일반논문 (Regular Paper)
방송공학회논문지 제25권 제3호, 2020년 5월 (JBE Vol. 25, No. 3, May 2020)
https://doi.org/10.5909/JBE.2020.25.3.428
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

Triplet CNN과 학습 데이터 합성 기반 비디오 안정화기 연구

양 병 호[a], 이 명 진[b)‡]

Study on the Video Stabilizer based on a Triplet CNN and Training Dataset Synthesis

Byongho Yang[a] and Myeong-jin Lee[b)‡]

요 약

영상 내 흔들림은 비디오의 가시성을 떨어뜨리고 영상처리나 영상압축의 효율을 저하시킨다. 최근 디지털 영상처리 분야에 딥러닝이 본격 적용되고 있으나, 비디오 안정화 분야에 딥러닝 적용은 아직 초기 단계이다. 본 논문에서는 Wobbling 왜곡 경감을 위한 triplet 형태의 CNN 기반 비디오 안정화기 구조를 제안하고, 비디오 안정화기 학습을 위한 학습데이터 합성 방법을 제안한다. 제안한 CNN 기반 비디오 안정화기는 기존 딥러닝 기반 비디오 안정화기와 비교되었으며, Wobbling 왜곡은 감소하고 더 안정적인 학습이 이루어지는 결과를 얻었다.

Abstract

The jitter in the digital videos lowers the visibility and degrades the efficiency of image processing and image compressing. In this paper, we propose a video stabilizer architecture based on triplet CNN and a method of synthesizing training datasets based on video synthesis. Compared with a conventional deep-learning video stabilization method, the proposed video stabilizer can reduce wobbling distortion.

Keywords : video stabilization, convolutional neural network, wobbling distortion

a) 서보산전(Servo Industrial Systems Co., Ltd.)
b) 한국항공대학교 항공전자정보공학부(School of Electronics and Information Engineering, Korea Aerospace University)
‡ Corresponding Author : 이명진(Myeong-jin Lee)
E-mail: artistic@kau.ac.kr
Tel: +82-2-300-0421
ORCID: https://orcid.org/0000-0002-3136-2819

※ 이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업(2018R1D1A1B07050603)과 경기도 재원의 경기도 지역협력연구센터 사업(GRRC항공2019-B01)의 일환으로 수행된 연구결과임. (This work was partly supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No. NRF-2018R1D1A1B07050603), and by the GRRC program of Gyeonggi province (No. GRRC-KAU2019-B01)).
· Manuscript received March 16, 2020; Revised April 13, 2020; Accepted April 13, 2020.

Ⅰ. 서론

디지털 이미징 기기의 급격한 증가와 함께 아마추어들이 촬영하는 디지털 영상 또한 가파르게 증가하고 있다. 그런데 개인 저작 영상은 촬영 환경의 열악함과 촬영자의 미숙함 등으로 인해 가시성이 많이 떨어지는데, 손 떨림에 의해 유발된 영상의 흔들림이 가시성을 저하시키는 주요 요인으로 지적되고 있다[1]. 또한, 야외 레저활동 과정에서 액션 캠, 드론, 로봇 등으로 촬영한 영상은 필연적으로 흔들림이 발생하므로 이들 영상을 시청하거나 혹은 영상의 활용 시 흔들림 보정이 필요하다[2].

영상의 흔들림을 보정하는 기술은 기계적 보정, 광학적 보정, 전자적 보정, 디지털 보정(digital image stabilization; DIS) 등으로 나누어 볼 수 있다. 기계적 보정은 무게 추, gimbal 등의 기계 장치를 이용한 흔들림 보정 기술이다. 근래의 기계적 보정은 관성센서를 이용한 흔들림 추정과 actuator 구동을 통해 CCD (Charge-Coupled Device) 플랫폼을 움직여 흔들림을 보정한다. 광학적 보정은 관성센서를 통한 흔들림 추정 후 특수 제작 렌즈를 움직여 흔들림을 보정한다[3]. 전자적 보정은 카메라 이미지를 분석하거나 관성센서를 통한 흔들림 추정 후 CCD 투영 이미지를 움직여 흔들림을 보정한다. 마지막으로 DIS는 흔들림 추정과 보정을 디지털 이미지 처리를 통해 수행한다.

스마트폰과 같은 소형/저가 장치에서는 크기와 가격 등의 요소로 인해 디지털 보정을 채택한다. 종래의 DIS는 특징점이나 카메라 궤적 추적에 기반한 방식들인데, 많은 연산량과 영상 보정 성능의 한계가 존재한다[6-8,11]. 이러한 문제를 해결할 수 있는 방법 중 하나가 CNN(Convolution Neural Network) 기반의 딥러닝의 도입이나, 비디오 안정화 분야에 딥러닝 적용은 아직 초기 단계이다.

본 논문에서는 Wobbling 왜곡을 경감시킬 수 있는 triplet 형태의 CNN 기반 비디오 안정화기 구조를 제안한다. 제안한 CNN 기반 비디오 안정화기가 좋은 성능을 보이기 위해서는 양질의 학습데이터가 필요하며, 학습을 위해 별도로 촬영할 필요가 없는 영상합성 기반 학습데이터 확보 방법을 제안한다.

본 논문의 구성은 다음과 같다. 제2절에서는 비디오 안정화를 위한 디지털 영상처리와 딥러닝 기반 기존 연구들에 대해 설명한다. 제3절에서는 triplet 형태의 CNN 기반 비디오 안정화기 구조를 제안한다. 제4절에서는 제안한 구조에 대한 비디오 안정화 실험 결과를 제시하고, 제5절에서 결론을 맺는다.

Ⅱ. 기존의 비디오 안정화 연구

1. 디지털 이미지 처리 기반 비디오 안정화

DIS는 디지털 이미지 처리를 통해 흔들림을 추정하고 흔들림을 보상한다. DIS는 움직임 추정, 움직임 필터링, 이미지 보상의 단계로 구성된다. DIS의 첫 단계인 움직임 추정은 연속으로 입력되는 비디오 프레임을 비교하여 각 특징점들의 지역움직임(Local Motion Vectors, LMV)을 추정한다[5-8]. 추정된 LMV에는 영상내 움직이는 물체, 조명변화 등으로 인해 실제 카메라 움직과는 다른 벡터들이 존재하기 때문에, 추정된 LMV들로부터 Low pass filter나 Gaussian smoothing 등을 이용하여 카메라의 전역 움직임(Global Motion Vector, GMV)를 추정한다[1,11]. Filtering 후 최종 GMV를 추정하기 위해 RANSAC[12] 알고리즘이나 그 변형들이 사용된다. 추정된 GMV로부터 3D 공간상의 perspective motion을 표현하는 homography를 추출할 수 있다[13]. 카메라 이동시 추정된 GMV에는 카메라 이동 벡터와 흔들림 벡터가 중첩되어 있어서 카메라 궤적 추적 방법을 통해 GMV에 카메라 이동 궤적이 배제된 흔들림만 남는다[13]. 흔들림 보상의 마지막 단계는 homography 변환시 발생한 테두리의 검은 영역 제거 작업이며, 이미지의 중심 부분을 검은 영역이 보이지 않을 정도로 확대하는 방법들이 사용된다.

2. 비디오 안정화를 위한 딥러닝 기법

비디오 안정화 분야에서 딥러닝 연구는 최근 시작되었는데, 신경망 여러 개를 병렬로 배치하고 각 신경망 출력 결과를 비교함으로써 입력 이미지에 대한 분류나 비교 문제를 푸는 방법[15,16]과 네트워크에서 이미지의 공간 방향의 다양한 변환을 고려한 모델[17]이 사용되었다.

Zagoruyko은 같은 weight를 공유하는 두 개의 병렬 신경망 출력 결과를 비교함으로써 두 이미지의 일치 여부를 판단하는 Siamese network을 제안하였다[15]. Siamese network은 positive sample은 true, negative sample은 false로 학습함으로써 단순한 이미지 비교만이 아니라 이미지 분류에도 활용할 수 있다. 기존의 딥러닝 모델에서는 대량의 학습데이터가 필요했는데 Siamese network는 소량의 기준 데이터 기반으로 일치여부를 판단하므로 소량의 데이터로 학습이 가능하다.

Siamese network은 feature learning에서 확장성을 제공할 수 있으나 분류 문제에서 기존 딥러닝 모델에 비해 정확도가 높지 않다. 이는 positive sample과 negative sample의 흔들림 정도를 계량화하여 판단할 수 없기 때문인데, Hoffer은 triplet network를 통해 이 문제의 해결방법을 제안하였다[16]. Triplet network는 동일한 weight를 공유하는 세 개의 병렬 신경망에 training sample, positive sample, negative sample을 입력하고 그 출력 사이의 거리를 loss 함수에 반영한다.

CNN 등 기존의 딥러닝 네트워크는 공간 영역에서 병진 또는 회전 변환된 이미지 처리시 제약이 있었다. Jaderberg은 이 문제를 해결하기 위해 네트워크의 중간에 transformer layer를 삽입하고 이 layer를 위한 적절한 gradient 함수를 제공함으로써 별도의 추가 작업 없이 backpropagation이 이루어짐을 보이고, 실제 딥러닝에서 성능이 개선됨을 보였다[17]. Jaderberg에 따르면 transformer layer는 CNN 등의 네트워크의 일부로 삽입하기만 하면 별다른 추가 작업 없이 이미지 변환에 강한 네트워크 구조를 얻을 수 있다.

Wang은 일반 비디오 카메라(unsteady camera)와 안정화 기가 장착된 비디오 카메라(steady camera)를 쌍으로 배치해 다양한 영상을 직접 촬영하여 학습 데이터 부족 문제를 해결했다[18]. Resnet과 Siamese network 기반으로 Stabnet을 설계하고, Stabnet의 출력인 homography를 이용한 Loss 함수를 정의함으로써 비디오 안정화 문제에 딥러닝을 적용하였다. Stabnet 입력은 그림 1과 같이 B_t와 B_{t-1} 두 개의 비디오 프레임 그룹이다. B_t는 unsteady 비디오 프레임 $\tilde{F}_t$와 안정화된 비디오 프레임 $\bar{F}_{t-30}$, $\bar{F}_{t-24}$, $\bar{F}_{t-18}$, $\bar{F}_{t-12}$, $\bar{F}_{t-6}$ 5장으로 구성된다. B_t에 포함된 비디오 프레임들 사이의 흔들림이 감소되도록 네트워크 훈련을 위해, Loss 함수는 안정화된 비디오 프레임과 ground truth (GT) 프레임 사이의 화소 차이와 특징점 간 거리를 포함하고, wobbling 왜곡을 경감시키기 위해 인접한 프레임 간의 화소 왜곡을 포함한다. 또한, 실제 사용을 가정할 때 초기 프레임들은 안정화를 위한 비교 대상이 없는 문제가 있는데 이는 첫 프레임을 시퀀스만큼 복제하여 사용하는 방법을 사용하여 해결하였다.

Ⅲ. 제안하는 비디오 안정화기

본 장에서는 Wobbling 왜곡 억제에 효과적인 딥러닝 네트워크 구조와 소량의 훈련데이터를 이용해 대량의 흔들림

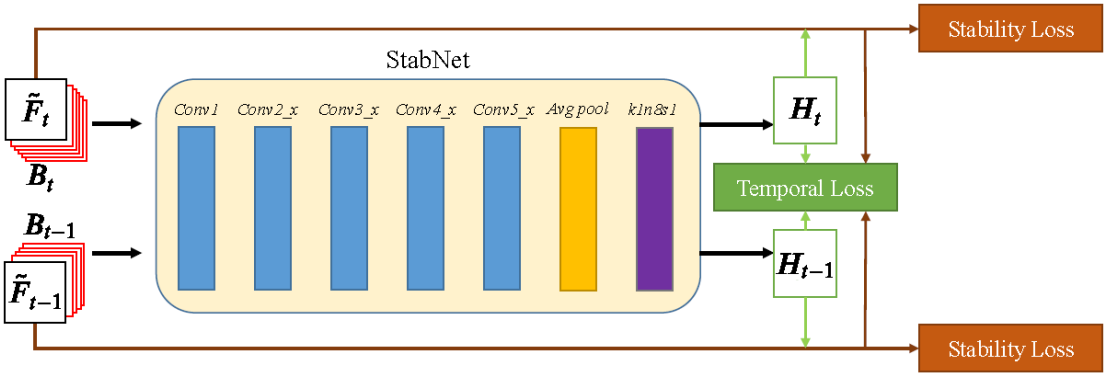

그림 1. Stabnet 구조 [18]
Fig. 1. The architecture of Stabnet [18]

영상을 합성하여 훈련하는 방법을 제안한다.

1. 제안하는 Wobbling 왜곡 억제를 위한 네트워크

Wang은 wobbling 왜곡 제거를 위해 현재 비디오 프레임($\tilde{F}_t$)와 직전 비디오 프레임($\tilde{F}_{t-1}$)를 이용한 시간방향 손실 함수를 도입했다. 본 논문에서는 여기에 추가로 history frame을 이용해 wobbling 왜곡을 제거하는 네트워크 구조와 손실함수를 제안한다. 시간방향 손실함수 구성 시 바로 이전 비디오 프레임만을 이용할 경우 Wobbling 왜곡 억제가 효과적이지 못하기 때문에 $\tilde{F}_t$, $\tilde{F}_{t-1}$, $\tilde{F}_{t-2}$를 입력으로 하는 그림 2의 3-branch 네트워크 구조를 제안한다. 2-branch 네트워크 구조를 이용해 단순히 현재 프레임을 이전 프레임과 유사하도록 학습하는 경우 왜곡이 누적되어 한순간 강한 흔들림으로 나타나는데 3-branch 네트워크 구조는 연속된 프레임의 유사성만 아니라 이전의 차이점과 현재의 차이점을 최소화하도록 학습함으로써 왜곡 편차를 줄일 수 있다.

제안한 3-branch 네트워크에서 각 branch는 Resnet 구조를 가지며, 각 branch 입력은 흔들림이 존재하는 비디오 프레임 $\tilde{F}_t$과 과거에 안정화된 비디오 프레임 $\overline{F}_{t-30}$, $\overline{F}_{t-24}$, $\overline{F}_{t-18}$, $\overline{F}_{t-12}$, $\overline{F}_{t-6}$로 구성된 비디오 프레임 그룹 B_t이다. 흔들림이 존재하는 비디오 시퀀스 $\tilde{S}_u$는 흔들림이 없는 비디오 시퀀스 S_S와 homography 시퀀스 S_H를 이용하여 생성한다. 제안한 3-branch 네트워크는 각 branch별로 비디오 안정화를 위한 homography $\overline{H}_t$를 출력하며, 이를 이용하여 흔들림이 존재하는 인접 프레임들의 안정화된 비디오 프레

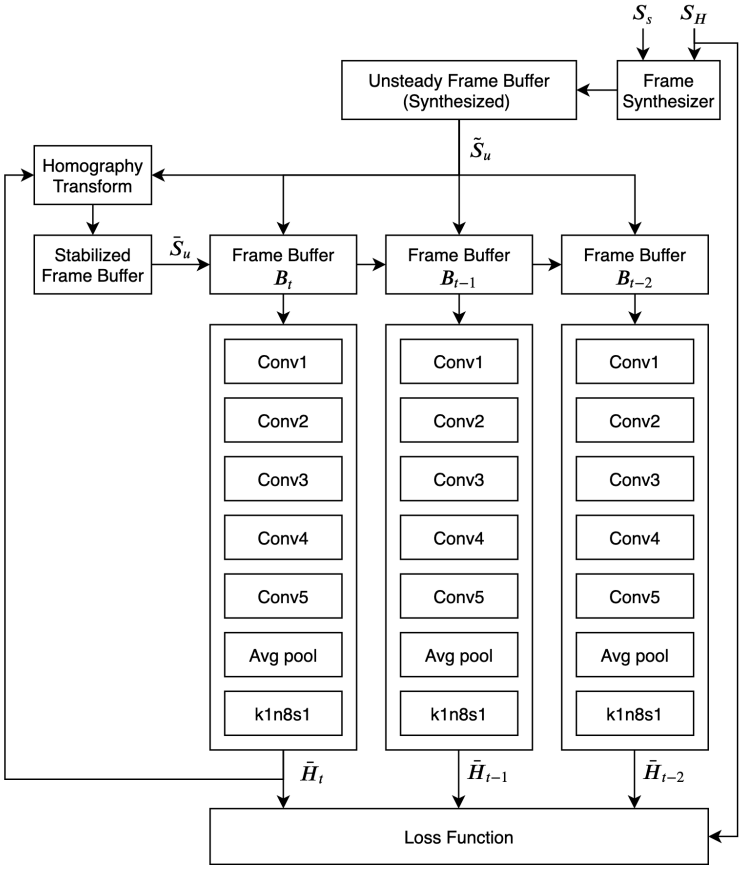

그림 2. 제안하는 3-branch 비디오 안정화기 네트워크 구조
Fig. 2. The proposed 3-branch network architecture

임들 사이의 wobbling 왜곡을 줄이고, 비디오 안정화를 위한 homography 추정의 정확도를 높이는 방향으로 학습을 진행한다.

안정화된 인접 비디오 프레임을 최대한 유사하게 유지하여 Wobbling 왜곡을 억제하기 위한 인접 비디오 프레임 손실 함수는 다음과 같다.

$$L_{temp_1}(\overline{H}_t, \overline{H}_{t-1}, \tilde{F}_t, \tilde{F}_{t-1}) = \frac{1}{D} \| \overline{H}_t * \tilde{F}_t - \omega(\overline{H}_{t-1} * \tilde{F}_{t-1}) \|_2^2 \quad (1)$$

여기에서 $\tilde{F}_t$와 $\tilde{F}_{t-1}$는 현재 시각 t와 이전 시각 $t-1$의 비디오 프레임이고, $\overline{H}_t$와 $\overline{H}_{t-1}$는 안정화를 위한 현재 비디오 프레임과 이전 비디오 프레임의 변환 행렬이다. D는 비디오 프레임 내 화소 수이고, $\omega(.)$는 인접한 과거 steady 비디오 프레임에서 현재 steady 비디오 프레임으로 변환하는 함수로써 optical flow를 이용해 계산된다.

$\tilde{F}_t$와 $\tilde{F}_{t-1}$의 거리와 $\tilde{F}_{t-1}$와 $\tilde{F}_{t-2}$의 거리는 독립적으로 계산되기 때문에 둘의 편차가 클 경우 Wobbling 왜곡이 심해지므로 $\tilde{F}_t$와 $\tilde{F}_{t-1}$의 거리와 $\tilde{F}_{t-1}$와 $\tilde{F}_{t-2}$의 거리를 유사하게 유지하기 위한 인접프레임 손실 비교를 다음과 같이 수행한다.

$$L_{temp_2}(\overline{H}_t, \overline{H}_{t-1}, \overline{H}_{t-2}, \tilde{F}_t, \tilde{F}_{t-1}, \tilde{F}_{t-2}) = \frac{1}{D}(\| \overline{H}_t * \tilde{F}_t - \omega(\overline{H}_{t-1} * \tilde{F}_{t-1}) \|_2 - \| \overline{H}_{t-1} * \tilde{F}_{t-1} - \omega(\overline{H}_{t-2} * \tilde{F}_{t-2}) \|_2) \quad (2)$$

Wobbling 왜곡을 제거하는 시간방향 손실 함수는 인접 프레임 손실 함수와 인접프레임 손실 비교 함수를 이용하여 다음과 같이 정의한다.

$$L_{temp}(\overline{H}_t, \overline{H}_{t-1}, \overline{H}_{t-2}, \tilde{F}_t, \tilde{F}_{t-1}, \tilde{F}_{t-2}) = L_{temp1}(\overline{H}_t, \overline{H}_{t-1}, \tilde{F}_t, \tilde{F}_{t-1}) + L_{temp1}(\overline{H}_{t-1}, \overline{H}_{t-2}, \tilde{F}_{t-1}, \tilde{F}_{t-2}) + L_{temp2}(\overline{H}_t, \overline{H}_{t-1}, \overline{H}_{t-2}, \tilde{F}_t, \tilde{F}_{t-1}, \tilde{F}_{t-2}). \quad (3)$$

제안하는 비디오 안정화기 구조는 online 사용을 목표로 하기 때문에 미래 프레임은 사용하지 않는다. 그러나, off-line 용도로 사용하거나 online을 사용하더라도 출력을 지연시킬 경우 미래 프레임을 사용할 수 있다.

Wang은 GT로 steady 비디오 프레임을 사용기 때문에 특징점 간 거리 측정 손실함수($L_{feature}$)와 안정화된 프레임과 GT 프레임 간 화소 에러 함수(L_{pixel})를 사용했는데 비디오 안정화 성능이 만족스럽지 못하다. 또한, steady한 GT 프레임 확보가 어렵기 때문에 대량의 학습데이터를 확보하기 힘들다.

본 논문에서는 $L_{feature}$과 L_{pixel} 대신에 homography를 GT로 사용하는 L_{trans}를 제안한다. Wang은 $L_{feature}$과 L_{pixel}이 학습된 homography를 이용해 unsteady 비디오 프레임을 변환하고 변환된 비디오 프레임과 GT로 입력된 비디오 프레임을 비교하는 간접적인 접근법으로 인해 학습에 한계가 있었다. 본 논문에서는 homography를 직접 비교하여 학습함으로써 이러한 한계를 극복할 수 있다. 이를 위해서는 GT로 사용할 흔들림 보정을 위한 homography를 확보할 수 있어야 하는데 이는 다음 절에서 설명한다.

제안하는 homography 기반 loss함수는 다음과 같다.

$$L_{trans}(\overline{H}_t, H_t) = \Sigma_{i=1}^{S} \| \overline{h}_t^i - h_t^i \|_{2'}^2 \quad (4)$$

여기에서 $\overline{h}_t^i$는 학습된 homography 행렬 $\overline{H}_t$의 요소, h_t^i는 GT homography 행렬 H_t의 요소이다.

제안하는 3-branch 네트워크 학습을 위한 loss함수는 다음과 같다.

$$L = \Sigma_{i=t,t-1,t-2} a L_{trans}(\overline{H}_i, H_i) + L_{temp}(\overline{H}_t, \overline{H}_{t-1}, \overline{H}_{t-2}, \tilde{F}_t, \tilde{F}_{t-1}, \tilde{F}_{t-2}) \quad (5)$$

여기에서 a는 균형항으로써 본 논문에서는 10을 사용한다.

2. 대량의 학습데이터 생성 방법

Wang이 제안한 데이터셋 확보 방법은 실시간으로 steady 비디오 프레임과 unsteady 비디오 프레임을 확보할 수 있고 직관적으로 비교할 수 있으나 다음 문제들이 존재

한다. 첫째, 이 데이터셋 확보 방법은 steady 카메라와 unsteady 카메라 사이에 시차가 존재하고, 이 시차 때문에 steady 카메라 영상은 unsteady 카메라 영상의 진정한 GT가 될 수 없다. 둘째, GT로 사용되는 steady 영상의 신뢰도가 steady 카메라가 장착된 안정화기의 성능에 의존적이다. 비디오 안정화기의 딥러닝 네트워크는 unsteady 영상을 안정화된 steady 영상으로 변환하도록 훈련을 받는데, 최적의 훈련을 받더라도 그 결과는 steady 영상을 제공하는 기계적 안정화된 영상으로 수렴한다. 실제 Wang 이 제공한 steady 영상에는 사람 걸음걸이에 의한 상하로 끄덕이는 움직임이 남아 있는데 이를 GT로 훈련을 받은 네트워크로 안정화된 영상에도 동일한 움직임이 존재한다.

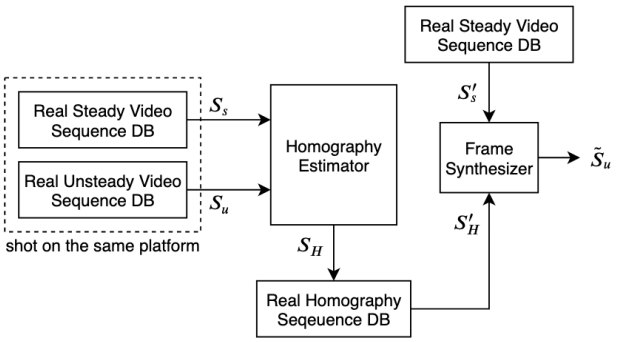

그림 3. 제안하는 학습데이터 생성 방법
Fig. 3. The proposed training data synthesis method

제안하는 학습데이터 생성방법은 그림 3과 같다. S_u와 S_s는 Wang의 두 대의 카메라를 이용하여 촬영한 unsteady 비디오 시퀀스와 steady 비디오 시퀀스다. S_s는 비디오 안정화 장치에 연결된 카메라에서 촬영되었다. 동일 시점에 촬영된 비디오 프레임 쌍인 S_u의 $\tilde{F}_t$와 S_s의 F_t를 이용하여 steady 비디오 프레임으로부터 unsteady 비디오 프레임으로의 변환행렬 H_t를 추출하여 homography 시퀀스 S_H를 생성한다. 학습데이터 시퀀스 생성을 위해 임의의 steady 비디오 시퀀스 $S_s^{'}$와 임의의 homography 시퀀스 $S_H^{'}$을 이용하여 unsteady 비디오 시퀀스 $S_u^{'}$를 합성한다.

제안 방법을 사용하면 쉽게 확보 가능한 steady 비디오 시퀀스와 기 확보된 homography 시퀀스를 이용하여 unsteady 비디오 시퀀스를 합성할 수 있다. Homography 시퀀스는 Wang 이 사용한 두 대의 카메라를 이용하는 방법으로부터 추출하거나, 카메라의 흔들림을 모델링하여 사용하여 생성 가능하다. Steady 비디오 시퀀스를 먼저 확보한 뒤 별도로 생성된 homography sequence를 이용해 steady 비디오 시퀀스를 unsteady 비디오 시퀀스로 합성한다. 합성된 unsteady 비디오 시퀀스와 homography sequence를 이용하여 제안한 네트워크를 훈련시킨다. 제안방법은 비디오 안정화기 학습을 위한 데이터와 정확한 GT를 확보할 수 있다.

Ⅳ. 실험결과 및 비교분석

제안한 비디오 안정화기 네트워크의 학습과 테스트를 위해 Redmon[19]이 구현한 Darknet 딥러닝 플랫폼에서 제안 구조와 Stabnet을 구현하였다.

표 1. 실험에 사용한 파라미터 셋
Table 1. The parameter-set used in this test

parameter	value	description
Image size	448x288	the input image size for the network
Max iteration	90000	the maximum number of iterations for network training
Batch size	8	the number of images used for each network training
Learning rate	0.001	the learning rate used for network training
a	10	the balancing factor between network loss functions, L_{trans} and L_{temp}

1. Wobbling 왜곡 억제를 위한 네트워크 학습결과 및 비교분석

Wobbling 왜곡 억제 효과를 비교하기 위해 1개, 2개, 3개의 branch를 갖는 네트워크를 각각 구현하여 학습하였다. Branch의 숫자만 다를 뿐 그 외 모든 파라미터는 동일하게 사용하였다.

그림 4와 그림 5는 n개의 branch를 갖는 네트워크를 학습시킨 후 입력된 unsteady 비디오 프레임을 안정화한 결과이다. 그림 4에서 빨강색 원을 관찰하면 1-branch는 프레임이 위로 움직였다가 아래로 움직이고, 2-branch는 프레임이 아래로 점점 내려가고, 3-branch는 아래로 아주 천천히 내려가는 것을 볼 수 있다. 프레임간 유사도가 증가할수록 평균

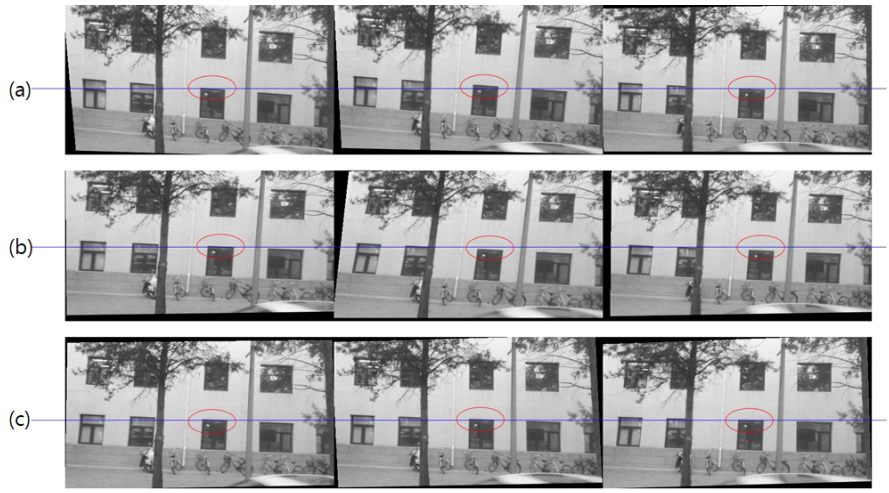

그림 4. N-branch 네트워크 별 비디오 안정화 결과, (a) 1-branch 네트워크, (b) 2-branch 네트워크, (c) 3-branch 네트워크
Fig. 4. The video stabilization results for each n-branch network, (a) 1-branch, (b) 2-branch, (c) 3-branch

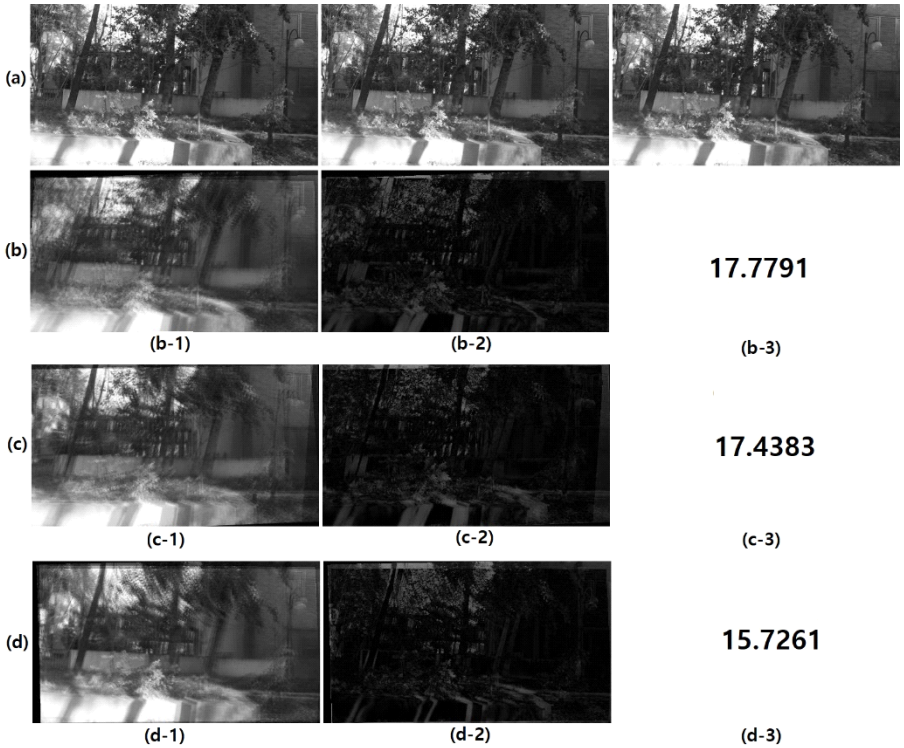

그림 5. N-branch 네트워크 별 비디오 안정화 성능 (a) 입력 unsteady 비디오 프레임, (b) 안정화된 비디오 프레임 (1-branch), (c) 안정화된 비디오 프레임 (2-branch), (d) 안정화된 비디오 프레임 (3-branch); n-branch의 제1열: 안정화된 비디오 프레임의 평균, n-branch의 제2열: 안정화된 비디오 프레임의 표준편차, n-branch의 제3열: 각 branch별 모든 pixel의 표준편차의 평균
Fig. 5. Video stabilization performance for n-branch networks, (a) input unsteady video frame (b) stabilized video frame (1-branch) (c) stabilized video frame (2-branch) (d) stabilized video frame (3-branch); The 1st column of n-branch: average of stabilized video frames, the 2nd column of n-branch: standard deviation of stabilized video frames, the 3rd column of n-branch: the average of standard deviation

이미지에 블러링이 줄어들고 시간방향으로 화소 별 표준편차가 줄어들기 때문에 그림 5로부터 branch가 증가할수록 프레임간 유사도가 증가하고 있다. 그림 4와 그림 5의 결과로부터 branch가 증가할수록 wobbling 왜곡이 줄어드는 것을 확인하였다.

표 2. 각 branch별 모든 pixel의 표준편차의 평균
Table 2. the average of standard deviation for each number of branches

the type of a network	1-branch	2-branch	3-branch
the average	17.7791	17.4383	16.7261

2. 학습데이터셋 생성, 학습결과 및 비교분석

본 절에서는 제안한 학습데이터 합성 방법을 구현하고 이들을 이용한 학습시 Stabnet과 학습 경향을 비교 분석하였다.

2.1 본 논문의 제안 구현

Qu[20]과 Lu[21]은 비디오 안정화 알고리즘을 평가하기 위한 방법으로 random한 움직임과 기존 영상으로부터 추출한 움직임을 통한 합성 영상 사용을 제안하고 있다. 본 논문에서는 Wang 이 제공한 unsteady 영상으로부터 추출한 흔들림을 사용하여 Stabnet의 결과와 비교하였다.

2.2 Stabnet 구현

Stabnet 구현에서 세 가지 변형 및 구체화가 이루어졌다. 첫째, Wang은 균형 가중치 λ를 30으로 제안했는데 그럴 경우 학습이 제대로 이루어 지지 않아 본 논문에서는 3으로 설정하였다. 둘째, Darknet에 구현을 하기위해서는 loss함수에 대한 미분이 필요해 이 부분을 별도 계산하여 구현하였다. 마지막으로 Wang은 online 사용을 가정하여 online 예측 시 history 비디오 프레임에 발생하는 black border를 학습에 반영하기 위해 학습을 위해 history 비디오 프레임에 random black border를 삽입하였는데, 본 논문에서 이 과정은 적용하지 않았다.

2.3 학습과정 비교

제안방식과 Stabnet의 학습과정을 비교 분석하였다. 먼저 훈련 시 training loss가 포화되는 시점을 비교하였고, 이의 보완으로 validation 셋을 이용한 loss 그래프를 비교하였다. Training 및 validation은 모두 Wang 이 제공한 61개 비디오 시퀀스의 44,000장의 비디오 프레임을 사용했으며 이를 7:3의 비율로 나누어 training 집합과 validation 집합을 구성했다. Training 집합은 43개 비디오 시퀀스의 29,990장의 비디오 프레임으로 구성되었으며 validation 집합은 18개 클립의 13,595장의 비디오 프레임으로 구성되었다.

학습 시 그림 6과 같이 Stabnet은 loss 포화지점에 매우 빠르게 도달한다. Batch size 8을 이용해 학습하였고 약 4,000회 학습을 하고 나면 이미 한번 학습훈련에 사용한 비디오 프레임을 재 사용한다. 이로 인해 Stabnet은 약 30,000회 학습을 진행하고 나면 포화상태에 도달한다. 반면 제안한 네트워크는 포화상태에 도달하지 않고 계속 훈련이 진행되는 것을 볼 수 있다. 이는 기본 steady 비디오 프레임은 동일하더라도 계속 새로운 homography로 변형된 unsteady 비디오 프레임들이 학습에 이용되기 때문이다.

Stabnet과 제안한 네트워크의 이러한 학습과정의 Loss 수렴 경향은 validation 결과를 보면 더욱 명확하다. Validation은 각 10, 20, …, 100, 200, 1000, 2000, …, 10000, 20000, …, 90000 회 학습의 결과를 저장한 뒤 그 결과를 이용하여 validation 집합을 예측하고 검증했다. Stabnet은 약 20000회 정도가 되면 validation loss가 오히려 증가하는 overfit경향을 보여주는 데 반해, 제안한 네트워크는 validation loss도 꾸준히 감소하는 것을 볼 수 있다.

추가로 그림 6의 (c)는 branch 비교를 위해 실험한 결과를 보여주는데 그림 6의 (b)와 같은 이미지 DB를 사용했으나 training/validation set 분할이 달라진 데이터 셋을 사용했다. 그림 6의 (c)에서도 그림 6의 (b)처럼 포화상태에 도달하지 않고 훈련이 진행되는 것을 관찰할 수 있다.

이상의 training loss와 validation loss 실험을 통해 Stabnet이 제안한 네트워크에 비해 포화상태에 느리게 도달함을 확인하였다. 이로부터 본 논문에서 제안한 unsteady 비디오 프레임 합성방법이 잘 작동하는 것을 확인할 수 있다.

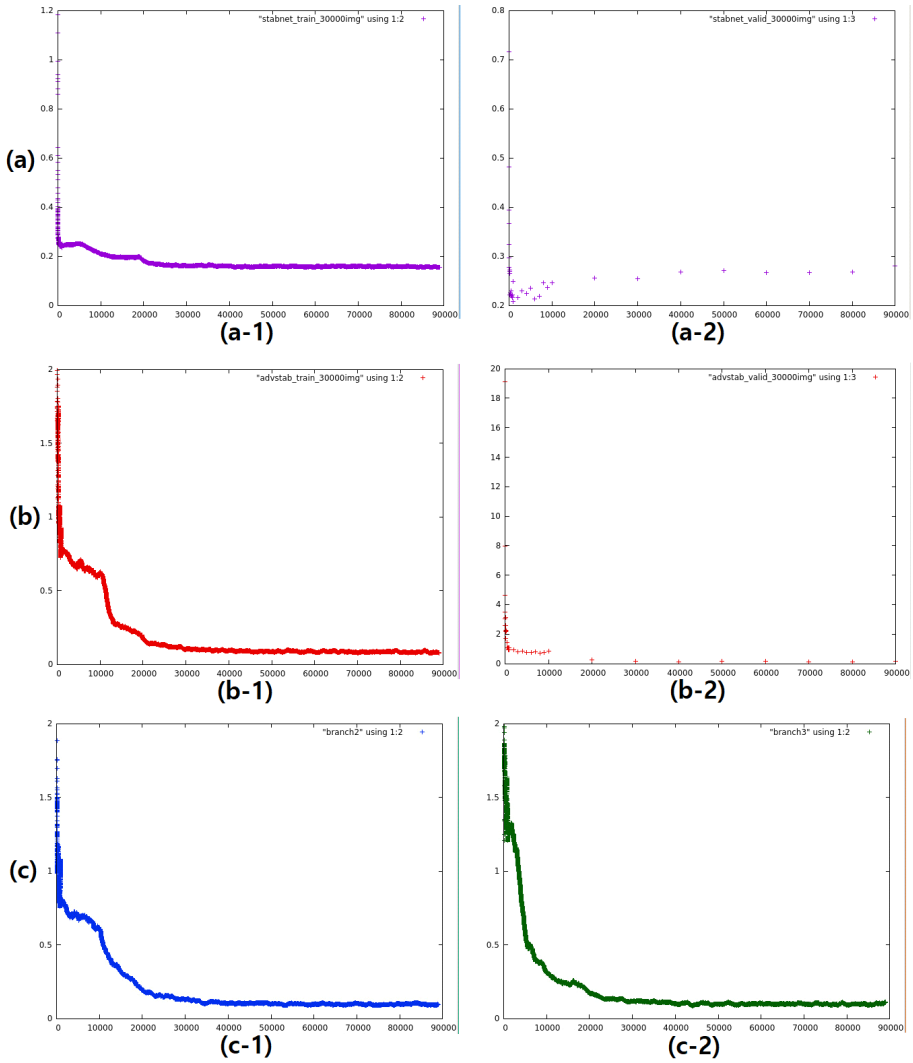

그림 6. Stabnet과 제안한 네트워크의 training loss 및 validation loss 비교, (a) Stabnet, (b), (c) 제안한 네트워크, (a-1), (b-1) Stabnet과 제안한 3-branch 네트워크의 training loss, (a-2), (b-2) Stabnet과 제안한 3-branch 네트워크의 validation loss, (c-1) 제안한 2-branch 네트워크의 training loss, (c-2) 제안한 3-branch 네트워크의 training loss

Fig. 6. Comparison training and validation losses between the Stabnet and the proposed network, (a) Stabnet, (b), (c) proposed network, (a-1), (b-1) training losses of the Stabnet and the proposed 3-branch network, (a-2), (b-2) validation losses of the Stabnet and the proposed 3-branch network, (c-1) training loss of the proposed 2-branch network, (c-2) training loss of the proposed 3-branch network

V. 결론 및 추후 연구과제

본 논문에서는 triplet 네트워크 구조를 응용해 프레임간 차이를 유사하게 유지함으로써 wobbling 왜곡을 억제하는 비디오 안정화기 네트워크를 제안하고 실험을 통해 그 성능을 입증하였다. 또한, 비디오 안정화기 학습을 위해 steady 비디오 시퀀스와 기 확보된 homography 시퀀스로부터 unsteady 비디오 시퀀스를 합성하는 방법을 제안하였다. 제안한 비디오 안정화기는 다음과 같은 경우에 대한 검증과 추가 연구가 필요하다. 첫째, 고정형 카메라가 제한된 범위내에서 흔들리는 경우 병진운동과 회전운동만 고려한 학습이 더 좋은 성능을 기대할 수 있다. 둘째, 이동형 카메

라가 흔들리는 경우 카메라 이동에 따른 영상의 변화와 카메라 흔들림에 따른 변화를 구분한 학습과정의 손실함수 설계가 필요하다.

참 고 문 헌 (References)

[1] P. Rawat and J. Singhai, "Efficient Video Stabilization Technique for Hand Held Mobile Videos," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, Vol. 6, No. 3, pp.17-32, Jun. 2013.

[2] W. J. Freeman, *Digital Video Stabilization with Inertial Fusion*, Master's Thesis of Virginia Polytechnic Institute, VA, 2013.

[3] F. L. Rosa *et al*., *Optical Image Stabilization (OIS)*, White paper. STMicroelectronics, 2015.

[4] S. Bayrak, *Video Stabilization: Digital and Mechanical Approaches*, Master's Thesis of Middle East Technical University, Ankara, Turkey, 2008.

[5] M. J. Smith *et al*., "Electronic Image Stabilization using Optical Flow with Inertial Fusion," *Proceeding of IEEE/RSJ International Conference on Intelligent Robots and Systems*, Taipei, Taiwan, pp.1146-1153, 2010.

[6] J. Xu *et al*., "Fast feature-based video stabilization without accumulative global motion estimation," *IEEE Transactions on Consumer Electronics*, Vol. 58, No. 3, pp. 993-999, Sep. 2012, https://ieeexplore.ieee.org/document/6311347.

[7] B. Pinto and P. R. Anurenjan, "Video stabilization using Speeded Up Robust Features," *Proceeding of International Conference on Communications and Signal Processing*, Calicut, India, pp. 527-531, 2011.

[8] S. Battiato *et al*., "SIFT Features Tracking for Video Stabilization," *Proceeding of 14th International Conference on Image Analysis and Processing*, Modena, Italy, pp. 825-830, 2007.

[9] C. Harris and M. Stephens, "A combined corner and edge detector," *Proceeding of Fourth Alvey Vision Conference*, Manchester, England, pp. 147-151, 1988.

[10] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91-110, 2004, https://doi.org/10.1023/B:VISI.0000029664.99615.94.

[11] K. Veon, *Video Stabilization using SIFT Features, Fuzzy Clustering, and Kalman Filtering*, Master's Thesis of University of Denver, Denver, CO, 2011.

[12] M. A. Fischler and R. C. Bolles "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Comm. ACM*, Vol. 24, No. 6, pp. 381-395, Jun. 1981, https://doi.org/10.1145/358669.358692.

[13] C. Yin *et al*., "Removing Dynamic 3D Objects from Point Clouds of a Moving RGB-D Camera," *Proceeding of International Conference on Information and Automation*, Lijiang, China, pp. 1600-1606, 2015.

[14] K. He *et al*., "Deep Residual Learning for Image Recognition," *Proceeding of Computer Vision and Pattern Recognition*, Las Vegas, NV, pp. 770-778, 2016.

[15] S. Zagoruyko and N. Komodakis, "Learning to Compare Image Patches via Convolutional Neural Networks," *Proceeding of Computer Vision and Pattern Recognition*, Apr. 2015.

[16] E. Hoffer and N. Ailon, "Deep metric learning using Triplet network," *Proceeding of Computer Vision and Pattern Recognition*, Mar. 2015.

[17] M. Jaderberg *et al*., "Spatial Transformer Networks," *Proceeding of the 28th International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 2017-2025, 2015.

[18] M. Wang *et al*., "Deep Online Video Stabilization," *arXiv,* Feb. 2018.

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once:Unified, Real-Time Object Detection," *Proceeding of Computer Vision and Pattern Recognition*, Las Vegas, NV, pp. 779-788, 2016.

[20] H. Qu, L. Song, and G. Xue, "Shaking video synthesis for video stabilization performance assessment," *Proceeding of Visual Communications and Image Processing*, Kuching, Malaysia, pp. 1-6, 2013.

[21] S.-P. Lu *et al*., "Synthesis of Shaking Video Using Motion Capture Data and Dynamic 3D Scene Modeling," *Proceeding of 25th IEEE International Conference on Image Processing*, Athens, Greece, pp. 1438-1442, 2018.

[22] M. Grundmann, V. Kwatra, and I. Essa, "Auto-Directed Video Stabilization with Robust L1 Optimal Camera Paths," *Proceeding of Computer Vision and Pattern Recognition*, Colorado Springs, CO, pp. 225-232, 2011.

저 자 소 개

양 병 호

- 1994년 2월 : KAIST 전산학과 학사
- 2019년 2월 : 한국항공대학교 항공전자정보공학과 석사
- 2018년 2월 ~ 현재 : 서보산전 수석연구원
- ORCID : https://orcid.org/0000-0003-2241-0580
- 주관심분야 : 딥러닝, 영상처리, 모터제어

이 명 진

- 2001년 8월 : KAIST 전자전산학부 박사
- 2001년 3월 ~ 2004년 2월 : 삼성전자 System LSI 사업부 책임
- 2004년 3월 ~ 2007년 2월 : 경성대학교 전기전자공학전공 조교수
- 2007년 3월 ~ 현재 : 한국항공대학교 항공전자정보공학부 교수
- ORCID : https://orcid.org/0000-0002-3136-2819
- 주관심분야 : 영상통신, 영상처리, 임베디드시스템