ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
https://www.jiisonline.org
J Intell Inform Syst 2023 June: 29(2): 115~127 http://dx.doi.org/10.13088/jiis.2023.29.2.115

동영상 안정화를 위한 옵티컬 플로우의 비지도 학습 방법

이보희
성균관대학교 DMC공학과
(btotheh@skku.edu)

김광수
성균관대학교 소프트웨어융합대학
(kim.kwangsu@skku.edu)

동영상 안정화 기술은 최근 1인 미디어 시장이 거대화됨에 따라 그 중요성이 점점 커지고 있는 카메라 기술 중 하나이다. 딥러닝 기반의 기존 방법들에서는 안정화 전/후 동영상 데이터 쌍을 사용하였으나 동영상의 특성상 동기화된 안정화 전/후 데이터를 만드는 것은 많은 시간과 노력이 필요하다. 최근 이러한 문제를 완화하기 위하여 안정화 전 데이터만을 사용하는 비지도 학습 방법이 제시되고 있다. 본 논문에서는 비지도 학습 방법의 하나인 Convolutional Autoencoder 구조를 사용하여 안정화 전/후 동영상 데이터 쌍 없이 안정화 전 영상만으로 안정화 궤적을 학습하는 네트워크 구조를 제안한다. 네트워크 입력 및 출력으로 옵티컬 플로우를 사용하고 네트워크 경량화 및 노이즈 최소화를 위해 옵티컬 플로우를 Grid 단위로 맵핑하여 사용했다. 또한 비지도 학습 방법으로 안정화된 궤적을 생성하기 위해 옵티컬 플로우를 부드럽게 만드는 손실함수를 정의하였고 결과 비교를 통해 손실함수의 의도대로 부드러운 궤적을 생성하도록 네트워크가 학습되었음을 확인했다.

주제어 : 동영상 안정화, 딥러닝, 비지도학습, 옵티컬 플로우

논문접수일 : 2023년 4월 20일 논문수정일 : 2023년 5월 10일 게재확정일 : 2023년 5월 12일
원고유형 : Fast Track 교신저자 : 김광수

1. 개요

동영상 안정화란 동영상 촬영 시 발생한 의도하지 않은 떨림을 제거하고 의도된 움직임만을 남기는 기술이다. 촬영 시 발생하는 흔들림을 제거하기 위해 스테디캠(Steadicam), 짐벌(Gimbal)과 같은 장비를 사용하기도 하지만 이런 장비를 사용할 경우 추가 장비를 구매해야 하고 해당 장비를 소지하고 다녀야 하는 불편함이 발생한다. 따라서 최근 출시되는 카메라는 자체 동영상 안정화 솔루션을 탑재하고 있으며 사용자가 의도하지 않은 흔들림을 촬영과 동시에 제거할 수 있도록 제공한다. 하지만 단말에 탑재된 기술의 경우 대부분 자이로 센서(Gyroscope sensor)와 같은 센서 데이터를 사용하는데 이런 추가 데이터는 따로 저장이 되지 않기 때문에 저장 후의 영상은 추가 편집이 어렵다는 단점이 있다. 어도비 프리미어(Adobe Premiere), 구글 유투브(Google YouTube) 등의 동영상 편집 프로그램들은 후처리로 동영상 안정화 기능을 지원하고 있지만, 이미지의 잔상이 남는 고스트 현상과 울렁임 같은 왜곡이 발생할 가능성이 있다.

전통적인 동영상 안정화 연구는 일반적으로 움직임 추정, 궤적 안정화, 결과 이미지 생성 단계를 포함하고 있다. 또한, 기존 연구들은 크게 3D, 2D 방식으로 분류되는데, 3D 방식은 3D 평면상에서 카메라의 위치를 모델링하고 부드럽게 만드는 방법(Liu et al., 2009)으로 높은 보정성능을

보여주지만 모델이 무겁고 실시간으로 사용하기 어렵다. 2D 방식은 카메라의 위치를 2D Affine transformation 등으로 표현하고 부드럽게 만드는 방법(Grundmann et al., 2011, Liu et al., 2013)으로, 깊이 정보를 처리하지 못하지만 빠르고 강건한 점에서 이점을 보이기 때문에 본 논문은 2D 방식을 사용한다.

딥러닝 기반 동영상 안정화 연구는 StabNet (Wang et al., 2018)을 통해 처음으로 제안되었는데 CNN을 사용해 동영상 안정화가 가능함을 보였다. StabNet은 카메라 경로를 나타내지 않고 Mesh-grid warping transformation을 학습하는 방식으로 처리 속도가 빠르고 저화질 동영상 처리에 강건함을 보였다. 하지만 StabNet은 지도 학습 방법으로 안정화 전/후 동영상 데이터 쌍이 필요했는데, 이 데이터 쌍을 직접 제작하기 위해 동일한 카메라 2대 중 한대를 짐벌에 연결하고 나머지 한대를 짐벌 옆에 부착한 상태로 안정화 전/후 상태를 동시에 촬영했다. 또한, 두 영상의 시작과 끝 지점을 맞추기 위해 추가적인 동기화 작업을 진행해 데이터셋 제작의 어려움을 보여주었다.

이를 보완하기 위해 최근에는 흔들리는 영상만으로 학습이 가능한 비지도 학습 방식이 제안되고 있다. 비지도 학습 방식인 DUT (Xu et al., 2022)는 안정화 전 영상만으로 움직임을 추정하는 네트워크와 궤적 안정화를 학습하는 네트워크를 각각 제안했다. DUT는 전통적인 방식으로 궤적을 명시하고 단계를 나누어 궤적을 안정화했지만 본 논문에서는 궤적이 아닌 옵티컬 플로우를 입력으로 받은 뒤 안정화하는 네트워크를 제안해 단계를 축소한다.

합성곱 오토인코더 구조를 사용하는 방법(Xu et al., 2018)은 안정화 전 이미지와 생성된 안정화 후 영상을 네트워크 입력으로 사용하는 딥러닝 프레임워크를 제시했다. 이 방법은 오토인코더의 각 Layer에 Spatial transformer 네트워크를 추가하여 Affine transformation을 계산하고 이를 영상의 Warping에 사용했다. 그러나 이 방법은 오토인코더의 손실함수에서 안정화 후 영상을 Ground truth로 사용하는 지도학습 방식이기 때문에 학습을 위한 안정화 전/후 데이터 쌍이 필요했다. 본 논문에서는 합성곱 오토인코더 구조를 사용하면서 안정화 후 영상을 사용하지 않는 손실함수를 새롭게 정의해 비지도학습을 가능하게 한다.

또한, 픽셀 단위 Warp field를 계산하는 네트워크 구조들(Yu et al., 2020, Chen et al., 2021) 중 옵티컬 플로우의 정확도를 향상시키는 방법으로 안정화 성능을 끌어올린 방법(Yu et al., 2020)은 옵티컬 플로우의 사용이 좋은 성능을 보임을 입증하였다. 따라서 본 논문에서는 옵티컬 플로우를 네트워크 입력 및 출력으로 사용한다.

본 논문은 안정화 전/후 데이터 쌍 제작의 어려움을 해결하기 위해 안정화 전 동영상만을 사용하는 비지도 학습 방법을 제안하고, 빠르고 정확한 보정성능을 위해 2D 방식으로 카메라 궤적을 계산한다. 본 논문에서 제안하는 동영상 안정화 방법은 다음과 같다.

1. 비지도 학습 구조를 사용하여 안정화 전/후 동영상 데이터 쌍 없이 안정화 전 동영상만으로 학습이 가능한 네트워크 구조를 제안한다.
2. 입력과 출력의 차이를 최소화하여 부드러운 궤적을 생성하는 네트워크 손실함수를 정의하고 이를 학습한다.
3. 옵티컬 플로우를 $N \times N$ 크기의 Grid 단위로 나누고 이 구조를 네트워크의 입력 및 출력으로 사용함으로써 노이즈 영향을 최소화한다.

2. 관련 연구

2.1 전통적인 동영상 안정화 방법

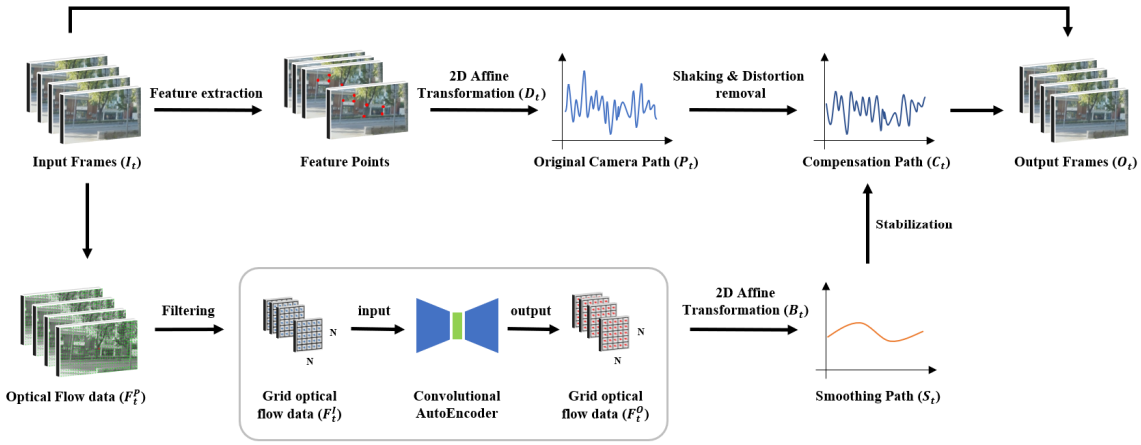
동영상 안정화 기법은 크게 2D, 3D 방법으로 분류할 수 있다. 이 중 3D 방법은 3D 공간상에서 카메라 위치를 모델링하고 추정(Liu et al., 2012, Bradley et al, 2021)하는데, 공간상에서 원본 카메라 경로를 복구하고 Linear 혹은 Quadratic 경로로 카메라 궤적을 피팅하여 영상을 안정화한다. 콘텐츠 보존을 위한 3D 비디오 안정화 방법(Liu et al., 2009)은 안정화된 경로를 기반으로 장면의 왜곡과 뒤틀림을 복구하여 출력 프레임을 만들며 기존 3D 기술에서 보였던 움직이는 물체의 고스팅 현상을 피하면서 2D 기술보다 안정적인 결과를 보여준다. 하지만 3D 방법은 빠른 모션, 빠른 장면 변화, 큰 폐색이 있는 경우와 같이 20 프레임 이상의 긴 객체 추적이 필요한 경우 처리하기 어렵다는 단점을 보인다.

이를 해결하기 위해 3D 기술처럼 고품질의 결과를 유지할 수 있는 강력한 2D 모델(Liu et al., 2013)이 제안되었는데, 이 방법은 공간적으로 변하는 여러 개의 카메라 궤적을 번들 카메라 궤적 모델로 제안하고 이를 유지하는 방법을 제시했다. 이미지 각 위치에 고유한 카메라 궤적을 계산하여 유연한 모델을 가질 수 있도록 하고 시차 및 롤링 셔터 효과로 인한 비선형 모션을 처리할 수 있도록 했다. 이것은 두 개의 연속 프레임 사이의 특징 대응만 필요하므로 3D 방법의 견고성과 단순성을 유지할 수 있다. 또 다른 2D 방법으로, 긴 객체 추적이 필요한 3D 방법의 단점을 보완하기 위해 고비용의 장면 3D 재구성 없이 후처리로 동작하는 실시간 처리 방법(Grundmann et al., 2011)이 제안되었다. 이 방법은 원하지 않은 흔들림을 제거하기 위해 전문 촬영 감독이 사용하는 카메라 모션을 모방하여 Constant, Linear, Parabolic 세 모션으로 구성된 L1-optimal 카메라 궤적을 계산한다. 이것을 위해 카메라 궤적의 1차, 2차, 3차 도함수를 최소화하는 선형 프로그래밍 프레임워크를 기반으로 한 알고리즘을 제안했다. 1차 도함수는 Static 카메라와 같은 움직임이 없는 상태를 만들기 위해 Constant path를 만들고 패닝이나 Dolly shot과 같은 속도가 일정한 궤적을 만들기 위하여 2차 도함수를 사용했다. 마지막으로 점점 빨라지거나 느려지는 변화처럼 가속도가 일정한 궤적을 만들기 위하여 3차 도함수를 사용했다. 여기에 Inclusion, Proximity, Saliency 제약사항을 추가하고 각 도함수에 가중치를 부여함으로써 Trend 필터링을 달성했다. 이것은 고주파 성분만 억제하는 기존의 필터링을 넘어서 비디오 안정화를 가능하게 했다.

2.2 딥러닝 기반 동영상 안정화 방법

딥러닝 기반 동영상 안정화 기법은 StabNet(Wang et al., 2018)에 의해 처음으로 제안되었는데 CNN을 사용해 동영상 안정화 알고리즘이 학습 가능함을 보였다. 이 방법은 CNN 구조에 미래 프레임을 사용하지 않는 실시간 모델이다. 카메라 경로를 나타내지 않고 Multi-grid warping transformation을 학습하는 구조인 StabNet을 제안하여 처리 속도가 빠르고 저화질 동영상 처리에 강건함을 보였다. 지도 학습 구조이므로 직접 카메라 2대로 안정화 전/후 데이터 쌍을 취득하였는데 30초 이내의 총 60쌍의 동기화된 비디오 데이터를 획득했다. 이 중 44쌍을 학습에 사용하고 8쌍을 검증에, 나머지 8쌍을 테스트에 사용했다.

이후 옵티컬 플로우를 사용한 딥러닝 방법(Yu

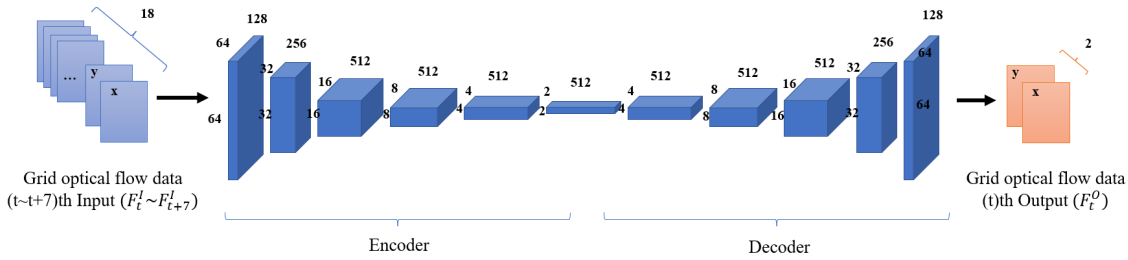


〈Figure 1〉 Pipeline of the proposed method.

et al., 2020)으로 픽셀 단위 Warp field를 계산하는 네트워크 모델을 제안했다. 이 모델은 옵티컬 플로우로 모션을 계산하고 옵티컬 플로우 데이터의 정확도를 향상시키는 모션 인페인팅(Inpainting) 방법과 주파수 도메인에서의 손실함수를 정의해서 안정화 성능의 정확도를 향상시켰다. 그 결과 다른 접근 방법에서 처리하기 어려웠던 움직이는 물체, 가려짐 및 옵티컬 플로우 부정확성에 대해 강건한 결과를 보였다. 이 방법은 안정화 성능 향상을 위해 옵티컬 플로우를 사용해서 모션을 정확하게 계산하는 것에 초점을 맞추었으며 옵티컬 플로우를 사용해서 모션 벡터를 추정하는 것이 정확도가 높다는 결과를 보였다. 또한 계산 복잡도를 극복하여 빠른 수행 속도(570ms/frame)를 달성했다.

최근에는 안정화 전 영상만으로 학습이 가능한 비지도 학습 방법들이 제안되고 있다. DUT(Xu et al., 2022)는 DNN 기반으로 비지도 학습 구조를 구성하고 전통적인 방식으로 움직임 추정, 궤적 안정화 두 단계를 학습하는 네트워크를 제안했다. 우선 움직임 추정 단계에서는 이미지를 Grid 기반으로 보정하기 위해 각 Grid의 Motion vector를 계산하고 grid의 궤적들을 복원했는데, 여기서 Motion vector는 Motion initialization module과 Motion refinement module의 두 단계를 거쳐 계산되었다. Motion initialization module은 옵티컬 플로우를 계산하는 PWCNet(Sun et al, 2018)과 옵티컬 플로우를 신뢰할 수 있는 데이터로 가공하는 RFNet(Shen et al, 2019)으로 구성되어 있는데, RFNet의 결과를 K-means clustering으로 군집화함으로써 Motion vector를 계산했다. Motion refinement module은 옵티컬 플로우를 조정하는 네트워크로 Local motion에 의한 영향성을 제거했다. 이후 궤적 안정화 네트워크에서 각 Grid의 궤적을 부드럽게 만드는 손실함수를 정의해 부드러워진 궤적을 생성했고, 부드러운 Grid 궤적을 기반으로 입력 이미지를 Grid 단위로 Warping해 안정화된 결과 영상을 생성했다. 이 방식은 궤적 계산의 정확도를 향상시켰고 궤적 기반의 안정화를 진행하여 안정화 성능을 극대화하는 것에 초점을 맞추었다.

또 한편으로는, 동영상 안정화를 위해 필요했던



〈Figure 2〉 Network architecture.

Cropping으로 인한 화각 손실을 극복하기 위해 손실되는 영역을 딥러닝을 통해 채우는 Full-frame 방법(Choi et al., 2020, Choi et al., 2021, Liu et al., 2021)이 제안되었다. 이 방법들은 생성형 딥러닝 구조를 사용해 안정화 전 프레임을 기반으로 이웃 프레임 보간 방법을 제시했지만 시차가 큰 동적 객체 및 건물 주변에 고스트 아티팩트가 나타나는 단점을 보였다. 이 외에도 센서 데이터와 이미지 모두를 사용하는 비지도 학습 DNN 구조 [13] 등의 방법도 제시되었다.

3. 제안 방법

3.1 파이프라인

<Figure 1>은 안정화 영상을 생성하기 위해 입력부터 출력까지의 흐름을 보여준다. 제안하는 방법은 크게 세 단계로 이루어져 있는데, 카메라의 궤적을 계산하는 궤적 분석 단계, 네트워크를 통해 부드러운 궤적을 생성하는 궤적 안정화 단계, 분석 궤적과 안정화 궤적을 이용하여 출력 이미지를 생성하는 이미지 생성 단계이다.

우선 궤적 분석 단계에서는 흔들리는 RGB 컬러 입력 영상에서 카메라 궤적을 예측한다. 카메라의 원본 궤적을 계산하기 위해 입력 영상에 대해 특징점을 추출하고 계산된 특징점으로 2D Affine transformation matrix를 계산한 뒤 이를 곱하여 원본 궤적을 예측한다(Grundmann et al., 2011). 시간 순서대로 입력되는 입력 이미지 $I_1, I_2, \ldots, I_t$에 대해 원본 궤적 P_t는 (I_{t-1}, I_t)의 쌍으로 계산된 Transformation matrix D_t의 곱으로 나타낸다.

$$P_t = D_1 D_2 \ldots D_t \qquad (1)$$

P_t는 카메라의 초기 위치에서부터 시간 t까지 누적된 이동량을 의미하므로 P_t의 이동, 회전, 기울기 값을 보정함으로써 입력 영상의 흔들림과 왜곡을 제거한다.

흔들림과 왜곡을 단순히 모두 제거한다면 영상의 움직임에 따라 Cropping 영역이 지나치게 커질 수 있다. 따라서 Cropping 영역을 적절히 조절하고 촬영 의도를 반영하기 위하여 궤적 안정화 단계에서 부드러운 궤적을 계산한다. 우선 입력 영상에 대해 픽셀 단위 옵티컬 플로우 F_t^P를 계산하고,

$$F_t^P = OpticalFlow(I_{t-1}, I_t) \qquad (2)$$

이것을 네트워크 입력으로 만들기 위하여 $N \times N$ Grid 구조에 매핑한다. 이때 F_t^P의 노이즈를 제거하기 위하여 중간값 필터를 적용하고

[0 ~ 1] 사이의 값으로 정규화하여 네트워크 학습이 가능하도록 한다. 네트워크 입력에 대해 출력된 $N \times N$ Grid를 이용하여 2D Affine transformation matrix B_t를 계산하고 수식 (3)과 같이 곱해,

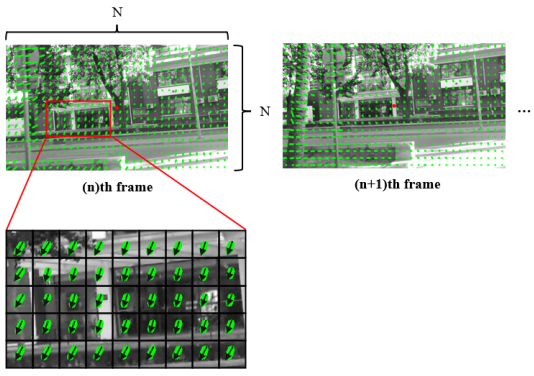
$$S_t = B_1 B_2 \ldots B_t \quad (3)$$

안정화된 경로 S_t를 얻을 수 있으며 S_t에 가우시안 필터를 적용하여 네트워크 출력에서 발생할 수 있는 노이즈를 제거한다. 네트워크와 입력 및 출력 구조에 대한 자세한 내용은 **3.2.**에서 다루었다.

결과 이미지 생성 단계에서는 궤적 분석 단계의 결과 P_t 와 궤적 안정화 단계의 결과 S_t를 이용하여 수식 (4)와 같이 보정량 C_t을 계산하고,

$$C_t = P_t^{-1} \bullet S_t \quad (4)$$

C_t를 입력 이미지 I_t에 적용하여 최종적으로 안정화된 RGB 컬러 이미지 O_t를 생성한다.



〈Figure 3〉 Grid-based optical flow data.

3.2 네트워크 구조

<Figure 2>는 네트워크 구조를 보여준다. 네트워크는 합성곱 오토인코더(Convolutional Autoencoder) 기반으로, (Xu et al., 2018)에서 제안한 구조를 사용했고 입력, 출력 데이터 및 손실함수를 변경하여 사용했다.

네트워크 입력을 만들기 위하여 우선 흔들리는 입력 영상에서 인접한 프레임을 분석하여 F_t^P를 계산한다. F_t^P에는 신뢰할 수 없는 데이터와 노이즈가 존재할 가능성이 있어 이러한 노이즈를 제거하기 위해 필터링 과정을 거친다. <Figure 3>은 F_t^P를 Grid 단위로 매핑하는 것을 보여준다. <Figure 3>의 초록색 선은 각 영역의 옵티컬 플로우의 크기와 방향을 의미한다. 이것은 입력 이미지와 크기가 같은 F_t^P를 $N \times N$ Grid로 균일하게 나눈 각 구역의 대푯값이다. 이후 $N \times N$ Grid 값에 중간값 필터를 적용하여 옵티컬 플로우 계산이 취약한 반복 무늬, 단색 영역 등에서 발생할 수 있는 오류를 최소화했다. 본 논문에서는 $N = 64$로 놓고 64×64크기의 데이터를 사용했다.

이렇게 계산된 네트워크 입력 F_t^I는 (t)번째 프레임부터 $(t+7)$번째 프레임의 총 18채널로 이루어진 x, y 값이며 네트워크 출력 F_t^O는 (t)번째 프레임의 총 2채널 x, y 값이다. 따라서 네트워크는 입력과 동일한 64×64 크기의 옵티컬 플로우 데이터를 출력한다.

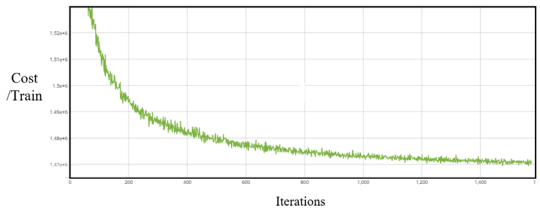
3.3 손실함수

네트워크의 학습에 사용한 손실함수는 다음과 같다.

$$Loss = \sum_{t=0}^{T}(w_1 * \left| F_t^O \right|_1 + w_2 * \sum_{f=0}^{7} \left| F_{t+f}^I - F_t^O \right|_1) \quad (5)$$

손실함수는 두 항으로 이루어져 있는데, 우선 네트워크 입력 $F_1^I, F_2^I, ..., F_t^I$과 출력 $F_1^O, F_2^O, ..., F_t^O$에 대해 시간 $(t \sim t+7)$ 사이의 움직임을 반영하는 항 $\sum_{f=0}^{7} \left| F_{t+f}^I - F_t^O \right|_1$이 있다. 이 항은 영상의 전체적인 이동량을 고려하며 F_t^O가 $F_t^I \sim F_{t+7}^I$의 평균값에 가까워지게 만든다. 또 다른 항은 최대한 안정화된 움직임을 만들 수 있도록 하는 항 $\left| F_t^O \right|_1$으로 출력 결과의 이동량을 작아지게 만든다. 여기서 두 항의 가중치를 의미하는 w_1, w_2를 조절하여 안정화 정도를 조절했다.

수식 (5)을 사용하여 학습을 충분히 진행하면 네트워크는 움직임이 작아지는 방향으로 안정화된 옵티컬 플로우를 출력한다.

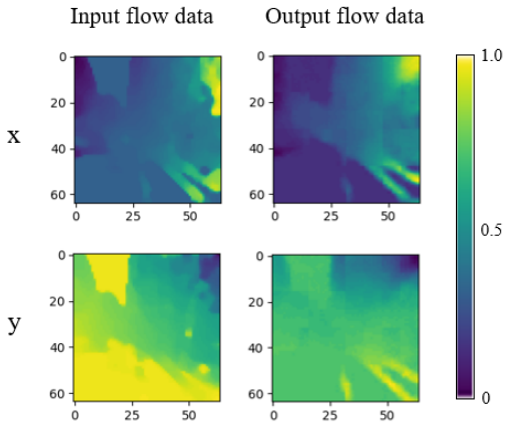


〈Figure 4〉 Cost/Trains per iterations.

4. 실험 및 결과

4.1 데이터셋

네트워크 학습 및 테스트를 위해 StabNet(Wang et al., 2018) 데이터셋을 이용했다. StabNet은 직접 카메라 2대로 안정화 전/후 데이터 쌍을 취득하였는데 총 30초 이내의 60 쌍의 동기화된 비디오 데이터를 획득했다. 이 중 우리는 안정화 전 동영상만을 이용하여 총 54개의 영상을 학습에 사용하고 6개의 영상을 테스트에 사용했다.



〈Figure 5〉 Network input and output.

4.2 평가 지표

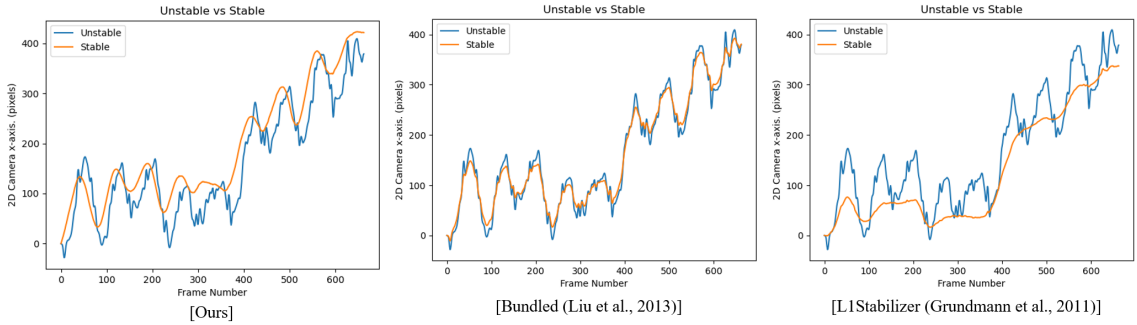
제안한 방법을 평가하기 위해 우리는 결과 영상에 대한 안정화 전/후의 궤적을 그래프로 나타냈다. 안정화 전/후 궤적은 안정화 전/후 영상에 각각 수식 (1)를 적용하여 계산했다.

또한, 안정화 전/후를 정량적으로 분석하기 위해 우리는 Stability, Distortion, Cropping Ratio의 세가지 평가지표(Liu et al., 2013, Zhang et al., 2017)를 사용하였으며 각 평가 방법은 다음과 같다.

Stability는 안정화 후 영상의 부드러움 정도를 의미하는데, 안정화 전/후 궤적에서 이동 및 회전 성분을 구한 뒤 각 성분의 가속도를 구하고 안정화 전/후 가속도의 비율로 나타낸다. Stability는 큰 값을 가질수록 가속도가 일정하고 결과 영상이 안정되었음을 의미한다.

〈Table 1〉 Quantitative comparison with existing methods.

	Stability(↑)	Distortion(↑)	Cropping Ratio(↑)
Bundled (Liu et al., 2013)	0.600	0.881	0.959
L1Stabilizer (Grundmann et al., 2011)	0.869	0.877	0.899
DUT (Xu et al., 2022)	0.719	0.960	0.949
StabNet (Wang et al., 2018)	0.736	0.980	0.855
Ours	0.820	0.917	0.886



[Ours] [Bundled (Liu et al., 2013)] [L1Stabilizer (Grundmann et al., 2011)]

[DUT (Xu et al., 2022)] [StabNet (Wang et al., 2018)]

〈Figure 6〉 Comparison of trajectories before and after stabilization with existing methods.

Distortion 값은 안정 후 영상에서 나타나는 왜곡 정도를 의미하는데 왜곡이 크게 발생하면 영상이 비틀리는 문제가 보인다. Distortion은 안정화 후 영상의 궤적에서 Affine transformation matrix를 구하고 이 Matrix의 가장 큰 두 Eigenvalue의 비율로 나타낸다. 모든 프레임에 대해 값을 구한 뒤 이 중 가장 최악의 값을 사용하고 값이 클수록 왜곡이 작은 것을 의미한다.

Cropping ratio는 안정화를 위해 잘라낸 영역을 의미하고 안정화 후 프레임과 안정화 전 프레임의 비율을 사용하여 계산한다. Cropping ratio도 Distortion 값과 마찬가지로 영상의 모든 프레임에 대해 값을 구한 뒤 가장 작은 최악의 값을 사용한다.

4.3 실험 결과

제안한 네트워크의 반복횟수에 따른 Cost/Train 값은 <Figure 4>와 같으며 학습이 진행됨에 따라 수렴하는 것을 볼 수 있다.

<Figure 5>는 충분히 수렴한 상태의 네트워크를 동작시켰을 때, 시간 t에서 네트워크 입력 F_t^I과 출력 F_t^O을 그림으로 나타낸 것이다. 여기서 F_t^O은 F_t^I에 비해 값이 작아졌고 경계선이 부드러워진 것을 볼 수 있다. 이것은 입력 $F_t^I \sim F_{t+7}^I$과 비슷하지만 최대한 작은 움직임이 되는 방향으로 출력이 부드러워진 것을 의미하고 손실함수의 의도와 같음을 보여준다.

동일한 입력 영상에 대해 기존 연구들과 비교한 결과를 <Figure 6> 및 <Table 1>에 나타내었다. <Figure 6>은 안정화 전/후 영상에서 안정화 전 궤적(Unstable)과 안정화 후 궤적(Stable)을 계산해 그래프로 나타낸 것이다. 비교에 사용한 기존 연구는 전통적인 방법인 Bundled(Liu et al., 2013), L1Stabilizer(Grundmann et al., 2011)의 두 방법과 딥러닝 방법 중 비지도 학습 방법인 DUT(Xu et al., 2022)와 지도학습 방법인 StabNet(Wang et al., 2018)의 두 방법이다. <Table 1>은 동일한 영상에 대해 4.2에서 설명한 평가지표를 사용해 각 방법의 Stability, Distortion, Cropping ratio 값을 계산한 것이다. Stability가 좋을수록 <Figure 6>의 그래프에서 안정화 후 궤적의 그래프가 부드럽게 나타나야하고 Cropping ratio가 좋을수록 안정화 전과 후 궤적의 차이가 적어야 한다.

L1Stabilizer. 방법은 원본 궤적의 추세를 따라가면서도 안정화 궤적을 최대한 부드럽게 그려주어 Stability 항목에서 가장 높은 점수를 받았다. 하지만 영상의 외곽에 Distortion이 크게 발생하여 Distortion 점수가 가장 작게 나타났다. DUT의 경우 영상 내에서 왜곡이 가장 작게 보였지만 흔들림이 많이 남아 있어 Stability 값이 가장 작았다. 우리가 제안한 방법의 경우 손실함수의 의도에 맞게 부드럽게 처리된 안정화 후 궤적을 볼 수 있었다.

Stability, Distortion 성능과 Cropping ratio는 Trade-off 관계에 있어 기존의 연구는 한가지 성능을 높여 다른 성능을 하락시키는 결과를 보였다. Bundled. 방법의 경우 Cropping ratio 성능이 가장 좋았지만 Stability 성능은 가장 낮은 결과를 보였고, Stability와 Distortion 성능이 가장 좋았던 L1Stabilizaer, StabNet은 Cropping ratio 성능에서 열세를 보였다. 하지만 우리가 제안한 방법의 경우 세 성능 모두에서 열세를 보인 항목이 없었다. 이것은 Stability 성능과 Cropping ratio 성능 사이의 적정선을 찾기 위해 수식 (5)의 가중치 w_1, w_2를 조절했음을 의미한다. 만약 w_1가 더 컸을 경우 Stability 성능이, w_2가 더 컸을 경우 Cropping ratio 성능이 높게 나왔을 것으로 예상한다.

또한, 우리가 제안한 방법의 결과에서는 다른 방법들의 결과와는 다르게 안정화 궤적이 왼쪽으로 Shift 되어 있는 것을 볼 수 있었는데, 이것은 시간 $(t \sim t+7)$에 대한 입력을 시간 t에 반영한 결과로 보인다.

5. 결론

본 논문은 동영상 안정화를 위한 데이터셋 제작의 문제를 비지도 학습 모델을 사용함으로써 해결하는 방안을 제안한다. 이를 위해 안정화 전 영상에서 옵티컬 플로우를 구한 뒤 이것을 Grid

단위의 작은 크기로 가공하고 필터링하여 합성곱 오토인코더 구조에서 학습한다. 여기서 비지도 학습을 위해 옵티컬 플로우를 안정화하는 새로운 손실함수를 정의하였다. 그 결과 정량적 비교를 통해 기존 전통적인 방법들보다 왜곡 보정 관점에서 효과적이고 기존 딥러닝 방법들보다 안정화 성능 관점에서 효과적임을 증명했다. 네트워크의 입력 및 출력이 작은 크기의 옵티컬 플로우이므로 해상도에 상관없이 기존에 존재하는 안정화 전 영상을 모두 학습에 사용할 수 있을 것으로 예상한다.

본 연구는 간단한 손실함수를 정의했기 때문에 다양한 씬에 유연하게 대응하지 못할 수 있다는 한계점이 존재한다. 이를 보완하기 위해 후속 연구에서는 다양한 씬을 고려한 손실함수를 사용해 성능을 검증할 필요가 있다. 또한, 최종적으로는 옵티컬 플로우가 아닌 이미지 자체를 입력을 받아 안정화 이미지를 출력하는 End-to-end 구조로의 연구가 필요하다.

참고문헌(References)

[국외 문헌]

Bradley, A., Klivington, J., Triscari, J., & van der Merwe, R. (2021). Cinematic-L1 video stabilization with a log-homography model. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1041-1049).

Chen, Y. T., Tseng, K. W., Lee, Y. C., Chen, C. Y., & Hung, Y. P. (2021, September). Pixstabnet: Fast multi-scale deep online video stabilization with pixel-based warping. In 2021 IEEE International Conference on Image Processing (ICIP) (pp. 1929-1933). IEEE.

Choi, J., & Kweon, I. S. (2020). Deep iterative frame interpolation for full-frame video stabilization. ACM Transactions on Graphics (TOG), 39(1), 1-9.

Choi, J., Park, J., & Kweon, I. S. (2021). Self-supervised real-time video stabilization. arXiv preprint arXiv:2111.05980.

Grundmann, M., Kwatra, V., & Essa, I. (2011, June). Auto-directed video stabilization with robust l1 optimal camera paths. In CVPR 2011 (pp. 225-232). IEEE.

Liu, F., Gleicher, M., Jin, H., & Agarwala, A. (2009). Content-preserving warps for 3D video stabilization. ACM Transactions on Graphics (ToG), 28(3), 1-9.

Liu, S., Wang, Y., Yuan, L., Bu, J., Tan, P., & Sun, J. (2012, June). Video stabilization with a depth camera. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 89-95). IEEE.

Liu, S., Yuan, L., Tan, P., & Sun, J. (2013). Bundled camera paths for video stabilization. ACM transactions on graphics (TOG), 32(4), 1-10.

Liu, Y. L., Lai, W. S., Yang, M. H., Chuang, Y. Y., & Huang, J. B. (2021). Hybrid neural fusion for full-frame video stabilization. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2299-2308).

Shen, X., Wang, C., Li, X., Yu, Z., Li, J., Wen, C., ... & He, Z. (2019). Rf-net: An end-to-end image matching network based on receptive field. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8132-8140).

Shi, Z., Shi, F., Lai, W. S., Liang, C. K., & Liang, Y. (2022). Deep online fused video stabilization. In Proceedings of the IEEE/CVF Winter

Conference on Applications of Computer Vision (pp. 1250-1258).

Sun, D., Yang, X., Liu, M. Y., & Kautz, J. (2018). Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8934-8943).

Wang, M., Yang, G. Y., Lin, J. K., Zhang, S. H., Shamir, A., Lu, S. P., & Hu, S. M. (2018). Deep online video stabilization with multi-grid warping transformation learning. IEEE Transactions on Image Processing, 28(5), 2283-2292.

Xu, S. Z., Hu, J., Wang, M., Mu, T. J., & Hu, S. M. (2018, October). Deep video stabilization using adversarial networks. In Computer Graphics Forum (Vol. 37, No. 7, pp. 267-276).

Xu, Y., Zhang, J., Maybank, S. J., & Tao, D. (2022). DUT: learning video stabilization by simply watching unstable videos. IEEE Transactions on Image Processing, 31, 4306-4320.

Yu, J., & Ramamoorthi, R. (2020). Learning video stabilization using optical flow. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8159-8167).

Zhang, L., Chen, X. Q., Kong, X. Y., & Huang, H. (2017). Geodesic video stabilization in transformation space. IEEE Transactions on Image Processing, 26(5), 2219-2229.

Abstract

Deep Video Stabilization via Optical Flow in Unstable Scenes

Bohee Lee* · Kwangsu Kim**

Video stabilization is one of the camera technologies that the importance is gradually increasing as the personal media market has recently become huge. For deep learning-based video stabilization, existing methods collect pairs of video datas before and after stabilization, but it takes a lot of time and effort to create synchronized datas. Recently, to solve this problem, unsupervised learning method using only unstable video data has been proposed. In this paper, we propose a network structure that learns the stabilized trajectory only with the unstable video image without the pair of unstable and stable video pair using the Convolutional Auto Encoder structure, one of the unsupervised learning methods. Optical flow data is used as network input and output, and optical flow data was mapped into grid units to simplify the network and minimize noise. In addition, to generate a stabilized trajectory with an unsupervised learning method, we define the loss function that smoothing the input optical flow data. And through comparison of the results, we confirmed that the network is learned as intended by the loss function.

Key Words : Video Stabilization, Deep Learning, Unsupervised Learning, Optical Flow.

Received : April 20, 2023 Revised : May 10, 2023 Accepted : May 12, 2023
Corresponding Author : Kwangsu Kim

* Department of Digital Media and Communications Engineering, Sungkyunkwan University
** Corresponding Author: Kwangsu Kim
College of Computing and Informatics, Sunkyunkwan University,
2066 Seobu-ro Jangan-gu, Suwon-si, 16419, Gyeonggi-do, Republic of Korea
Tel: +82-031-290-7969, E-mail: kim.kwangsu@skku.edu

저 자 소 개

이 보 희

현재 성균관대학교 DMC공학과 석사과정에 재학중이며, 삼성전자에서 연구원으로 재직중이다. 국민대학교 전자공학부에서 학사를 취득했다. 주요 관심사는 Computer Vision과 Machine Learning이다.

김 광 수

현재 성균관대학교 소프트웨어학과 교수로 재직중이며, 성균관대학교 인공지능 연구소 소장을 겸하고 있다. 한양대학교에서 전자공학과 학사를 취득했으며, University of Southern California에서 공학석사와 박사를 취득했다. 주요 관심분야는 Computer Vision과 Explainable AI 및 AI Application이다.