한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 30, No. 3, P. 99~107

적응형 블러 기반 비디오의 수평적 확장 여부 판별 네트워크

김민선°[1] 서창욱[2] 윤현호[1] 노준용*[1]

[1]카이스트 비주얼 미디어 연구실 [2]애니그마 테크놀로지스

[1]{sunnykimhappy, yunhhp12, junyongnoh}@kaist.ac.kr [2]lgtwins@anigma-ai.com

Video classifier with adaptive blur network to determine horizontally extrapolatable video content

Minsun Kim°[1] Changwook Seo[2] Hyun Ho Yun[1] Junyong Noh*[1]

[1]KAIST, Visual Media Lab [2]Anigma Technologies

요 약

기존에 존재하는 비디오 영역을 가로 혹은 세로로 확장하는 비디오 확장 기술에 대한 수요가 높아지고 있지만, 최신 기술로도 모든 비디오를 성공적으로 확장할 수는 없다. 따라서 비디오 확장을 시도하기 전에 해당 비디오가 잘 확장될 수 있을지 판단하는 것이 중요하다. 이를 통해 불필요한 컴퓨팅 자원 낭비를 줄일 수 있기 때문이다. 이 논문은 비디오가 수평 확장에 적합한지 판별하는 비디오 분류기를 제안한다. 이 분류기는 광학 흐름과 적응형 가우시안 블러 네트워크를 활용하여 흐름 기반 비디오 확장 방식에 적용할 수 있다. 학습을 위한 라벨링은 유저 테스트 및 정량적 평가를 거쳐 엄격하게 이루어졌다. 이렇게 라벨링된 데이터셋으로 학습한 결과, 주어진 비디오의 확장 가능성을 분류하는 네트워크를 개발할 수 있었다. 제안된 분류기는 광학 흐름과 적응형 가우시안 블러 네트워크를 통해 비디오의 특성을 효과적으로 포착함으로써, 단순히 원본 비디오나 고정된 블러만을 사용하는 경우보다 훨씬 정확한 분류 성능을 보였다. 이 분류기는 향후 다양한 분야에서 활용될 수 있으며, 특히 몰입감 있는 시청 경험을 위해 장면을 자동으로 확장하는 기술과 함께 사용될 수 있을 것으로 기대된다.

Abstract

While the demand for extrapolating video content horizontally or vertically is increasing, even the most advanced techniques cannot successfully extrapolate all videos. Therefore, it is important to determine if a given video can be well extrapolated before attempting the actual extrapolation. This can help avoid wasting computing resources. This paper proposes a video classifier that can identify if a video is suitable for horizontal extrapolation. The classifier utilizes optical flow and an adaptive Gaussian blur network, which can be applied to flow-based video extrapolation methods. The labeling for training was rigorously conducted through user tests and quantitative evaluations. As a result of learning from this labeled dataset, a network was developed to determine the extrapolation capability of a given video. The proposed classifier achieved much more accurate classification performance than methods that simply use the original video or fixed blur alone by effectively capturing the characteristics of the video through optical flow and adaptive Gaussian blur network. This classifier can be utilized in various fields in conjunction with automatic video extrapolation techniques for immersive viewing experiences.

키워드: Video extrapolation, Video classification, Optical flow
Keywords:

*corresponding author: Junyong Noh/ KAIST, Visual Media Lab(junyongnoh@kaist.ac.kr)

Received : 2024.06.15./ Review completed : 1st 2024.07.02. / Accepted : 2024.07.05.
DOI : 10.15701/kcgs.2024.30.3.99
ISSN : 1975-7883(Print)/2383-529X(Online)

1 서론

대형 영화관 스크린부터 개인 기기에 이르기까지 다양한 크기의 디스플레이 비율에 대한 수요가 증가하고 있다. 관객들은 영화관에서 몰입감 있는 경험을 기대하며, 이에 따라 디스플레이 형식이 돌비, IMAX, ScreenX와 같은 더 크고 고품질의 버전으로 진화하고 있다. 특히 ScreenX [1]는 270도의 고품질 시각 콘텐츠를 표시하여 몰입감을 크게 향상시키는 파노라마 영화 시청 경험을 제공한다(그림 1). 한편, 개인 기기에서는 유튜브의 Shorts와 인스타그램의 Reels와 같은 세로 비디오의 인기가 급증하면서 콘텐츠 소비 방식에 변화가 일어나고 있다. 사용자는 이러한 비디오를 세로로 된 스마트 기기뿐만 아니라 가로로 된 화면에서도 시청할 수 있어 화면 비율을 맞추기 위해 양쪽에 필러 박스를 생성하거나 비율을 조정하는 유연성이 요구된다.

Figure 1: ScreenX [1] is a technology designed to transform ordinary movie theaters into multi-projection environments. It expands the cinematic experience by using the left and right walls of a theater as supplementary projection surfaces. ScreenX-enabled movies can then be projected onto these walls, creating an immersive viewing experience.

이러한 수요에 대응하여 비디오 확장 기술의 중요성이 증가하고 있다. 비디오 확장 기술은 기존에 존재하던 비디오 영역을 가로 혹은 세로로 확장함으로써 콘텐츠 제작자가 하드웨어 한계를 극복하고 원하는 비율의 비디오 콘텐츠를 제작할 수 있도록 한다. 그러나 기존 비디오 확장 기술은 확장 영역의 품질과 관련된 문제에 직면하는 경우가 많다. 기존 비디오 확장 기술의 품질에 대한 사용자 평가를 진행하였고, 다음 세 가지 문제점을 발견하였다(그림 2 참조). 전경(foreground) 객체가 확장 경계에서 잘리거나(case1), 확장된 영역에 동적 객체의 잔상이 생기거나(case2), 복원된 전경 또는 배경이 비디오의 컨텍스트에 맞지 않았다(case3).

이렇듯 비디오 확장은 실패로 이어지는 경우가 많다. 이때, 수많은 비디오에 대해 일일이 화면을 확장하고 결과를 확인하는 것은 시간과 자원 낭비로 이어진다. 따라서 비디오가 성공적으로 확장될 수 있는지 사전에 판단하는 것이 중요하다. 비디오가 자연스럽게 확장될 수 있을지 사전에 판단하려면 사용자가 각 장면을 시각적으로 분석해야 한다. 그러나 인력에 의존하는 이러한 과정은 시간이 많이 걸리고 지루하며 전문 지식이 필요하다.

이 연구에서는 실제 확장을 수행하지 않고도 주어진 비디오가 자연스럽게 확장될 수 있는지 자동으로 판단하는 방법을 제안한다. 앞에서 제시한 3가지 경우의 확장 실패의 주요 원인이 모두 확장 경계에서의 객체의 움직임에 기인한다는 것에 착안하여, 이 연구는 이를 효과적으로 감지할 수 있는 방법을 제시한다. 전경 객체의 움직임을 감지하기 위해서는 광학 흐름을 활용한다. 또한 분류기가 확장될 영역에 집중할 수 있도록 하기 위해 foveated view [2, 3, 4, 5, 6, 7] 개념에서 영감을 받은 적응형 가우시안 블러 네트워크를 도입한다. 이 네트워크는 각 프레임에 대해 동적으로 관심 영역(Region of Interest, ROI)을 다르게 생성한다.

본 연구의 기여는 세 가지로 요약할 수 있다. 첫째, 수평으로 확장 가능한 비디오 콘텐츠를 사전에 예측하는 네트워크를 처음으로 고안하였다. 둘째, 사용자 테스트 및 엄격한 정량적 평가를 통해 흐름 기반(flow-based) 비디오 확장 방법에 특화된 데이터셋 라벨링을 수행하였다. 셋째, 광학 흐름 및 적응형 블러 기법을 통합하여 효율적인 비디오 분류를 위한 네트워크를 개발하였다.

2 관련 연구

2.1 비디오 확장(Video extrapolation)

Aides et al. [9]과 Avraham et al. [10]은 확장된 픽셀 세트를 채우기 위해 completion algorithm을 사용하는 foveated view 기반 비디오 확장 방식을 제안하였다. 이 방식은 중심 시야에서 디테일을 우선시하고 주변부로 갈수록 해상도를 점차 감소시키는 인간 시각 시스템에 기반한다. 그러나 사용자는 비디오 콘텐츠에서 주변 환경을 탐색하는 데 흥미를 느끼므로 확장되어 새로 생성된 부분이 흐릿한 foveated view 접근 방식은 한계를 가진다.

Lee et al. [11]은 structure-from-motion(SFM) [12, 13] 을 기반으로 복구된 3D 장면 정보를 활용하여 균일한 품질의 콘텐츠를 생성하는 새로운 방법을 제안함으로써 이 한계를 해결하려고 하였다. 이 방식은 인접 프레임을 왜곡하고 혼합하여 고품질로 비디오를 확장시킨다. 그러나 장면의 추정된 3D 기하학에 크게 의존하므로 프레임 불일치로 인해 확장이 실패할 수 있다.

이 문제를 해결하기 위해 Gao et al. [14]과 Dehan et al. [8]은 딥러닝 기술에 적용하여 흐름 기반 정보를 활용한다. Dehan et al. [8]은 비디오 인페인팅 및 광학 흐름 추정을 위해 여러 딥러닝 모델을 활용하였다. 이는 다양한 상황에서 고품질 비디오 확장을 가능하게 한다(그림 3 참조). 본 연구는 영상 확장 기술 중 현재 가장 좋은 결과를 생성하는 Dehan et al. [8]의 방법을 기반으로, 주어진 비디오가 흐름 기반 비디오 확장 방법을 사용했을때 확장이 잘 되는지 판별하는 분류기를 제안한다.

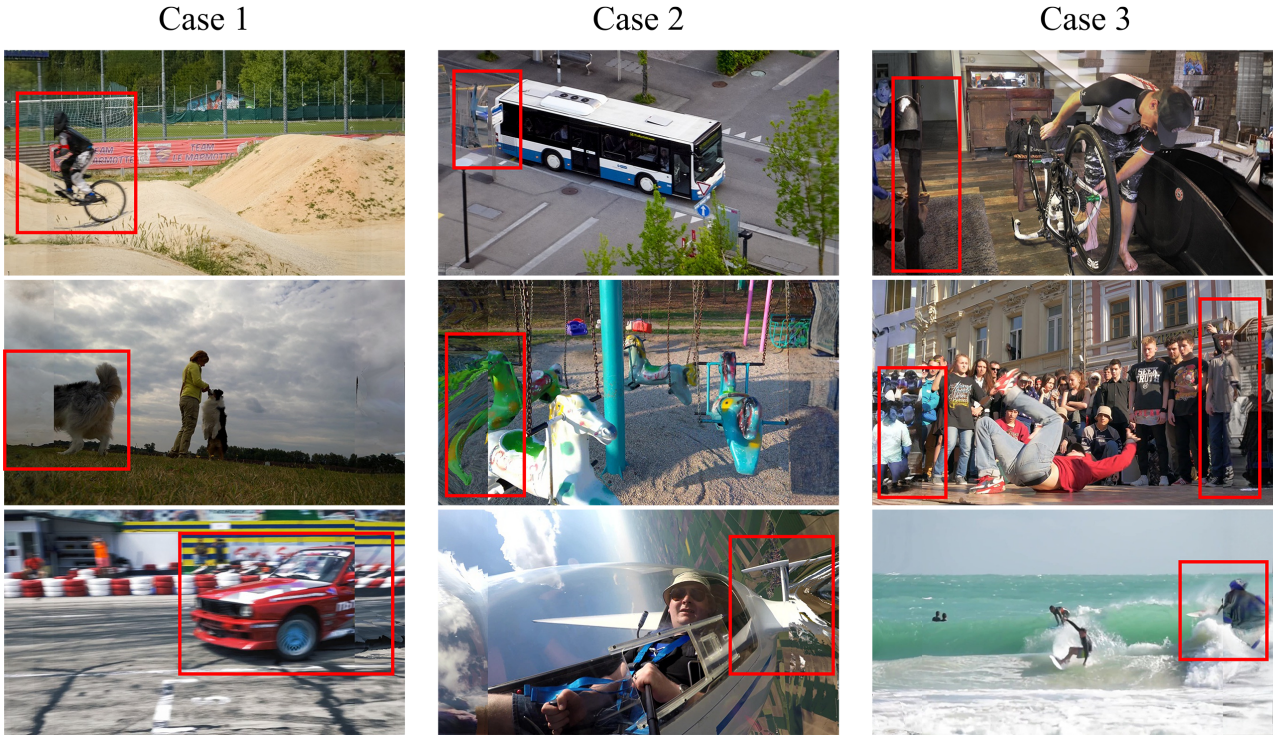

Figure 2: Based on the user test conducted using the DAVIS dataset, extrapolation results that users judged as *unnatural* fall in one of three cases. A video scene that passes all conditions is tagged as *natural*.

Figure 3: Extrapolation result by Dehan et al. [8]. In the experiment, we cropped each side of the frame 12.5% and extrapolated it back.

2.2 행동 인식(Action recognition)

컴퓨터 비전의 주요 과제인 행동 인식은 비디오에서 전체 행동 시퀀스를 분석하여 인간의 행동을 식별하는 것이다 [15]. 최근 몇 년 동안 딥러닝 방법론이 높은 일반화 능력을 가져 강력한 특징을 학습할 수 있어 관심을 끌고 있다 [16, 17, 18, 19]. 행동은 역동적이며 비디오에서 몇 초 동안만 지속되는 경우가 많다. 따라서 비디오의 모든 프레임을 입력으로 사용하는 것은 학습 중 불필요한 계산 비용을 초래한다. 많은 딥러닝 접근 방식은 프레임을 서브샘플링하여 이 부담을 완화하였지만, 행동 인식에 중요한 핵심 프레임을 놓칠 가능성이 있다 [20]. 이를 극복하기 위해 Liu et al. [20]은 학습 중 모든 비디오 프레임을 활용하고 Hamming distance를 기반으로 한 시간 클러스터링 알고리즘을 사용하여 유사한 프레임을 식별하는 전략을 채택하였다. 본 연구에서는 이 방법론을 베이스라인으로 하여 분류기를 설계하였다.

3 방법

3.1 데이터셋

비디오가 확장 가능한지 자동으로 판별하는 본 연구에 대한 엄격한 평가를 수행하기 위해, densely annotated video segmentation dataset (DAVIS) 2017 [21]을 평가 데이터셋으로 선정하였다. DAVIS 데이터셋은 150개의 고화질 비디오로 구성되어 있으며,

Table 1: The top part of the table explains how we will designate the datasets henceforth. For both training and evaluation, we have ground truth denoted as a raw video dataset, the cropped video where the sides were cut off, and the extrapolated dataset. The bottom part of the table indicates the final count of labels categorized either as *natural* or *unnatural* in the train and evaluation datasets.

Dataset purpose		**Training**	**Evaluation**
Ground truth		$Train_{raw}$	$Eval_{raw}$
Crop 25%		$Train_{crop}$	$Eval_{crop}$
Extrapolated back		$Train_{extra}$	$Eval_{extra}$
Tag	natural	81	34
	unnatural	255	116
Total		336	150

비디오 연구에 널리 사용되는 오픈 데이터셋으로, 본 연구의 평가에 적합하다. 학습 데이터셋은 다양한 장면과 촬영 스타일을 포함하는 6편의 영화에서 336개의 샷을 추출하여 사용하였다.

3.2 데이터셋 라벨링

본 연구의 주요 목표는 비디오가 자연스럽게 확장될 수 있는지를 예측하는 것이다. 이를 위해, 비디오가 자연스럽게(*natural*) 혹은 부자연스럽게(*unnatural*) 확장되었는지 라벨링하여 학습 및 평가 데이터셋을 구성하였다. 데이터셋의 각 비디오는 Dehan et al. [8]의 흐름 기반 비디오 확장 방법을 사용하여 확장하였다. 각 비디오를 라벨링하기 위해 정량적 및 정성적 평가를 모두 수행하였다. 원본 비디오에서 양측에 대해 12.5%를 잘라내고, 이를 원래 비율로 다시 확장하는 방식을 취하였다(그림 3 및 그림 5 참조).

엄격한 데이터셋 라벨링을 위해, 두 단계를 거쳤다. 먼저, LPIPS의 임계값을 0.1로 설정하고, 0.1 미만인 장면은 *natural*, 그렇지 않은 경우 *unnatural*로 분류하였다. 다음으로 사용자 테스트를 실시하였다. LPIPS 점수가 균등하게 분포되도록 평가용 데이터셋을 세 가지 다른 버전으로 생성하였다. 각 버전에 대해 5명씩, 총 15명의 평가자가 참여하였다. 비디오 영상은 누구나 시청할 수 있음을 고려하여 평가자는 특정 그룹에 중점을 두지 않았다. 사용자 테스트 결과, 확장 후 *unnatural*로 평가된 비디오는 다음과 같은 세 가지 특성을 나타냈다(그림 2 참조).

사례1. 전경 객체가 확장 경계에서 잘린 경우.

사례2. 확장된 영역에 동적 객체의 잔상이 생긴 경우.

사례3. 복원된 전경 또는 배경이 기존 비디오의 컨텍스트에 맞지 않는 경우.

세 가지 경우 모두에 해당하지 않아 *natural*이라고 평가된 비디오를 최종적으로 *natural*로 라벨링하였다. 학습 데이터셋도 동일한 방식으로 라벨링하였다. 라벨링 과정의 최종 분포는 표 1와 같다.

3.3 데이터 전처리

본 연구에서는 *unnatural*로 라벨링된 세 가지 사례가 어떤 상황에서 발생하는지 분석하였다. 사례 1은 객체가 확장된 경계 방향으로 계속해서 이동할 때 발생하고, 사례 2는 움직이는 객체가 경계에 걸칠 때 발생하며, 사례 3은 확장 경계에서 시각 정보가 불충분할 때 발생한다. 이를 바탕으로, 본 연구에서는 동적 전경 객체의 존재(*특징1*)와 확장되는 영역의 경계선에서 해당 객체의 출현(*특징2*)을 확장 품질에 영향을 미치는 두 가지 주요 특징으로 결론내렸다. 다음 방법을 사용하여 분류기가 각 특징을 감지하도록 했다.

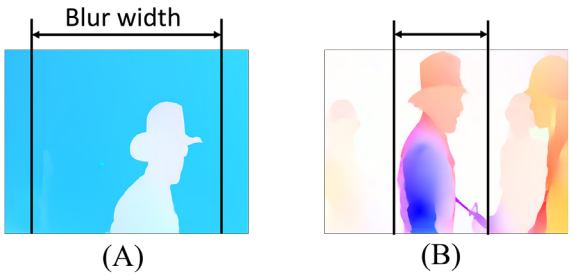

Figure 4: Different frame scenarios require different blur widths. (A) shows a stationary object in the center, allowing for a wider blur width due to the absence of significant periphery artifacts. On the other hand, (B) displays numerous temporal artifacts at the boundary, requiring a narrower blur width and expanded ROI. These varying scenarios reflect the concept of foveated vision [5] and have informed our design of an adaptive Gaussian blur network.

네트워크로 하여금 전경 객체의 움직임에 주목하여 *특징 1*을 잘 감지하도록 하기 위해 밀집 광학 흐름을 사용했다. *특징 2*에 대해서는 foveated vision [4]에서 영감을 받아 가우시안 블러를 프레임의 중심에 적용했다. 이 접근 방식은 비관심 영역(non-ROI)에 블러를 처리해 낮은 해상도를 적용하는 동시에 관심 영역(ROI)에 더 많은 연산 자원을 할당하여 네트워크가 가장자리의 정보에 주목할 수 있도록 한다 [22].

각 프레임 내에서 객체의 수와 움직임의 강도가 다양하기 때문에 가우시안 블러의 적합한 폭인 W_{blur}를 정의하는 것은 복잡하다. 따라서 본 연구에서는 그림 4에서 표기한 것처럼 각 장면의 특성에 따라 블러 폭을 다르게 하는 적응형 블러 네트워크(N_{ab})를 설계했다. 이 개념은 foveated vision의 원리와 일치하며, 네트워크가 유연하게 다양한 시나리오에 주목할 수 있도록 한다.

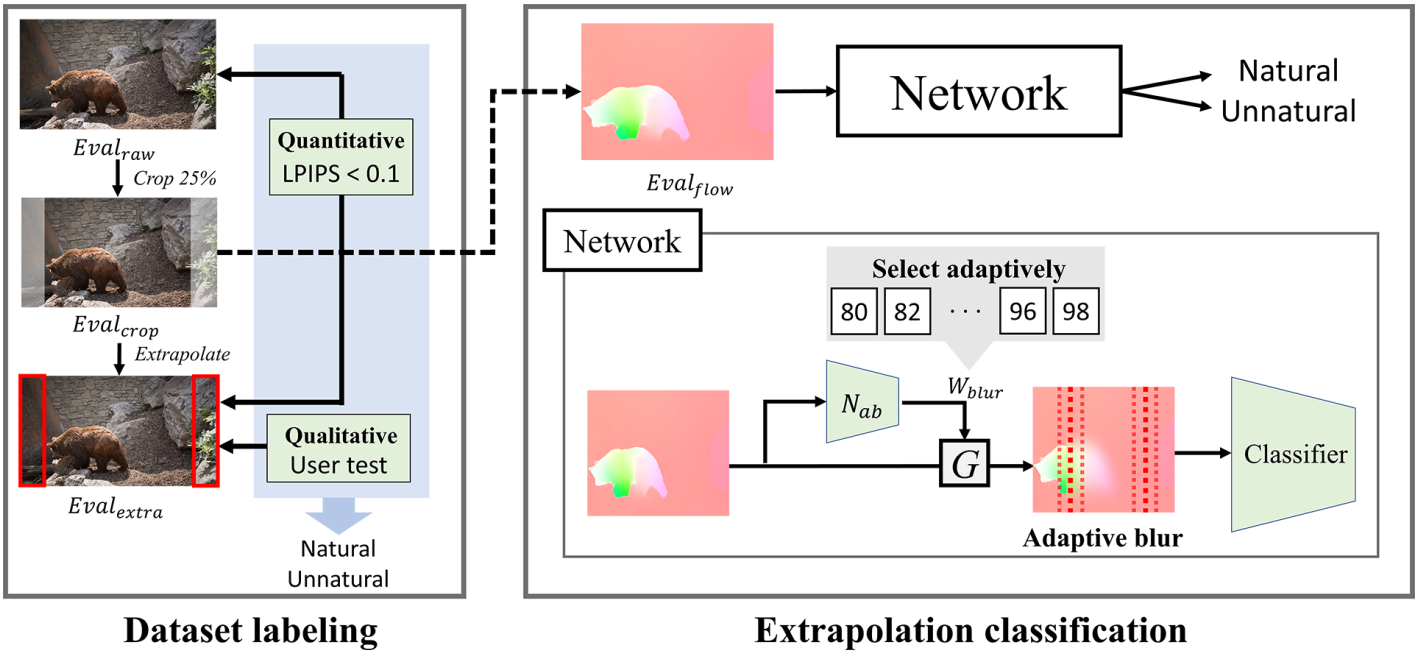

Figure 5: Our work is divided into two key parts: dataset labeling and classification for extrapolatable video content. Dataset labeling initially relies on a quantitative evaluation, where an LPIPS value of 0.1 serves as the threshold. This is followed by a qualitative evaluation using a user test. We establish three cases of extrapolation failure and assign the final *natural* label to videos that pass all these cases. For classification, the optical flow of the cropped dataset is initially used as input. Following this, the Adaptive Blur Network(N_{ab}) selects the most suitable blur width(W_{blur}) for each frame. The W_{blur} list is chosen from one of the top 10 best-performing intervals from the fixed Gaussian blur results. Afterward, frames are blurred by G, and they subsequently pass through a classifier based on Liu et al., which ultimately determines whether the video will be extrapolated well.

3.4 네트워크 구조

실험의 베이스라인 아키텍처는 ResNet-18 [23]을 백본으로 사용하는 TSM [24]모델을 채택하였고, 임의 길이 비디오 처리를 위해 Liu et al. [20]의 방법을 사용하였다. 전체 네트워크는 그림 5에 나타나 있다. 프레임은 ResNet-18로 입력되기 전 적응형 가우시안 블러 네트워크(N_{ab})를 거친다. 이 네트워크는 각 프레임마다 0부터 9까지의 인덱스를 선택하고, 이 인덱스에 해당하는 W_{blur} 값을 미리 정의된 리스트에서 가져온다. 리스트에는 실험적으로 결정된 최적 W_{blur} 값을 중심으로 2 픽셀씩 차이나는 10개의 W_{blur} 값들이 포함되어 있다. 선택된 W_{blur} 값에 따라 함수 G가 프레임에 가우시안 블러를 적용하여 최종적으로 블러 처리된 프레임(I_{blur})을 얻게 된다.

4 실험

평가 데이터셋($Eval_{extra}$)과 학습 데이터셋($Train_{extra}$)은 *natural* 또는 *unnatural*로 라벨링되었다. 학습은 각각 양 옆이 12.5% 잘린 $Train_{crop}$을 사용하여 진행되었고, 모델의 성능은 $Eval_{crop}$으로 평가되었다. 각 프레임의 크기는 224×224로 조정되었으며, 광학 흐름, 고정 가우시안 블러, 적응형 가우시안 블러를 각각 적용한 후 분류 정확도를 평가하였다.

다양한 학습 방법에 따른 성능 차이를 비교하기 위해, 먼저 원본 RGB($Train_{raw}$)와 광학 흐름 RGB($Train_{flow}$) 입력을 사용하여 각각 모델을 학습시켰다. 이후 7개의 서로 다른 고정 W_{blur} 값으로 가우시안 블러를 적용하여 가장 높은 점수를 갖는 최적의 블러 폭 범위를 찾았다(표 3 참조). 이 범위는 이후 적응형 블러 네트워크가 어떤 범위로 블러를 취할지 결정하는 기준이 된다. 성능은 정확도, 평균 정밀도(AP), F1 점수로 평가되었으며, 이는 네트워크의 출력과 3.2에서 얻은 라벨을 비교하여 측정되었다.

표 2에서 볼 수 있듯이, 광학 흐름 RGB 입력이 원본 RGB 입력보다 더 높은 AP와 F1 점수를 보였다. 이는 광학 흐름 RGB 입력이 모델의 성능 향상에 기여한다는 것을 의미한다. 고정 블러 실험에서는 W_{blur} 값에 따라 상이한 결과가 관찰되었으며(표 3), 가장 우수한 성능은 80에서 100 사이의 값에서 나타났다. 그에 따라, 이 범위에서의 W_{blur}를 갖는 적응형 블러를 적용하였다. 그 결과, 광학 흐름 RGB 입력에 적응형 블러를 적용한 방법이 고정 블러보다 우수한 성능을 보였으며, 가장 높은 정확도(89.333%), AP(0.584), F1 점수(0.704)를 달성하였다.

적응형 블러는 고정 블러보다 *unnatural* 비디오를 효과적으로 분류하였다. 그림 6 은 W_{blur}를 96으로 고정하여 학습한 분류기가 *unnatural* 비디오를 *natural*로 잘못 판단한 세 가지 시나리오를 보여준다. 시나리오 (A)에서는 객체가 경계에서 부자연스럽게 잘린 것이 명확하게 보인다. 매우 역동적인 객체가 포함된 (B)의 경

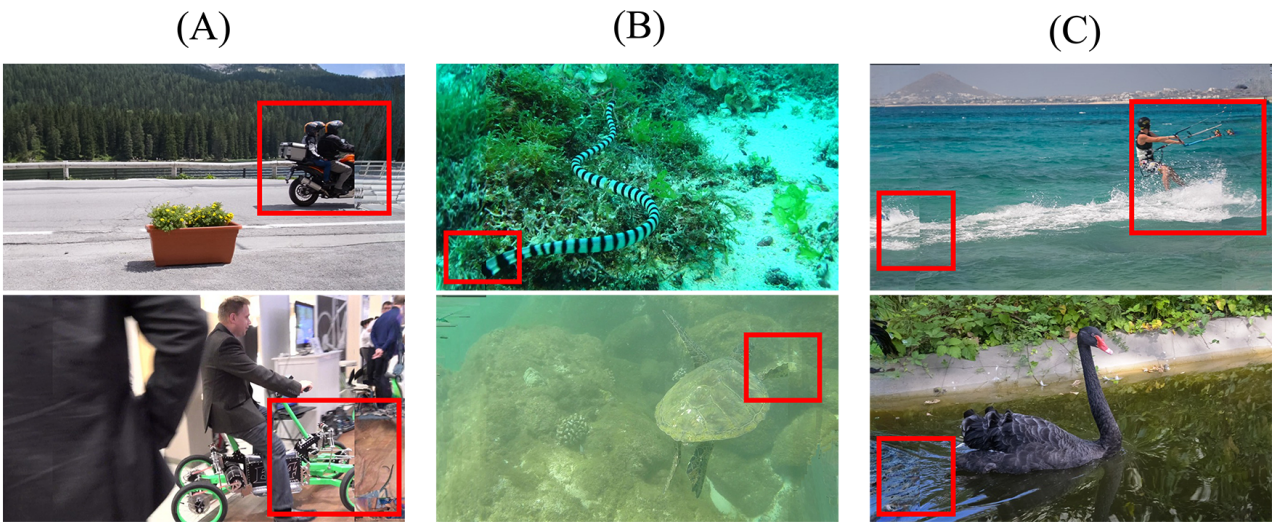

Figure 6: These examples show the cases where the fixed blur classifier misjudged the results as *natural* while the adaptive blur classifier correctly judged them as *unnatural*.

Table 2: The performance enhanced in the order of Raw RGB, Flow RGB, and Flow RGB + Adaptive Blur. The performance of the classifier was improved by using the proposed components.

DAVIS dataset	Accuracy	AP	F1
Raw RGB	80%	0.2	0
Flow RGB	77.333%	0.356	0.409
Flow RGB + Adaptive blur	**88.667%**	**0.536**	**0.655**

Table 3: The accuracy, Average Precision (AP), and F1 score varied depending on the blur width. The total horizontal range of each frame is 224 pixels, with the possibility of a blur width from 0 to 112 pixels from the center. The interval between 80 and 100 demonstrated superior performance.

Width	Accuracy	AP	F1
42	78.667%	0.358	0.444
56	82.667%	0.343	0.33
70	83.333%	0.443	0.54
88	**86.667%**	**0.456**	**0.523**
92	**82%**	**0.393**	**0.523**
96	**86.667%**	**0.473**	**0.583**
112	78.667%	0.213	0

우, 객체의 끝 부분이 경계에서 갑자기 잘려 나간다. 마지막으로 (C)의 경우, 확장 경계에서 파도의 연속성이 끊어져 부자연스러운 모습이 된다. 반면, 적응형 블러는 이를 효과적으로 *unnatural*로 분류하였다.

5 논의

5.1 본 방법의 실패 사례

적응형 블러 방식은 광학 흐름 입력이 비디오의 특징을 정확히 포착하지 못할 때 분류에 어려움을 겪었다. 그림 7에 나타난 것처럼, 자전거를 타고 빠르게 화면 밖으로 이동하는 사람이나 빠르게 날아가는 새와 같이 전경 객체가 프레임에서 갑자기 사라지는 상황에서 광학 흐름은 이를 감지하지 못했다. 프레임이 심하게 흔들리는 경우에도 이와 유사한 문제가 발생했다. 이는 광학 흐름 기술 자체의 한계로 볼 수 있다. 이러한 문제를 해결하기 위해서는 더 높은 해상도와 프레임 레이트의 비디오를 활용하고, 보다 정확한 광학 흐름 추출 방법을 사용하는 것이 도움될 것으로 예상한다.

광학 흐름과 다른 특징 추출 기술을 함께 사용하는 하이브리드 접근 방식을 통해 본 분류기의 성능을 개선할 수 있을 것이다. 한 가지 방안으로, 최신 이미지 분할 모델인 Segment Anything Model(SAM) [25]을 활용하면 프레임의 가장자리에 위치한 주요 관심 객체에 보다 집중할 수 있다. 또한, 비디오 깊이 추정 방법 [26]을 통합하므로써 깊이가 깊은 객체의 움직임에는 낮은 가중치를, 전경에 있는 객체의 움직임에는 높은 가중치를 부여할 수 있다. 이러한 전략을 통해 입력 데이터에 보다 상세한 객체 움직임 정보를 포함시킴으로써 분류기의 정확도와 신뢰도를 높일 수 있을 것으로 기대된다.

Figure 7: Limitations of our method are demonstrated in failure cases where the whole object exits the frame or the frame experiences severe shaking. In such instances, the optical flow fails to accurately detect these features, highlighting the constraints of our classifier.

5.2 본 비디오 분류기의 활용 가능성

Dehan et al. [8]이 제안한 방법에 기반을 둔 본 분류기는 더욱 발전된 비디오 확장 기술이 개발됨에 따라 잠재적 한계에 대한 우려를 가질 수 있다. 그러나 Gao et al. [14]과 같은 다른 흐름 기반 비디오 확장 기법들도 Dehan et al.의 방법과 유사한 특성을 공유하고 있다(그림 8 참조). 이러한 유사성 때문에 흐름 기반 기법들에서 *natural* 혹은 *unnatural* 로 라벨링되는 기준 또한 서로 비슷하다. 이는 본 분류 방식이 Dehan et al.의 방법뿐만 아니라 다른 흐름 기반 비디오 확장 기술에도 적용 가능함을 시사한다. 따라서 향후 더욱 발전된 흐름 기반 확장 기법이 등장하더라도, 그 기법이 Dehan et al.의 방법과 유사한 실패 사례를 보인다면 본 분류기를 보편적으로 적용 가능할 것으로 예상된다.

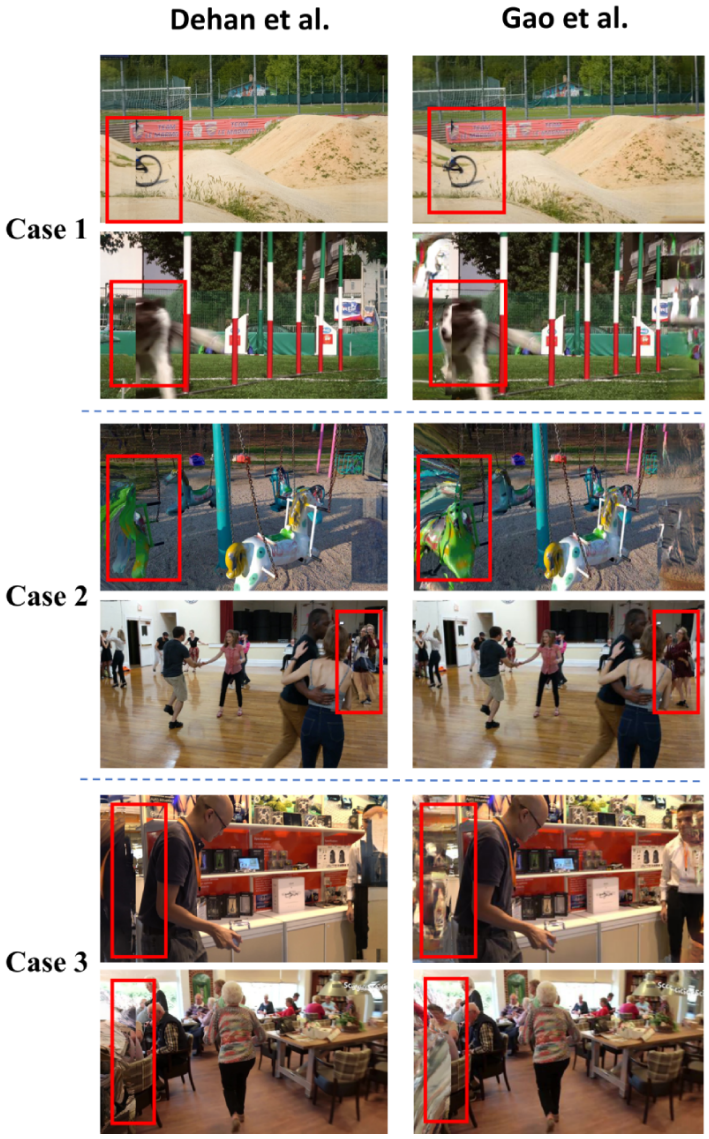

Figure 8: Flow based extrapolation methods Dehan et al. [8] and Gao et al. [14] share similar failure cases. The scenarios of failure cases are the same as those mentioned in Figure 2.

6 결론

이 논문은 흐름 기반 비디오 확장 방식을 적용하기에 앞서 주어진 비디오가 해당 방식에 적합한지 여부를 판단하는 비디오 분류기를 제안하였다. 이는 비디오 확장 분야에서 기존에 다루지 않았던 과제를 해결하는 데 중요한 진전을 보여준다. 이 분류기는 광학 흐름과 적응형 가우시안 블러 네트워크를 활용하여 효과적으로 확장 가능한 비디오를 분류한다.

제안된 분류기의 타당성을 검증하기 위해, 널리 사용되는 DAVIS 데이터셋을 활용하여 각 비디오 클립의 확장 시 자연스러움 여부를 나타내는 레이블링을 수행하였다. 학습에 사용된 영화 데이터셋 또한 동일한 기준으로 레이블링되었다. 분류기는 사용자 테스트를 통해 도출된 확장 실패의 특징을 반영하여 광학 흐름을 입력으로 사용하였으며, 인간의 지각 특성에 기반한 적응형 가우시안 블러 네트워크를 통해 성능을 개선하였다. 그 결과, 제안된 분류기는 기존 방식 대비 우수한 성능을 달성하였다.

이 논문에서 제시한 비디오 분류기는 광학 흐름과 적응형 가우시안 블러 네트워크를 결합함으로써 고품질 확장에 적합한 비디오를 실용적이고 효과적으로 식별할 수 있게 해준다. 그에 따라 몰입감 있는 시청 경험을 위해 장면을 자동으로 확장하는 기술과 함께 사용될 수 있을 것으로 기대된다.

7 감사의 글

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2023년도 문화체육관광 연구개발사업으로 수행되었음 (과제명: 아바타 개성 표현을 위한 유니버설 패션 창작 플랫폼 기술개발, 과제번호: RS-

2023-00228331, 기여율: 100%)

References

[1] Jungjin Lee, Sangwoo Lee, Younghui Kim, and Junyong Noh. Screenx: Public immersive theatres with uniform movie viewing experiences. *IEEE transactions on visualization and computer graphics*, 23(2):1124–1138, 2016.

[2] Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. Foveated 3d graphics. *ACM transactions on Graphics (tOG)*, 31(6):1–10, 2012.

[3] Sanghoon Lee, Marios S Pattichis, and Alan C Bovik. Foveated video quality assessment. *IEEE Transactions on Multimedia*, 4(1):129–132, 2002.

[4] Cornelius Weber and Jochen Triesch. Implementations and implications of foveated vision. *Recent Patents on Computer Science*, 2(1):75–85, 2009.

[5] Mohammed Yeasin and Rajeev Sharma. Foveated vision sensor and image processing–a review. *Machine Learning and Robot Perception*, pages 57–98, 2005.

[6] David V Wick, Ty Martinez, Sergio R Restaino, and BR Stone. Foveated imaging demonstration. *Optics Express*, 10(1):60–65, 2002.

[7] Zhou Wang and Alan C Bovik. Foveated image and video coding. *Digital Video, Image Quality and Perceptual Coding*, pages 431–457, 2006.

[8] Loïc Dehan, Wiebe Van Ranst, Patrick Vandewalle, and Toon Goedemé. Complete and temporally consistent video outpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 687–695, 2022.

[9] Amit Aides, Tamar Avraham, and Yoav Y Schechner. Multiscale ultrawide foveated video extrapolation. In *2011 IEEE International Conference on Computational Photography (ICCP)*, pages 1–8. IEEE, 2011.

[10] Tamar Avraham and Yoav Y Schechner. Ultrawide foveated video extrapolation. *IEEE Journal of Selected Topics in Signal Processing*, 5(2):321–334, 2010.

[11] Sangwoo Lee, Jungjin Lee, Bumki Kim, Kyehyun Kim, and Junyong Noh. Video extrapolation using neighboring frames. *ACM Transactions on Graphics (TOG)*, 38(3):1–13, 2019.

[12] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[13] Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A survey of structure from motion*. *Acta Numerica*, 26:305–364, 2017.

[14] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 713–729. Springer, 2020.

[15] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

[16] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[17] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[18] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[19] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[20] Xin Liu, Silvia L Pintea, Fatemeh Karimi Nejadasl, Olaf Booij, and Jan C Van Gemert. No frame left behind: Full video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14892–14901, 2021.

[21] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference*

on computer vision and pattern recognition, pages 724–732, 2016.

[22] Emre Akbas and Miguel P Eckstein. Object detection through search with a foveated visual system. *PLoS computational biology*, 13(10):e1005743, 2017.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[24] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019.

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[26] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics (ToG)*, 39(4):71–1, 2020.

〈저자소개〉

김민선
- 2019 ~ 2024 대구경북과학기술원 기초학부 학사
- 2024 ~ 현재 한국과학기술원 문화기술대학원 석사과정
- https://orcid.org/0009-0009-8212-4982

서창욱
- 2011 ~ 2017 교토예술대학 예술학부 예술학사
- 2019 ~ 2021 서강대학교 미디어공학 공학석사
- 2021 ~ 2024 한국과학기술원 문화기술대학원 공학박사
- 2024 ~ 현재 애니그마테크놀로지스 Principal researcher
- https://orcid.org/0000-0002-3809-9515

윤현호
- 2009 ~ 2013 일리노이대 어바나-샴페인 원자핵공학 학사
- 2014 ~ 2015 맨체스터대 전기전자공학 석사
- 2022 ~ 2024 한국과학기술원 문화기술대학원 공학석사
- https://orcid.org/0000-0002-7723-8143

노준용
- 2002 University of Southern California 전산학 박사 취득
- 2003 ~ 2006 Rhythm and Hues Studios 그래픽스 사이언티스트 재직
- 2006 KAIST 문화기술대학원 교수 부임
- 2011 KAIST 문화기술대학원 석좌교수 추대
- 2016 ~ 2020 KAIST 문화기술대학원장
- 2022 ~ 2023 KAIST 실패연구소장
- https://orcid.org/0000-0003-1925-3326