RotoNet: Rotoscoping-Based Artistic Style Transfer Networks

최서윤

Division of Software, Data Science

Sookmyung Women's University

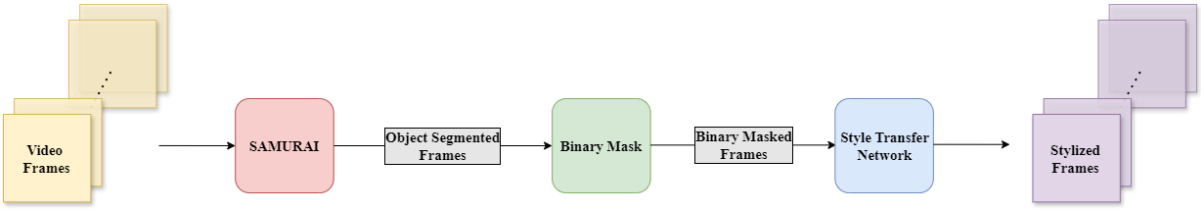

Figure 1. Overview of RotoNet.

Abstract

기존 비디오 스타일 변환 기술은 전체 프레임에 일괄 적용되어 특정 객체의 선택적 변환이 어렵다. 본 연구에서는 로토스코핑 기반의 객체별 스타일 변환을 가능하게 하는 새로운 딥러닝 프레임워크 RotoNet을 제안한다. RotoNet은 객체 추적 네트워크와 스타일 변환 네트워크로 구성되며, 비디오 내 특정 객체에 선택적으로 스타일을 적용하는 것을 목표로 한다. RotoNet이 기존 스타일 변환 모델의 한계를 극복하고, 예술적 표현의 자유도를 확장하며, 영화 및 게임 산업의 제작 파이프라인 혁신에 기여할 가능성을 탐색하고자 한다.

1. Introduction

로토스코핑은 실제 영상 속 특정 객체의 움직임을 추출하여 프레임별로 따라 그리는 애니메이션 제작 기법으로, 자연스러운 움직임을 표현하는 데 널리 활용된다. 특히, 실사 영상을 기반으로 정밀한 동작을 재현할 수 있어 사실적인 애니메이션 제작에 효과적이다. 그러나 로토스코핑은 상당한 시간과 노동력을 요구하는 작업으로, 몇 가지 실무적 제약을 수반한다. 우선, 원하는 장면을 구현하기 위해 사전 실사 촬영이 필요하며, 이는 제작 일정과 비용에 상당한 영향을 끼친다. 또한, 영상의 각 프레임을 개별적으로 처리해야 하므로 높은 인적 자원이 소모되며, 제작 소요 시간이 길어진다. 예를 들어, 밴드 A-ha의 Take On Me 뮤직비디오는 단 4분 길이의 영상으로, 약 3000개의 프레임을 로토스코핑하는 데 16주 이상이 소요되었다[1]. 이러한 한계로 인해 로토스코핑은 대규모 프로젝트에서 효율적으로 적용하기 어려운 기법으로 인식된다.

본 연구는 로토스코핑 기법의 기존 한계를 극복하고, 비디오 스타일 변환 기술의 효율성을 향상시키고자 한다. 전통적인 로토스코핑은 수작업으로 각 프레임을 따라 그리는 방식으로, 제작 소요 시간과 비용이 매우 높고, 대규모 프로젝트에서 실용적인 적용이 어려운 문제가 있다. 이를 해결하기 위해, 본 연구에서는 딥러닝 기반의 객체별 스타일 변환 프레임워크인 RotoNet을 제안한다.

2. Methods

Figure 1은 RotoNet의 전체 구조를 시각적으로 나타낸다. RotoNet은 비디오 내에서 특정 객체를 정확하게 추적하고, 선택적으로 스타일을 적용하기 위한 두 가지 주요 컴포넌트로 구성되어 있다. 첫 번째 컴포넌트인 객체 추적 네트워크는 사용자가 지정한 대상 객체를 초기 프레임에서 인식하고, 이후 전체 비디오 프레임에 걸쳐 해당 객체를 일관되게 분할 및

추적하는 역할을 수행한다. 두 번째 컴포넌트인 스타일 변환 네트워크는 객체 추적 네트워크로부터 전달받은 이진 마스크를 활용하여, 비디오 내 지정된 객체 영역에만 선택적으로 스타일을 적용한다.

2.1. Object Segmentation & Tracking

비디오에서 객체의 정확한 구분과 추적을 위해 SAMURAI를 사용하였다[2]. SAMURAI는 동작 기반 모델링과 동작 인식 메모리 선택 기법을 도입하여, 혼잡한 동적 환경에서도 뛰어난 객체 추적 성능을 발휘한다. 그리고 zero-shot 비디오 객체 분할을 지원하여 첫 번째 프레임에서 박스나 마스크와 같은 간단한 프롬프트만으로도 추가적인 학습 없이 전체 비디오에 걸쳐 대상 객체를 분할하고 추적할 수 있다. 또한, Segment Anything Model(SAM)을 기반으로 구축되어 강력한 분할 성능을 보장하며, 비디오 도메인에 특화된 시공간적 일관성을 제공한다.

2.2. Video Stylization

이미지 스타일 변환 모델을 비디오에 직접 적용할 경우, 프레임 간 일관성이 유지되지 않아 각 프레임마다 스타일이 상이하게 적용되는 'Popping 현상'이 발생하는 문제가 있다. 이러한 시간적 불연속성을 완화하기 위해, 본 연구에서는 현재 프레임과 스타일이 적용된 이전 프레임을 일정 비율로 블렌딩하는 방법을 적용하였다[3][4]. 이를 통해 잔상 효과(Ghost Effect)를 형성하고, 프레임 간의 시간적 연속성을 강화하고자 하였다. 결과적으로 영상 전반에 걸친 스타일 일관성을 확보하고, 스타일 전이에 따른 시각적 이질감을 감소시키고자 하였다.

3. Experiments

Figure 2. 입력 비디오 237번 프레임의 원본, 객체 분할, 이진 마스크, 스타일 변환 결과.

Figure 2는 RotoNet을 적용한 결과로, 각 처리 단계에서의 273번 프레임에 해당하는 이미지이다. 실험은 아래와 같은 순서로 수행되었다.

먼저, 스타일 변환을 적용할 대상 객체를 원본 비디오의 첫 프레임에서 선택한다. 선택된 객체의 위치 정보는 바운딩 박스 형식 (x, y, w, h)으로 정의되며, 이를 기반으로 객체 마스크를 생성한다. 이후 각 프레임마다 후보 마스크들을 생성하고, 객체, 동작 그리고 마스크를 고려한 하이브리드 스코어링 시스템을 통해 가장 적합한 마스크를 선택한다. 선택된 마스크는 현재 프레임의 객체로 확정되며, 이 과정을 반복 수행함으로써 객체 추적이 수행된다.

다음 단계에서는 객체와 배경을 구분하기 위해, 객체를 흰색(255), 배경을 검정색(0)으로 표현한 이진 마스크(Binary Mask)를 생성한다. 해당 마스크는 스타일 변환이 적용될 영역을 명확히 정의하며, 이를 기반으로 이미지 스타일 변환 모델은 객체 영역에만 선택적으로 스타일을 적용한다.

스타일 변환은 이미지의 내용과 스타일을 분리하고 재조합하는 알고리즘을 적용한다. 생성 이미지는 무작위로 초기화되며, 콘텐츠 이미지와 스타일 이미지의 표현과 유사해지도록 손실 함수를 최소화하면서 픽셀 값을 최적화한다. 네트워크의 가중치는 고정된 상태로 학습은 오직 이미지 자체에 대해 수행된다. 마지막으로 스타일이 적용된 객체 영역과 원본 배경을 합성하여 프레임을 구성하며, 이러한 방식으로 처리된 각 프레임을 합성함으로써, 특정 객체에만 스타일 변환이 적용된 최종 비디오를 생성한다.

4. Conclusion

로토스코핑 기법은 예술적인 애니메이션을 제작할 수 있지만, 제작 비용과 시간으로 인해 효율적인 제작 기법으로 평가받지 못한다. 로토스코핑 효과를 내기 위한 스타일 변환 기술은 다양하지만, 특정 객체의 선택적 변환은 어렵다. 본 연구에서는 로토스코핑 기법에 기반한 객체별 스타일 변환 프레임워크인 딥러닝 프레임워크 RotoNet을 제안하였다. 실험 결과 RotoNet은 기존의 스타일 변환 방식과 달리 객체 중심의 선택적 스타일 변환을 효과적으로 수행하였으며, 영상의 예술적 표현 가능성과 실용적 유연성을 동시에 확보할 수 있음을 확

인하였다. 이는 로토스코핑의 자동화 및 비디오 스타일 변환의 정밀한 제어 가능성 측면에서 실질적인 기여를 하여, 영상 제작 파이프라인의 효율성과 창의성 증대에 긍정적인 영향을 미칠 수 있다.

본 연구에서 객체 추적 모델로 SAMURAI만을 활용하여 실험을 진행하였다. 이에 따라 향후 연구에서는 대표적인 객체 검출 및 추적 알고리즘인 YOLO와 DeepSORT 기반의 결합 모델을 적용하여 성능을 비교·분석할 필요가 있다. YOLO는 실시간 객체 탐지에 특화된 모델로, 프레임 단위의 객체 위치를 정확히 탐지한다[5]. DeepSORT는 외관 정보와 모션 정보를 활용한 추적 알고리즘으로, 지속적이고 안정적인 객체 ID 유지에 효과적인 것으로 알려져 있다[6]. 이와 같은 조합은 다양한 객체 추적 성능을 높일 수 있어, SAMURAI 모델과의 정량적 비교가 필요하다. 또한, 본 실험에서 비디오 스타일 변환 모델을 직접 사용하지 않고, 개별 프레임에 대한 이미지 스타일 변환을 수행한 후 시간적 연속성을 보정하는 방식을 채택하였다. 따라서 향후 비디오 스타일 변환 모델을 적용함으로써, 보다 자연스럽고 연속적인 스타일링 결과를 도출하기 위한 추가적인 연구가 필요하다[7].

References

[1] I. Failes. A-ha's rotoscoped 1985 music video for 'Take On Me' has been watched…a lot, befores&afters. https://beforesandafters.com/2020/02/20/a-has-rotoscoped-1985-music-video-for-take-on-me-has-been-watched-a-lot, 2020.

[2] Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang and Jenq-Neng Hwang. SAMURAI: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. arXiv preprint arXiv:2411.11922, 2024.

[3] Leon A. Gatys, Alexander S. Ecker and Matthias Bethge. A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576, 2015.

[4] Tristan Jogminas and Brycen Westgarth. Neural Style Transfer Transition Video Processing, github. https://github.com/westgarthb/style-transfer-video-processor, 2021.

[5] Joseph Redmon, Santosh Divvala, Ross Girshick and Ali Farhadi. You only look once: Unified, real-time object detection. arXiv preprint arXiv: 1506.02640, 2015.

[6] Nicolai Wojke, Alex Bewley and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. arXiv preprint arXiv:1703.07402, 2017.

[7] Zixuan Ye, Huijuan Huang, Xintao Wang, Pengfei Wan, Di Zhang and Wenhan Luo. StyleMaster: Stylize your video with artistic generation and translation. arXiv preprint arXiv:2412.07744, 2024.