, 제N 다중 태스크 유닛에서 생성된 이전 타겟 프레임의 제N 화질개선 피처맵(N^{th} deblurred feature map)을 제1 다중 태스크 유닛과 제2 다중 태스크 유닛 등 다른 모든 다중 태스크 유닛에 제공하는 과정을 포함하도록 구현될 수 있다.

[0097] 또한, 상기의 프로그램은, 제1 다중 태스크 유닛의 제1 디테일 피처맵 네트워크에 의해 상기의 메인 입력을 받고 제1 디테일 피처맵을 출력하고; 제1 다중 태스크 유닛의 정합 네트워크의 모션 레이어에 의해 제1 디테일 피처맵에 구조 피처맵을 추가한 구조-주입 피처맵(structure-injected feature map)을 현재 프레임 피처맵으로 변환하고; 정합 네트워크의 모션 보상 모듈에 의해 현재 프레임 피처맵과 이전 타겟 프레임의 이전 프레임 피처맵과의 모션을 추정하고 추정된 모션에 기초하여 이전 타겟 프레임의 화질개선 피처맵을 현재 타겟 프레임에 대해 정렬하고; 정합 네트워크의 연결 네트워크를 통해 이전 타겟 프레임의 정렬된 화질개선 피처맵과 제1 디테일 피처맵을 채널 차원으로 연결한 제1 화질개선 피처맵을 제2 다중 태스크 유닛의 입력측으로 출력하는 일련의 과정

을 포함하도록 구현될 수 있다.

[0098] 또한, 상기의 프로그램은, 다중 태스크 유닛들을 포함한 네트워크의 훈련 시, 제1 다중 태스크 유닛의 제1 디블러링 네트워크의 디블러 레이어에 의해 제1 디테일 피처맵을 출력 잔차 이미지로 변환하고 디블러 레이어의 출력측에 연결되는 스킵 커넥션을 통해 현재 타겟 프레임을 더한 제1 보정 프레임을 출력하는 과정을 포함하도록 구현될 수 있다. 제1 보정 프레임은 제1 다중 태스크 유닛의 가중치를 업데이트하는데 이용될 수 있다.

[0099] 또한, 상기의 프로그램은 제N 다중 태스크 유닛에서 출력되는 현재 타겟 프레임에 대한 화질개선 프레임인 보정 프레임을 통해 계산된 손실함수나 목적함수를 이용해 비디오 화질개선 모델의 기계학습을 수행하는 과정을 포함하도록 구현될 수 있다.

[0100] 다시 말해서, 전술한 실시예들을 통해 설명한 머신러닝 기반의 비디오 화질개선 방법은 다양한 컴퓨터 수단을 통해 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 컴퓨터 판독 가능 매체에 기록되는 프로그램 명령은 본 발명을 위해 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용가능한 것일 수 있다.

[0101] 컴퓨터 판독 가능 매체의 예에는 롬(rom), 램(ram), 플래시 메모리(flash memory) 등과 같이 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러(compiler)에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터(interpreter) 등을 사용해서 컴퓨터에 의해 실행될 수 있는 고급 언어 코드를 포함한다. 상술한 하드웨어 장치는 본 실시예에 따른 머신러닝 기반의 비디오 화질개선 방법의 일련의 동작을 수행하기 위해 적어도 하나의 소프트웨어 모듈로 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0102] 전술한 실시예에 따른 비디오 화질개선 장치의 비디오 디블러링 결과를 각 모듈별 효과로써 나타내면 [표 2]와 같다.

표 2

[0103]

Components				Motion compensation		Deblurring	
L_{deblur}	$\mathcal{M}$	L_{motion}	motion layer	PSNR	SSIM	PSNR	SSIM
✓				-	-	29.79	0.915
✓	✓			24.59	0.851	29.93	0.916
✓	✓	✓		**26.10**	**0.887**	29.72	0.913
✓	✓	✓	✓	**26.10**	**0.886**	**30.05**	**0.918**

[0104] [표 2]는 본 실시예의 비디오 화질개선 방법의 모듈별 정략적 효과를 보여준다. 정량적 수치들은 2개의 다중 태스크 유닛들이 쌓인 네트워크들(N=2)에 의한 결과이다. [표 2]에서 볼 수 있듯이, 정합 네트워크(M)를 쓰는 것만으로도 디블러링 성능의 증가가 있음을 알 수 있다([표 2]의 왼쪽 2번째 열).

[0105] 추가적으로, 학습 시 정합 성능은 증가하지만 디블러링 성능은 하락한다([표 2]의 3번째 열). 하지만, [표 2]의 왼쪽 마지막 행(4행)에서 모션 레이어(motion layer)(도 3의 정합 네트워크 M^n 참조)를 사용함으로써 구조-디테일 분리 기반의 학습이 완성되어 디블러링 성능이 크게 증가함을 확인할 수 있다.

[0106] 다음, 본 실시예에 따른 비디오 화질개선 장치의 점진적 디블러링-정합에 대한 정량적 성능 평가 결과를 나타내면 도 5 및 [표 3]과 같다.

[0107] 도 5는 도 1의 비디오 화질개선 장치에 의한 점진적 디블러링-정합에 대한 정량적, 정성적 결과를 설명하기 위한 도면이다. 도 5에서 (a)는 현재 타겟 프레임에 대응하는 입력(Input), (b)는 제1 다중 태스크 유닛의 제1 디블러링 네트워크(D^1)에서의 출력 결과, (c)는 제2 다중 태스크 유닛의 제2 디블러링 네트워크(D^2)에서의 출력 결과, (d)는 제3 다중 태스크 유닛의 제3 디블러링 네트워크(D^3)에서의 출력 결과, (e)는 제4 다중 태스크 유닛의 제4 디블러링 네트워크(D^4)에서의 출력 결과, (f)는 제5 다중 태스크 유닛의 제5 디블러링 네트워크(D^5)에서의 출력 결과를 각각 나타낸다.

표 3

[0108]

	$(-, \mathcal{D}^1)$	$(\mathcal{M}^1, \mathcal{D}^2)$	$(\mathcal{M}^2, \mathcal{D}^3)$	$(\mathcal{M}^3, \mathcal{D}^4)$	$(\mathcal{M}^4, \mathcal{D}^5)$
Motion compensation	-	24.29	24.61	24.87	24.88
Deblurring	24.54	25.82	26.65	27.50	27.77

[0109] [표 3] 및 도 5에 나타낸 바와 같이, 정합 네트워크(M^1, M^2, M^3, M^4) 및 디블러링 네트워크(D^1, D^2, D^3, D^4, D^5) 중 적어도 어느 하나를 각각 구비한 5개의 다중 태크스 유닛들을 쌓은 비디오 화질개선 네트워크(N=5)에 대한 성능 평가 결과, 본 실시예에 따른 비디오 화질개선 장치에서 정합 성능과 디블러링 성능이 점진적으로 증가함을 알 수 있다.

[0110] 도 6은 도 1의 비디오 화질개선 장치와 비교예들에 대한 정량적, 정성적 결과를 비교하여 나타낸 도면이다.

[0111] 도 6 및 [표 4]를 참조하면, 10개의 다중 태스크 유닛들을 쌓은 본 실시예의 비디오 화질개선 네트워크(N=10)에 대한 성능과 비교예들의 성능을 비교한 결과, 4개의 다중 태스크 유닛들을 쌓은 본 실시예((e)Ours (4-stack))와 10개의 다중 태스크 유닛들을 쌓은 본 실시예((f)Ours (10-stack))의 연산량 및 크기가 비교예들((b)IFIRNN, (c)STFAN, (d)ESTRNN)보다 작음에도 불구하고 입력(a)에 대해 가장 높은 디블러링 성능을 보임을 알 수 있다.

표 4

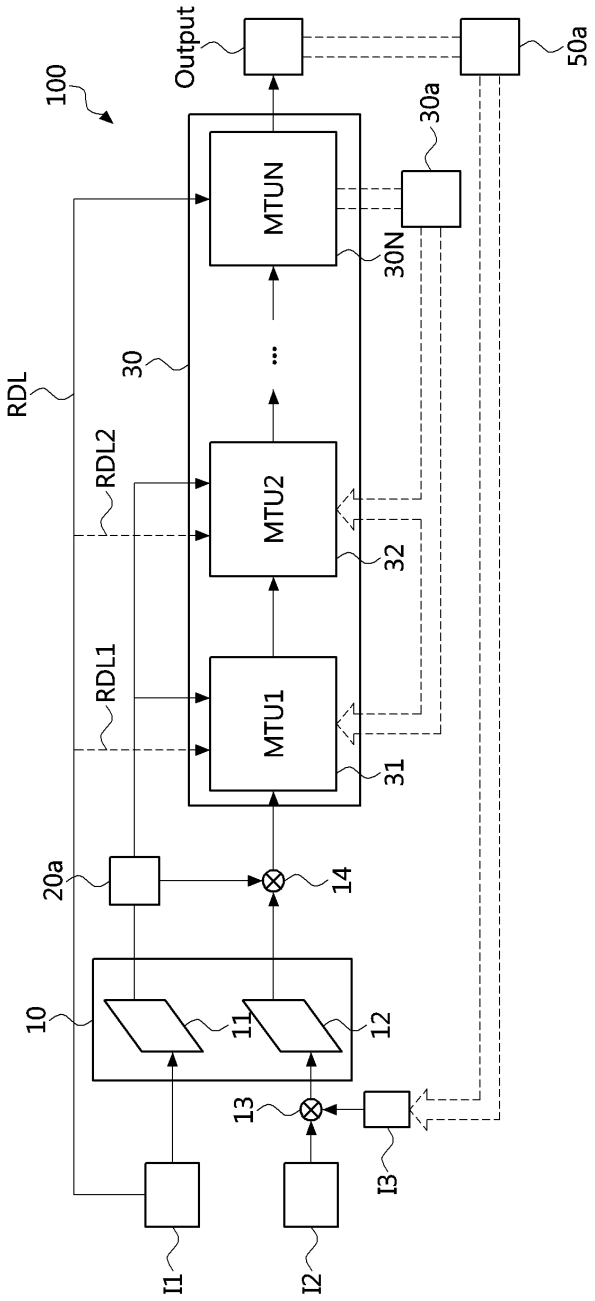
[0112]

	Nah *et al.*	Tao *et al.*	DVD	IFIRNN	STFAN	ESTRNN	Ours (n-stack)		
							2	4	10
PSNR (dB)	29.59	30.24	30.05	30.78	31.24	31.02	30.54	31.07	**31.56**
Params (M)	75.92	3.76	15.31	1.64	5.37	6.71	**0.92**	1.89	4.78
Time (ms)	1790	560	581	54	145	63	**9**	14	31

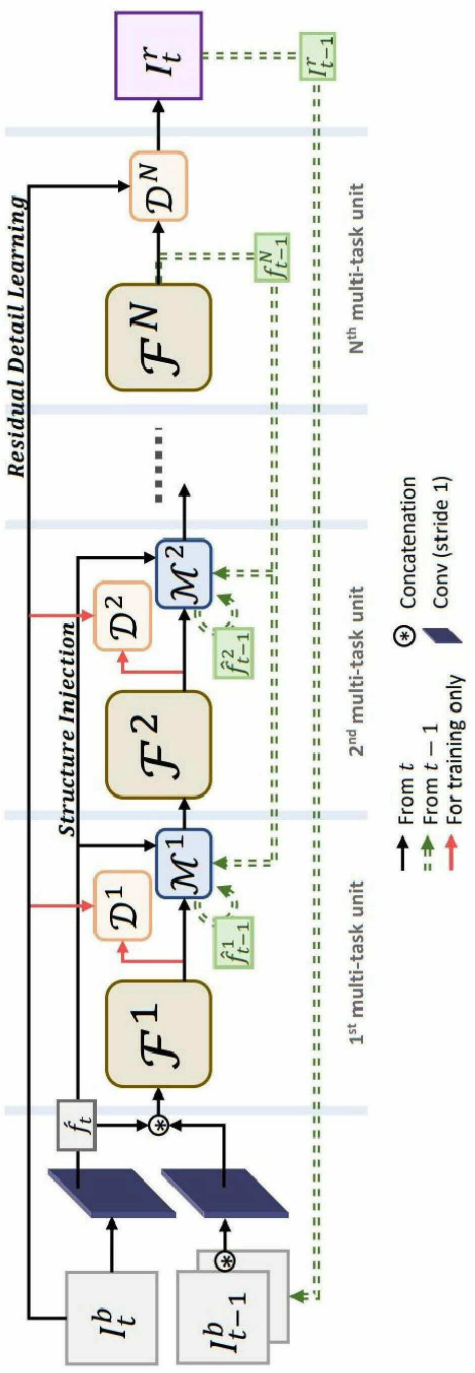
[0113] 이상 실시예를 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 청구범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

도면

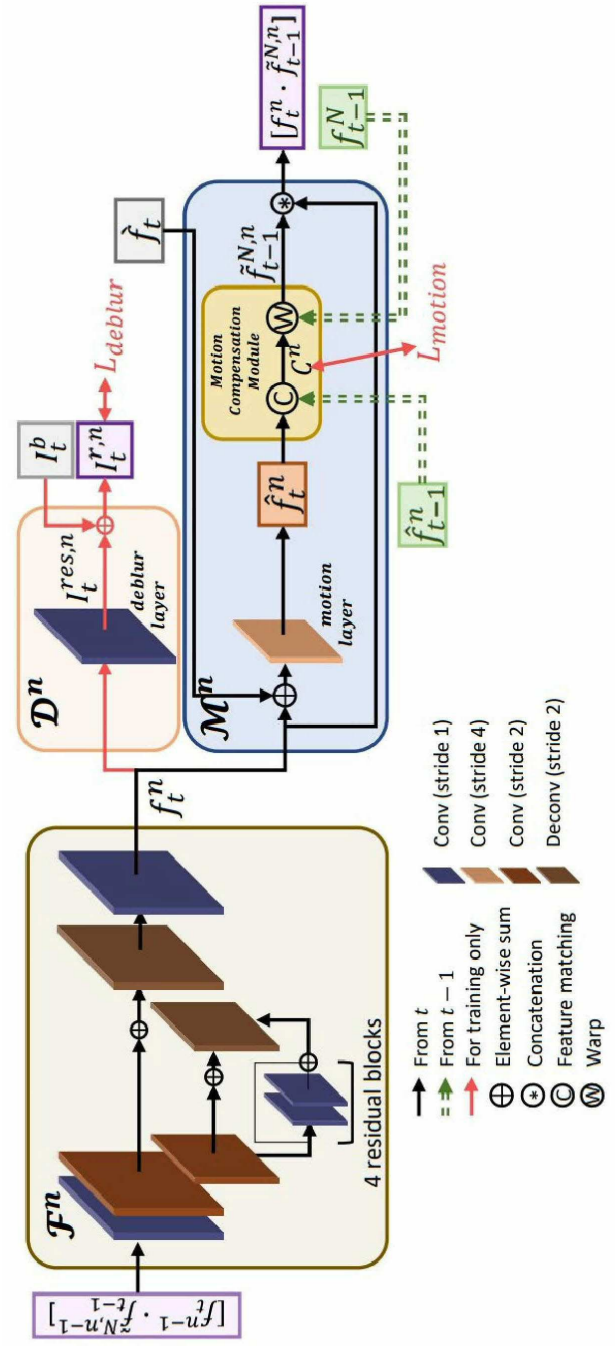
도면1


100
Output
50a
30a
MTUN
30N
30
RDL
RDL2
MTU2
32
RDL1
MTU1
31
20a
14
10
11
12
13
I3
I1
I2

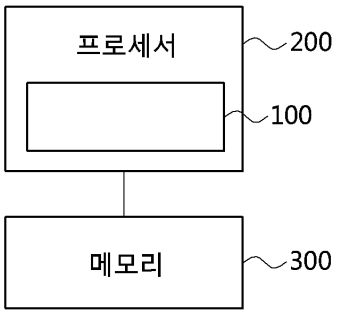
도면2


Residual Detail Learning
Structure Injection
1st multi-task unit
2nd multi-task unit
Nth multi-task unit
Concatenation
Conv (stride 1)
From t
From t − 1
For training only

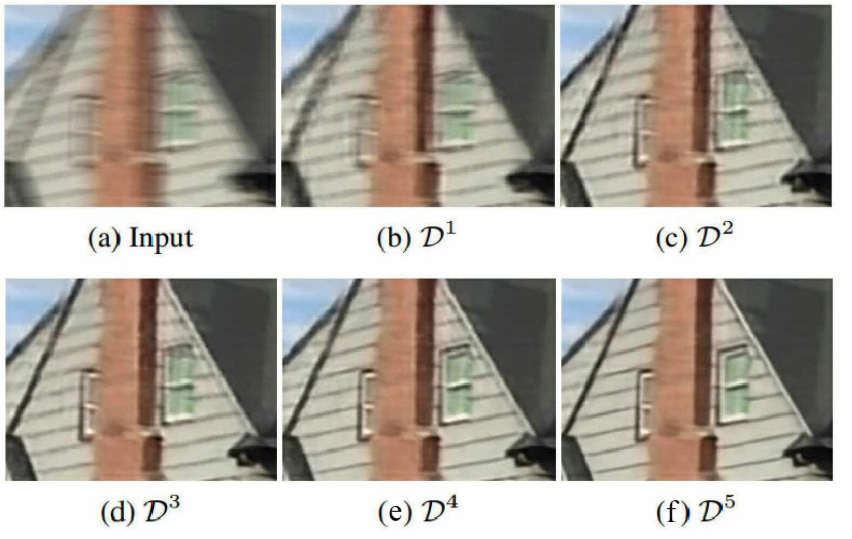
도면3


$[f_t^n \cdot \tilde{f}_{t-1}^{N,n}]$
f_{t-1}^N
$\hat{f}_t$
L_{deblur}
L_{motion}
$\tilde{f}_{t-1}^{N,n}$
Motion Compensation Module
C^n
I_t^b
$I_t^{r,n}$
$\hat{f}_t^n$
$\hat{f}_{t-1}^n$
$I_t^{res,n}$
deblur layer
motion layer
$\mathcal{D}^n$
$\mathcal{M}^n$
f_t^n
$\mathcal{F}^n$
4 residual blocks
$[f_t^{n-1} \cdot \tilde{f}_{t-1}^{N,n-1}]$
Conv (stride 1)
Conv (stride 4)
Conv (stride 2)
Deconv (stride 2)
From t
From $t-1$
For training only
Element-wise sum
Concatenation
Feature matching
Warp

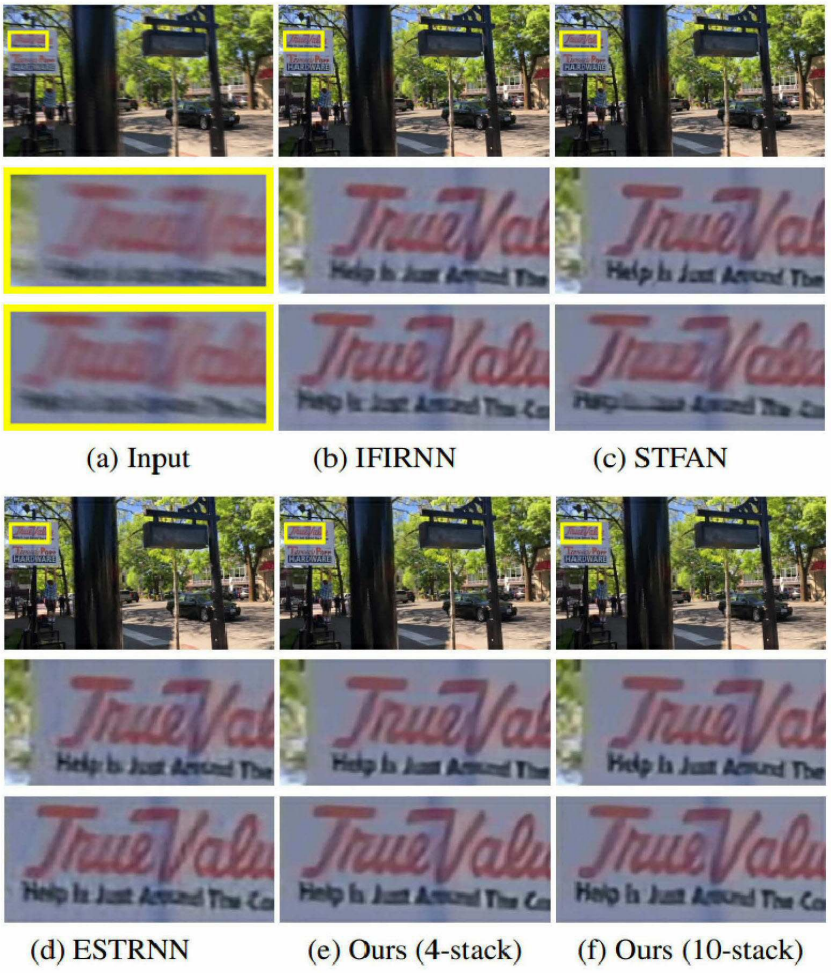
도면4


프로세서
200
100
메모리
300

도면5


(a) Input
(b) $\mathcal{D}^1$
(c) $\mathcal{D}^2$
(d) $\mathcal{D}^3$
(e) $\mathcal{D}^4$
(f) $\mathcal{D}^5$

도면6


(a) Input
(b) IFIRNN
(c) STFAN
(d) ESTRNN
(e) Ours (4-stack)
(f) Ours (10-stack)