Journal of KIIT. Vol. 23, No. 2, pp. 1-10, Feb. 28, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.2.1

교통사고 감지를 위한 Inflated 3D ConvNet 및 Transformer 기반 향상된 시공간 특징 탐지 모델

윤성안*, 이은성**, 조정호***

Enhanced Spatiotemporal Feature Detection Model based on an Inflated 3D ConvNet and Transformers for Traffic Accident Detection

Sung-An Yoon*, Eunsung Lee**, and Jeongho Cho***

이 논문은 2024년도 정부(교육부)의 재원으로 한국연구재단 기초연구사업(No.2021R1I1A3055973)의 지원을 받아 수행되었으며 2023학년도 순천향대학교 교수 연구년제에 의하여 연구하였음

요 약

영상 기반 모니터링 시스템은 도로 교통 관리, 특히 교통사고 감지에서 중요한 역할을 하지만, 여전히 만족할 만한 성능을 보이지 못하고 있다. 본 연구에서는 보다 정확한 교통사고 감지와 오탐지 감소를 위해 I3D(Inflated 3D ConvNet)와 MHSA(Multi Head Self-Attention) 기반 Transformer 분류기를 결합한 지도학습 기반 이상 상황 탐지 모델을 제안한다. 제안된 기법은 입력 비디오를 세그먼트로 나누고, I3D를 통해 각 세그먼트의 시공간 특징을 추출한 뒤, MHSA로 세그먼트 간 관계를 학습하여 비정상 상황에 대한 특징을 강화하고, 최종적으로 MLP(Multi-Layer Perceptron)을 통해 정상과 비정상 세그먼트를 분류한다. 이를 통해 기존 방식의 정상과 비정상 이벤트의 혼재 문제를 해결하였으며, 기존 접근 방식 대비 최대 5.75% 향상된 AUC 성능을 확인함으로써 제안된 모델의 우수성을 입증하였다.

Abstract

Video-based monitoring systems play a critical role in various fields, including road traffic management, particularly in the detection of traffic accidents. However, current systems still fall short of achieving satisfactory performance. This study proposes a novel supervised anomaly detection model that combines Inflated 3D ConvNet (I3D) and a Transformer classifier based on Multi Head Self-Attention(MHSA) to enhance traffic accident detection accuracy and reduce false positives. The proposed method segments the input video and extracts spatiotemporal features of each segment through I3D. Subsequently, MHSA is applied to learn inter-segment relationships, thereby emphasizing features associated with anomalous situations, and finally, a Multi-Layer Perceptron(MLP) classifies segments as normal or abnormal. This approach addresses the issue of intermingling normal and abnormal events inherent in traditional anomaly detection methods and demonstrates the model's effectiveness, achieving up to a 5.75% improvement in AUC performance over existing approaches.

Keywords
deep learning, anomaly detection, traffic accident detection, transformer

* 순천향대학교 전기통신시스템공학과
- ORCID: https://orcid.org/0009-0006-7984-4813
** 한국항공우주연구원 위성항법연구부
- ORCID: https://orcid.org/0000-0003-4565-5863
*** 순천향대학교 전기공학과(교신저자)
- ORCID: https://orcid.org/0000-0001-5162-1745

· Received: Oct. 21, 2024, Revised: Nov. 07, 2024, Accepted: Nov. 10, 2024
· Corresponding Author: Jeongho Cho
Dept. of Electrical Engineering, Soonchunhyang University, Korea
Tel.:+82-41-353-5698, Email: jcho@sch.ac.kr

Ⅰ. 서 론

영상 기반 모니터링 시스템은 실시간으로 데이터를 수집 및 분석하여 즉각적인 대응이 가능하며, 넓은 범위의 지역을 동시에 감지할 수 있어 보안 및 감시 분야에서 널리 사용된다. 특히, 도로 교통 관리 시스템에서는 영상 기반 모니터링 시스템을 도입함으로써 도로 안전을 강화하고, 교통사고를 신속하게 감지하는 중요한 역할을 수행하고 있다[1][2]. 교통사고의 신속한 감지는 인명 피해를 최소화하고 도로 교통의 원활함을 보장하는 데 중요한 역할을 한다. 기존의 사고 감지 방법은 주로 CCTV 감시나 목격자의 신고에 의존하기 때문에 CCTV 감시자의 부주의나, 차량의 유동량이 적은 경우 목격자의 부재로 인해 사고 감지가 지연될 수 있는 한계가 있다[3]. 이러한 문제를 해결하기 위해, 비디오 클립에서 정상 클립과 비정상 클립을 구분하여 이상 여부를 판단하는 영상 기반 이상치 탐지 기법이 연구되어왔다[4][5]. 최근에는 딥러닝의 비약적인 발전을 통해, 딥러닝 기반 이상치 탐지 모델이 등장하면서 사고 감지의 정확성과 신속성이 크게 향상되었다.

교통사고와 같은 이상 현상은 정상과 비정상 현상이 혼재되어 있고 경계가 모호하기 때문에 라벨링 작업에 많은 시간과 인력이 요구된다. 따라서 딥러닝 기반 이상치 탐지 기법은 일반적으로 약한 지도학습 기반 방식과 재구성 및 예측 기반 방식으로 이상 현상을 탐지한다. 약한 지도학습 기반 방식은 비디오 수준에서 라벨을 부여하고, 비정상 클립 내 세그먼트를 특정 기준에 따라 정상과 비정상으로 분류해 이상치를 탐지하는 방식이다. 이 방식은 시간 소모적인 라벨링 작업을 단순화 하였지만, 학습 중 비정상 세그먼트가 정상 세그먼트로 잘못 분류될 수 있는 단점이 있다. 재구성 및 예측 기반 방식은 정상 세그먼트를 입력받아 Auto-Encoder 기반 모델을 통해 입력된 세그먼트를 재구성하거나 다음 프레임을 예측하여 실제 프레임과의 오차를 통해 이상치를 탐지하는 방식이다. 이 방식은 정상 클립만을 학습에 사용하기 때문에 데이터 확보 문제와 라벨링 문제를 해결하였지만, 모델이 과도하게 학습하여 비정상 세그먼트도 재구성할 수 있다는 단점이 있다.

본 연구는 교통사고를 정확하게 탐지하기 위해 I3D(Inflated 3D ConvNet)[6]와 MHSA(Multi Head Self-Attention)기반 Transformer 분류기[7]를 결합한 새로운 지도학습 기반 이상치 탐지 모델을 제안한다. 제안하는 기법은 입력 비디오를 여러 세그먼트로 나눈 후, 각 세그먼트를 I3D를 사용해 특징 벡터로 변환한다. 변환된 특징 벡터는 MHSA를 통해 세그먼트 간의 관계를 학습하여 더욱 정교한 특징 벡터로 강화되며, 이후 MLP(Multi-Layer Perceptron)을 사용해 정상 세그먼트와 비정상 세그먼트를 분류한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 딥러닝 기반 이상치 탐지 기법에 관한 관련 연구를 소개하고, 3장에서는 본 연구에서 제안하는 기법에 대해 상세히 설명한다. 4장에서는 제안된 기법의 효과를 검증하기 위한 실험과 그 결과를 제시, 분석하며 5장에서는 본 논문의 결론과 향후 연구 방향을 서술한다.

Ⅱ. 관련 연구

딥러닝 기술은 영상 기반 모니터링에서 필수적인 도구이다. 일반적으로, 딥러닝 기반 이상치 탐지 기법은 비디오를 일정 간격의 세그먼트로 나눈 뒤, 각 세그먼트를 정상과 비정상으로 분류하는 방식이다. W. Sultani et al.[8]은 C3D(Convolutional 3-dimensional Network)을 통해 세그먼트 수준에서 이상 점수를 생성하고, 정상 비디오와 비정상 비디오에서 각각 최고 점수를 나타내는 세그먼트를 추출하여 정상과 비정상 세그먼트로 학습에 사용함으로써, 세그먼트에 대한 라벨 없이도 학습이 가능하도록 하였다. 그러나 비정상 비디오 내 정상 세그먼트가 비정상 세그먼트로 잘못 선택되는 단점이 있다. Y. Tian et al.[9]은 비정상 세그먼트의 특징 벡터 크기가 정상보다 높다는 점에 착안하여, 특징 추출기에서 추출된 비정상 세그먼트 중 top-K 개의 세그먼트를 선택하고 정상과 비정상을 구분하는 RTFM(Robust Temporal Feature Magnitude)을 제안했다. 하지만 복잡한 배경이나 환경 변화가 많은 경우, 정상 특징 벡터의 크기가 커져 잘못된 세그먼트가 선택되는 문제가 발생하는 단점이 있다.

Y. Chen et al.[10]은 Magnitude Contrastive Loss를 도입하여, 정상과 비정상 세그먼트에서 각각 top-K 개의 특징 벡터를 선택하고, 같은 라벨은 가깝게, 서로 다른 라벨은 멀리 배치하여 정상과 비정상을 구분하였다. 그러나 특징 벡터를 선택하는 과정에서 여전히, 정상과 비정상이 혼재되는 단점이 존재한다. Zhou et al.[11]은 비정상 비디오에서 비정상 세그먼트는 극히 일부분이기 때문에 전체 세그먼트의 평균은 중심 극한 정리에 의해 정상 세그먼트의 대한 정규 분포를 따르게 된다는 점을 활용한 BN-WVAD(BatchNorm-based Weakly Supervised Video Anomaly Detection)를 제안하였다. 이 방법은 명확한 기준을 통해 정상과 비정상을 구분하였으나, 비정상 비디오에서 정상 세그먼트 비율이 낮을 경우, 평균값이 정상 분포에서 벗어나 구분도가 낮아지는 한계가 있다. 약한 지도학습 기반 방식의 정상과 비정상이 혼재하는 단점을 해결하기 위해 Y. Li et al.[12]은 U-Net[13]과 Conv-LSTM(Convolutional Long Short-Term Memory)[14]을 결합한 형태의 비디오 이상 탐지 네트워크를 제안하였다. 제안된 네트워크는 연속된 프레임들을 입력받아 다음 프레임을 재구성하여 생기는 실제 프레임과의 오차를 통해 이상치를 탐지하였다. 학습에는 정상 비디오만 사용함으로써 정상 세그먼트가 잘못 선택되는 단점을 해결하였지만, 입력된 프레임의 길이가 길어질수록 Conv-LSTM의 기울기 소실 문제로 인해 긴 이상 현상에 대한 탐지 성능이 저하되는 단점이 있다. H. Yuan et al.[15]은 U-Net에 Conv-LSTM 대신 ViViT(Video Vision Transformer)[16]를 결합한 형태를 제안하였다. B. Guo et al.[17]는 ViViT의 Factorized 인코더를 활용한 VadVVT(Video Anomaly Detection with Video Vision Transformer)를 제안했다. [15][17]은 Transformer를 도입함으로써 프레임 사이의 시간적 관계를 학습하여 긴 이상 현상에 대한 탐지 성능을 향상시켰지만, 과도하게 학습된 모델이 비정상 세그먼트도 재구성할 수 있고 학습에 정상 비디오만 사용되기 때문에 정상과 비정상 데이터를 구분하는 능력이 제한되어 학습되지 않은 상황에서는 탐지 성능이 떨어지는 한계가 있다.

III. 교통사고 감지를 위한 시공간 특징 탐지 모델

일반적으로 이상치 탐지 모델은 정상 상황과 이상 상황 사이의 모호함으로 인해 약한 지도학습과 재구성 및 예측 기반 방식에 집중되어왔다. 위와 같은 문제 해결을 위해 약한 지도학습은 비디오 수준의 라벨링과 특징 기반 세그먼트 선택 알고리즘을 사용하여 정상과 비정상을 구분하였지만, 정상 특징이 비정상으로 선택되는 단점이 존재했다. 재구성 및 예측 기반 방식은 정상 비디오의 여러 프레임을 통해 다음 프레임을 재구성하여 실제 프레임과의 차이를 통해 비정상 프레임을 분류하였지만, 과도하게 학습된 모델에 의해 비정상 프레임이 높은 정확도로 재구성되는 단점이 존재했다. 이에 따라 본 연구에서는 교통사고의 발생 지점 이후를 모두 비정상으로 간주하고 라벨링 하여 정상과 비정상 세그먼트를 탐지하는 지도학습 기반의 이상 상황 탐지 모델을 제안한다. 제안하는 모델의 블록 다이어그램은 그림 1과 같다.

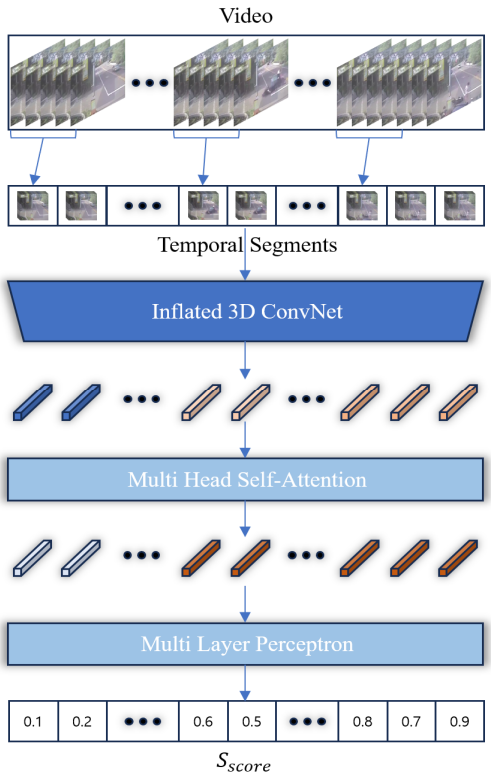

그림 1. 제안하는 모델의 블록 다이어그램
Fig. 1. Block diagram of the proposed model

교통사고와 같은 이상 상황은 여러 프레임에 걸쳐 발생하며, 사고 발생 시점의 앞뒤 맥락이 매우 중요하다. 이를 효과적으로 파악하기 위해, 먼저 비디오를 시간 축에 따라 여러 개의 시간적 세그먼트로 나누고, 각 세그먼트를 I3D에 입력하여 특징 벡터로 변환한다. I3D는 각 세그먼트를 다중 프레임을 입력받아 특징을 추출하기 때문에 시공간적 맥락을 반영한 특징 벡터가 된다. 그러나 일부 세그먼트는 비정상 프레임이 적게 포함되어 정상 특징 벡터와 유사하게 나타날 수 있다. 이를 해결하기 위해, MHSA를 통해 특징 벡터들 간의 관계를 학습하여 비정상 특징을 강조한다. 마지막으로, MLP를 사용해 정상과 비정상 세그먼트를 최종적으로 분류한다.

3.1 시공간 특징 벡터 추출 및 강화

일반적인 비디오 분류 모델은 비디오의 앞뒤 맥락을 파악하기 위해 주로 2D 컨볼루션과 RNN(Recurrent Neural Network)이 결합된 구조 또는 3D 컨볼루션을 사용하는 신경망 구조로 이루어져 있다. 2D 컨볼루션은 각 프레임의 공간적 정보를 효과적으로 추출할 수 있으나, 시간적 연속성을 처리하는 데 한계가 있어 이를 보완하기 위해 RNN을 함께 사용한다. 반면, 3D 컨볼루션은 시간 축을 포함한 다차원 특징을 추출할 수 있어 시공간적 맥락을 더 효과적으로 반영할 수 있다. 하지만 3D 컨볼루션은 계산 비용이 매우 높다는 단점이 있다. 이러한 문제를 해결하기 위해 J. Carreira et al.[6]은 기존의 2D 컨볼루션 필터를 3차원으로 확장한 구조의 I3D를 제안하였다. 2D 컨볼루션 필터($f_{2D} \in R^{n \times n}$)는 입력 이미지 $I \in R^{H \times W \times C}$와의 컨볼루션 연산을 통해 공간적 특징을 추출한다. 여기서 n은 필터의 크기이고 H, W, C는 각각 입력 이미지의 높이, 너비, 색상 채널이다.

그러나 비디오 분류에서 입력된 비디오 세그먼트는 시간 축이 포함된 형태인 $v \in R^{H \times W \times C \times T}$의 크기를 가진다. 기존의 2D 컨볼루션 필터를 비디오에 적용하기 위해, 2D 컨볼루션 필터의 가중치를 복사하여 3차원의 3D 컨볼루션 필터($f_{3D} \in R^{n \times n \times p}$)로 확장한다. 여기서 p는 필터가 한 번에 처리하는 프레임의 수를 나타내며, 3D 컨볼루션 필터는 p개의 프레임과 컨볼루션 연산을 통해 시공간 특징($S_s \in R^{1 \times d}$)을 추출한다. 이를 통해 기존의 ImageNet과 같은 대규모 2D 데이터 셋으로 사전 학습된 모델의 가중치를 비디오 분류에 적용할 수 있으며, 이는 교통사고와 같은 소규모 데이터 셋에서 높은 정확도를 달성할 수 있는 장점을 제공한다.

3.2 Transformer 분류기

제안하는 모델은 학습 시 입력된 비디오를 N개의 세그먼트로 나누고 I3D를 통해 각 세그먼트에서 시공간 특징을 추출한다. 비디오를 세그먼트로 나누는 과정에서 각 세그먼트가 포함하는 비정상 프레임의 비율이 다르며, 비율이 낮은 세그먼트는 정상 세그먼트와 유사한 시공간 특징을 가질 수 있다. 이를 효과를 보완하기 위해 본 연구에서는 MHSA 기반의 Transformer 인코더를 통해 시공간 특징을 강화한다. Transformer 인코더는 여러 층의 신경망 구조로 각 층은 MHSA과 FFN(Feed Forward Network)으로 이루어져 있다.

MHSA는 Self-Attention 알고리즘을 여러 개 사용하여 입력 데이터를 Query, Key, Value 세 가지로 변환한 후, 각 Value의 가중치를 계산하여 Query와 Key의 유사도를 결정하고 데이터 간의 관계를 모델링하는 알고리즘이다. MHSA의 핵심은 Self-Attention으로, 그림 2와 같이 세그먼트 집합($S = S_1, S_2, ..., S_s \in R^{s \times d}$)에 대해 G_q, G_k, G_v로 이루어진 Attention Head와의 곱셈을 통해 Q, K, V 세 가지 벡터로 변환한다. 여기서 G_q, G_k, G_v는 가중치 행렬을 의미하며, 각각 $d \times h$크기를 가진다. 다음으로, Q와 K의 벡터 내적을 통해 데이터 간의 유사도를 나타내는 유사도 행렬을 구성하고 $Softmax$을 통해 유사도 점수($Score_{att}$)을 다음과 같이 계산한다.

$$Score_{att} = Softmax\left(\frac{Q \cdot K'}{\sqrt{h}}\right) = \frac{\exp\left(\frac{Q \cdot K'}{\sqrt{h}}\right)_{lk}}{\sum_{m=1}^{s} \exp\left(\frac{Q \cdot K'}{\sqrt{h}}\right)_{lm}} \quad (1)$$

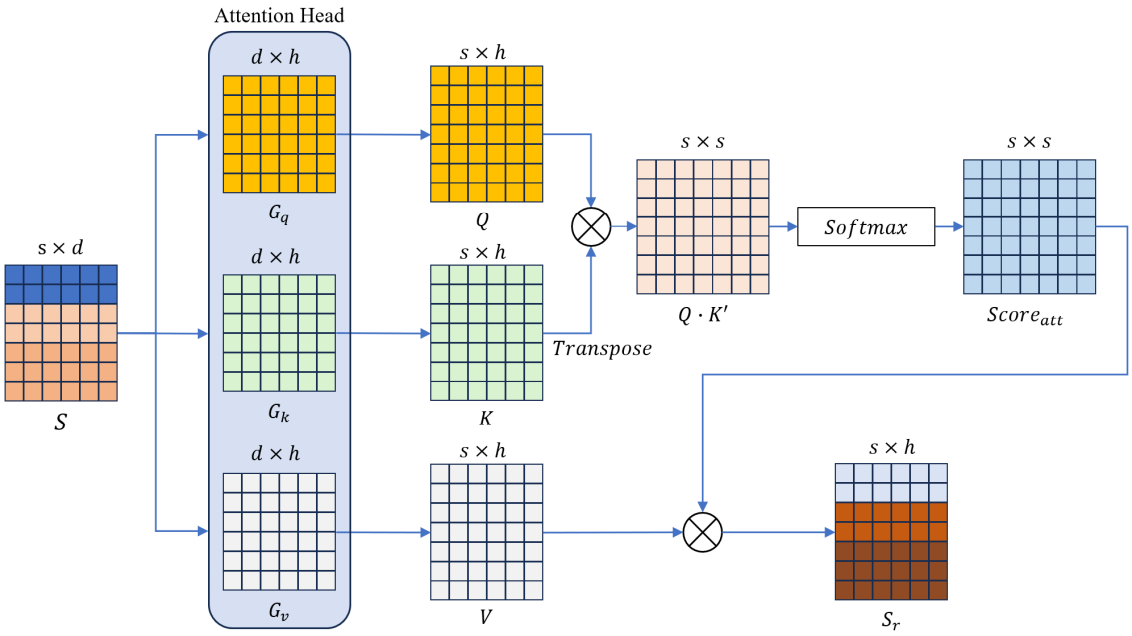

그림 2. Self-Attention 알고리즘
Fig. 2. Self-Attention algorithm

여기서 $Q \cdot K'$는 Q와 K'간의 내적을 통해 유사도 행렬을 구하는 과정이며, $\sqrt{h}$는 값의 크기를 스케일링하여 안정적인 학습을 보장한다. l,k는 유사도 행렬의 행과 열을 의미한다. s는 데이터의 총 개수를 나타내며, m은 각 l행에 대한 특정 열을 의미한다. $Softmax$을 통해 입력 데이터는 각 행에 대해 유사도 점수를 확률 분포로 변환하여 상대적인 중요도를 할당한다. 최종적으로 V와의 가중 합을 통해 입력 데이터 간의 관계를 학습하게 된다. MHSA는 Self-Attention 과정에서 가중치 행렬을 헤드 수만큼 나누어 병렬로 사용한 구조로 각 헤드는 독립적인 가중치 행렬을 가지게 된다. 이로 인해 헤드 별로 서로 다른 특징을 추출하게 되면서 다양한 특징을 학습하게 된다.

FFN은 앞서 추출된 특징 벡터($S_r \in R^{s \times h}$)에 비선형성을 추가하기 위한 네트워크로 두 번의 선형 변환과 정규화를 거쳐 특징을 더욱 강화한다. FFN의 과정은 다음과 같다.

$$S_{r1} = ReLU(S_r G_1 + b_1) \tag{2}$$

$$S_{r2} = S_{r1} G_2 + b_2 \tag{3}$$

$$S_{r3} = LayerNorm(S_r + S_{r2}) \tag{4}$$

여기서 G_1, G_2는 선형 변환을 위한 가중치이고 b_1, b_2는 바이어스이다. $ReLU$은 비선형성을 추가하기 위한 활성화 함수이다. $LayerNorm$은 입력된 벡터의 평균과 분산으로 각 요소에 대해 정규화 하는 함수로, 출력이 급격히 높아지는 현상을 방지하여 학습이 안정적으로 진행되게 하는 역할을 한다.

최종적으로 강화된 특징 벡터 S_{r3}는 다음과 같이 MLP를 통해 정상과 비정상 세그먼트로 분류된다.

$$S_{r4} = ReLU(S_{r3} G_3 + b_3) \tag{5}$$

$$S_{score} = Sigmoid(S_{r4} G_4 + b_4) \tag{6}$$

Ⅳ. 실험 및 성능 평가

4.1 실험 환경

제안된 기술은 NVIDIA RTX 3060, Intel Core i7-12700F CPU에서 구현되었으며, 이상 탐지 데이터 셋인 UCF-Crime[8]에서 교통사고 영상을 사용하여 데이터 셋을 구성하여 평가하였다. 데이터 셋은 총 62개의 학습 데이터와 23개의 테스트 데이터로 이루어졌다.

제안한 모델의 비디오 세그먼트 수는 32개이며, I3D의 입력 프레임 수는 16프레임이다. 이상치 탐지 성능은 AUC(Area Under Curve)와 Youden's Index[18]를 통한 기준 값 설정을 진행한 후, F1-Score, Recall, Precision을 통해 확인하였다.

AUC는 ROC(Receiver Operating Characteristic) 곡선의 아래 면적을 의미하며, ROC는 서로 다른 기준 값에서 TPR(True Positive Rate)과 FPR(False Positive Rate)의 비율을 나타내는 곡선을 의미한다. AUC는 모델의 전반적인 성능을 평가하는 중요한 지표로, 높은 AUC 값은 모델이 다양한 임계값에서 우수한 분류 성능을 보여준다는 것을 의미한다. 하지만 실제 운용되는 교통사고 모니터링 시스템의 경우 특정 임계값에서 동작하기 때문에 특정 임계값에서의 성능 또한 중요하다. 따라서 본 논문에서는 실제 운영 시 모델의 탐지 성능을 확인하기 위해 Youden's Index를 통해 기준 값을 설정하여 모델을 추가 검증하였다. Youden's Index(J)는 특정 임계값에서 이진 분류 문제에서 분류기의 성능을 평가하는 지표로 Sensitivity(Se)와 Specificity(Sp)를 결합하여 계산되며 다음과 같다.

$$J = Se + Sp - 1 \tag{7}$$

여기서 Se는 실제로 긍정인 샘플 중에서 올바르게 긍정을 예측한 비율로 $\frac{TP}{TP+FN}$로 계산된다. Sp는 실제로 부정인 샘플 중에서 올바르게 부정을 예측한 비율로 $\frac{TN}{TN+FP}$로 계산된다. 여기서 TP, TN,FP는 각각 True Positive, True Negative, False Positive를 의미한다. 결론적으로 Youden's Index는 Sensitivity와 Specificity를 모두 고려한 분류기의 균형 잡힌 성능을 의미하며, 값이 낮을수록 모델의 추측이 한쪽으로 치우쳐 있음을 의미한다. 본 논문에서는 모델의 모든 출력 값을 임계값으로 설정하여 J를 추출하고 가장 높은 J를 가지는 임계값을 이상치 판단 기준 값으로 설정하고 F1-Score, Recall, Precision을 측정하여 실제 응용 시 이상치 탐지 성능을 확인하였다.

F1-Score는 Recall과 Precision의 조화평균으로 $2\times\frac{Recall\times Precision}{Recall+Precision}$로 계산된다. 즉, F1-Score는 Recall과 Precision의 균형을 의미하며, 모두 높은 모델일수록 높은 값이 도출된다. Recall은 탐지해야 하는 이상치 중 제대로 탐지된 비율이며 Se와 같다. Precision은 교통사고로 탐지한 결과 중 옳은 탐지 비율을 의미하며 $\frac{TP}{TP+FP}$로 계산된다.

4.2 이상치 탐지 결과 비교

본 연구에서는 교통사고의 정확한 탐지와 낮은 오탐률을 줄이기 위해 다중 프레임을 입력받는 I3D와 Transformer 분류기를 결합한 시공간 특징 탐지 모델을 제안하였다. 제안한 모델의 성능 평가를 위해 최신 접근 방식인 C3D_MIL[8], MGFN[10], BN-WVAD[11]와 비교하였으며 표 1과 같이 요약하였다.

표 1. 접근 방식에 따른 이상치 탐지 결과 비교
Table 1. Comparison of anomaly detection performance evaluations by approach method

Method	AUC	Precision	Recall	F1-Score
C3D_MIL	65.82	0.3319	0.6411	0.43
MGFN	64.67	0.3796	0.6098	0.46
BN-WVAD	63.47	0.3872	0.6331	0.48
Proposed model	69.22	0.4358	0.6402	0.51

AUC 성능 평가 결과, 제안하는 모델은 69.22로 가장 높은 성능을 보여주었으며, 기존 기법인 BN-WVAD 대비 최대 5.75% 높은 성능을 보여주었다. 이는 모든 임계값에서 일관되게 정상과 비정상을 구분할 수 있음을 나타내고 결론적으로 실제 시스템 적용 시 일관된 결과를 제공할 수 있음을 의미한다.

F1-Score 및 Precision 평가에서도 제안된 모델은 각각 최대 3%, 10% 더 높은 성능을 보여주었으며, 이는 실제 시스템에서 운용 시 제안된 모델이 높은 탐지 성능과 낮은 오탐률을 보장할 수 있음을 의미한다. Recall의 경우, 제안된 모델은 0.6402로 C3D_MIL과 비슷한 성능을 보였으나, C3D_MIL은 Precision이 낮아 높은 오탐률을 가진다.

교통사고 감지 시스템에서 오탐률은 높을 경우 거짓 경고로 인해 운영자의 피로감을 유발하고 불필요한 자원 소모로 인해 실제 사고 대응이 지연될 수 있다. 제안된 모델은 유사한 탐지 성능을 보장함과 동시에 C3D_MIL 대비 약 10% 높은 Precision으로 낮은 오탐률을 보여주었다. 최종적으로 제안된 모델은 실제 시스템 적용 시 높은 신뢰성과 정확한 사고 탐지가 가능함을 확인하였다.

그림 3은 교통사고 탐지 결과를 예시로 보여준다. 1열부터 3열까지는 각 비디오 클립에 대해 세그먼트별로 예측된 이상 상황 점수를 나타낸다. 각 그래프의 x축은 비디오 클립의 프레임 수를, y축은 이상 상황 점수를 나타낸다. 빨간색 선은 Ground-Truth로 0은 정상 상황을, 1일 경우 비정상 상황을 의미한다. 파란색 선은 모델이 예측한 이상 상황 점수이며, 초록색 선은 Youden's Index를 통한 이상 상황 판단의 기준 값으로 점수가 기준 값을 초과하면 이상 상황으로 판단하게 된다. 그림 3의 각 행은 위에서부터 C3D_MIL, MGFN, BN-WVAD, 제안된 모델의 이상 상황 점수를 나타내며, 마지막 행은 각 비디오의 정상 및 비정상 상황을 보여주는 샘플 이미지이다.

그림 3의 첫 번째 열은 CCTV와 비교적 먼 거리에서 차량 사고가 일어난 상황이다. MGFN과 BN-WVAD는 특징 벡터의 크기를 기준으로 정상과 비정상 세그먼트를 선택하는 방식이다. 이러한 방식은 유동 차량의 양이 많을 경우 특징 벡터의 크기가 증가하면서 정상적인 차량 움직임이 비정상으로 잘못 인식될 수 있으며, 이로 인해 오탐지가 발생하는 것을 확인할 수 있다. 반면, 제안하는 모델은 지도학습 기반으로 학습 과정에서 정상과 비정상 세그먼트를 명확하게 학습하였기 때문에 보다 정확하게 정상과 비정상을 구분함을 확인할 수 있다.

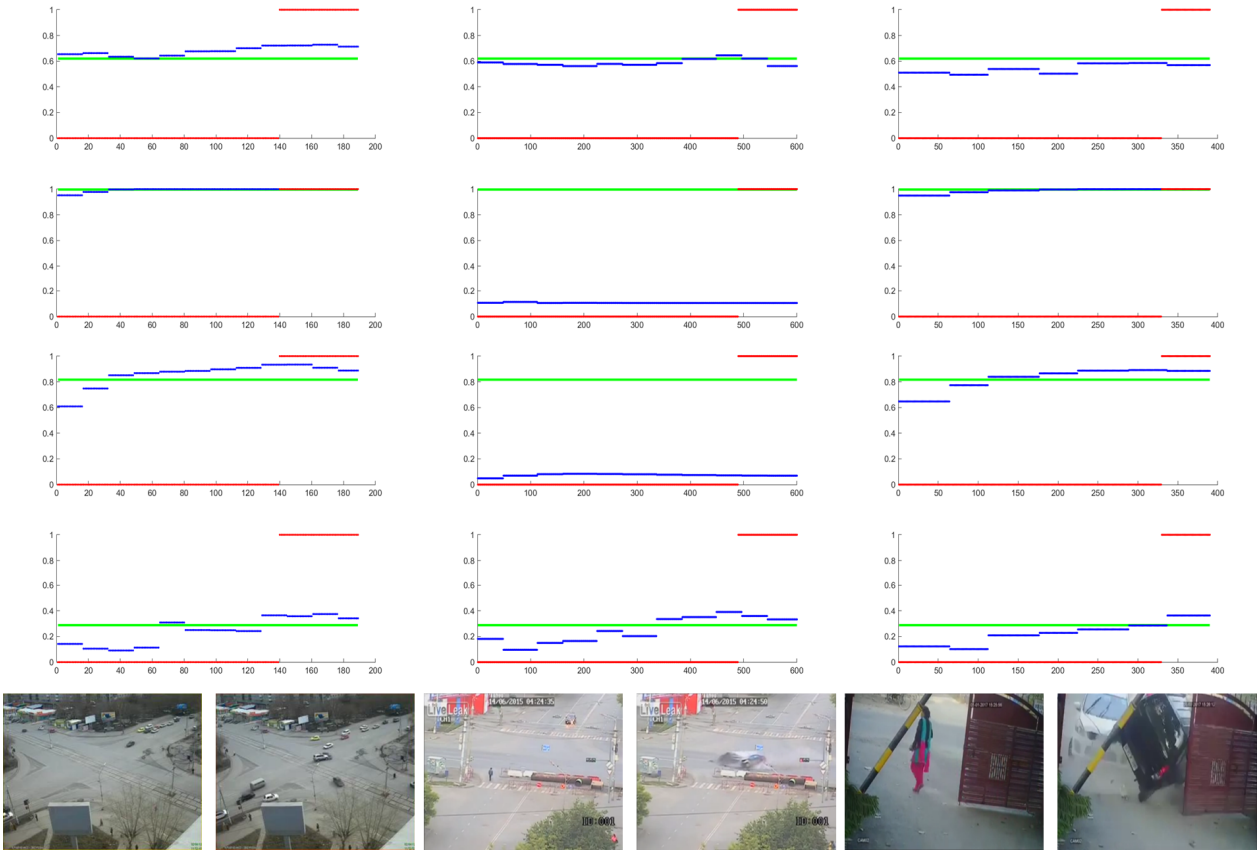

그림 3. 교통사고 탐지 결과 (위에서부터 C3D_MIL, MGFN, BN-WVAD, Proposed model)
빨간색: Ground-Truth, 파란색: 이상치 점수, 초록색: Youden's 기준 값
Fig. 3. Traffic accidents detection results (from the top, C3D_MIL, MGFN, BN-WVAD, Proposed model)
Red: Ground-Truth, Blue: Abnormal score, Green: Youden's threshold

두 번째 열은 비교적 먼 거리에서 사고 차량이 빠르게 화면에서 사라진 경우다. 이 경우 사고 장면이 짧게 노출되고 비정상 패턴이 약하게 포착되어 특징 벡터의 크기가 비교적 작아질 수 있다. MGFN과 BN-WVAD는 특징 벡터의 크기만을 기준으로 정상과 비정상을 구분하기 때문에 이러한 상황에서는 정상 세그먼트로 잘못 분류하여 전반적으로 낮은 이상치 점수를 확인할 수 있다. 반면, 제안하는 모델은 MHSA 기반 Transformer 분류기를 통해 세그먼트 간의 시공간적 관계를 학습하여 짧은 노출에도 불구하고 비정상 특징을 더욱 잘 포착할 수 있음을 확인하였다.

세 번째 열은 가까운 거리에서 사고가 발생한 상황이다. 이 경우 차량의 움직임이 화면의 대부분을 차지하여 비디오의 모든 구간에서 특징 벡터의 크기가 크기 때문에 비디오의 앞뒤 맥락이 중요하다. 기존 모델의 경우 특징 벡터의 크기가 크기 때문에 평균적으로 높은 이상치 점수를 보여준다. 반면, 제안하는 모델은 Transformer 분류기를 통해 이웃한 세그먼트 간의 시공간적 맥락을 학습하였기 때문에 안정적으로 사고 감지가 이루어졌음을 확인할 수 있다. 이러한 결과를 바탕으로 제안된 모델은 정상과 비정상 데이터의 정확한 구분뿐만 아니라, 낮은 오탐률을 통해 실제 교통사고 감지 시스템에서 안정적이고 신뢰성 있는 운영이 가능함을 확인하였다.

Ⅴ. 결론 및 향후 과제

영상 기반 모니터링 시스템은 도로 교통 관리에서 중요한 역할을 하며, 교통사고를 신속하게 감지하여 인명 피해를 최소화하고 도로 안전을 강화할 수 있다. 기존의 CCTV 감시나 목격자 신고에 의존한 방식은 사고 감지의 지연 가능성이 있지만, 딥러닝 기반 이상치 탐지 기법은 비디오 클립에서 정상과 비정상 클립을 구분하여 이러한 문제를 해결할 수 있다. 일반적으로 약한 지도학습 방식과 재구성 및 예측 기반 방식이 사용되나, 각각 비정상 세그먼트의 잘못된 분류나 과도한 학습으로 인한 문제를 가질 수 있다.

따라서 본 연구에서는 I3D와 MHSA 기반 Transformer 기반 분류기를 결합한 새로운 지도학습 기반 이상치 탐지 모델을 제안하였으며, 이 모델은 입력 비디오를 세그먼트로 나눈 후 I3D를 통해 특징 벡터를 추출하고 MHSA를 통해 세그먼트 간의 관계를 학습하여 비정상 특징을 강화한 뒤, MLP로 정상 및 비정상 세그먼트를 정확하게 분류한다. 제안된 모델은 모든 임계값에서 우수한 성능을 보였으며, 특히 Youden's Index를 기반으로 한 기준 값에서도 높은 성능을 기록하여, 실제 운용 시 안정적이고 신뢰성 있는 교통사고 탐지 성능을 보장할 수 있음을 확인하였다.

향후 연구에서는 제안된 모델의 성능을 더욱 향상시키기 위해 복잡한 환경에서의 이상 탐지 성능을 강화하고, 실시간 처리 성능을 높이기 위한 최적화 방안에 대한 연구가 필요하다. 또한, 다양한 사고 유형에 대응할 수 있는 데이터셋 확장과 다중 카메라 기반 탐지 모델의 도입을 통해 실질적인 교통 관리 시스템에서의 적용성을 더욱 높일 수 있을 것이다.

References

[1] T. H. Kim and J. Y. Seo, "Traffic Accidents Risk Forecasting based on Deep Learning Models using Spatiotemporal Data Learning", The Journal of Korean Institute of Information Technology, Vol. 22, No. 5, pp. 1-12, May 2024. https://doi.org/10.14801/jkiit.2024.22.5.1.

[2] M. Tahir, Y. Qiao, N. Kanwal, B. Lee, and M. N. Asghar, "Real-Time Event-Driven Road Traffic Monitoring System Using CCTV Video Analytics", IEEE Access, Vol. 11, pp. 139097-139111, Dec. 2023. https://doi.org/10.1109/ACCESS.2023.3340144.

[3] Z. U. Arifeen, J.-E. Hong, B.-S. Seo, and J.-W. Suh, "Traffic Accident Detection and Classification in Videos based on Deep Network Features", 2023 Fourteenth International Conference on Ubiquitous and Future Networks (ICUFN), Paris, France, pp. 491-493, Jul. 2023. https://doi.org/10.1109/ICUFN57995.2023.10199977.

[4] J. Fang, J. Qiao, J. Xue, and Z. Li, "Vision-Based Traffic Accident Detection and Anticipation: A Survey", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 34, No. 4, pp. 1983-1999, Apr. 2024. https://doi.org/10.1109/TCSVT.2023.3307655.

[5] K. Okokpujie, Q. B. Sodipo, A. V. Akingunsoye, and M. E. Awomoyi, "Deep-Learning Algorithms for Video Anomaly Detection: A Mini-Review", 2023 2nd International Conference on Multidisciplinary Engineering and Applied Science (ICMEAS), Abuja, Nigeria, pp. 1-5, Nov. 2023. https://doi.org/10.1109/ICMEAS58693.2023.10429877.

[6] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset", 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 4724-4733, Jul. 2017. https://doi.org/10.1109/CVPR.2017.502.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need", Advances in Neural Information Processing Systems, Vol. 30, Jun. 2017.

[8] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos", 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 6479-6488, Jun. 2018. https://doi.org/10.1109/CVPR.2018.00678.

[9] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning", 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 4955-4966, Oct. 2021. https://doi.org/10.1109/ICCV48922.2021.00493.

[10] Y. Chen, Z. Liu, B. Zhang, W. Fok, X. Qi, and Y.-C. Wu, "MGFN: Magnitude-Contrastive Glance-and-Focus Network for Weakly-Supervised Video Anomaly Detection", AAAI, Vol. 37, No. 1, pp. 387-395, Jun. 2023. https://doi.org/10.1609/aaai.v37i1.25112.

[11] Y. Zhou, Y. Qu, X. Xu, F. Shen, J. Song, and H. Shen, "BatchNorm-based Weakly Supervised Video Anomaly Detection", arXiv:2311.15367, Nov. 2023. https://doi.org/10.48550/arXiv.2311.15367.

[12] Y. Li, Y. Cai, J. Liu, S. Lang, and X. Zhang, "Spatio-Temporal Unity Networking for Video Anomaly Detection", IEEE Access, Vol. 7, pp. 172425-172432, Nov. 2019. https://doi.org/10.1109/ACCESS.2019.2954540.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", arXiv:1505.04597, May 2015. https://doi.org/10.48550/arXiv.1505.04597.

[14] X. Shi, et al., "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting", NIPS'15: Proceedings of the 29th International Conference on Neural Information Processing Systems, Montreal Canada, Vol. 1, pp. 802-810, Dec. 2015.

[15] H. Yuan, Z. Cai, H. Zhou, Y. Wang, and X. Chen, "TransAnomaly: Video Anomaly Detection Using Video Vision Transformer", IEEE Access, Vol. 9, pp. 123977-123986, Aug. 2021. https://doi.org/10.1109/ACCESS.2021.3109102.

[16] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A Video Vision Transformer", 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 6816-6826, Oct. 2021. https://doi.org/10.1109/ICCV48922.2021.00676.

[17] B. Guo, M. Liu, Q. He, and M. Jiang, "Video Anomaly Detection with Video Vision Transformer", 2023 8th International Conference on Signal and Image Pro2023, pp. 131-135, Jul. 2023. https://doi.org/10.1109/ICSIP57908.2023.10270932.

[18] R. Fluss, D. Faraggi, and B. Reiser, "Estimation of the Youden Index and its Associated Cutoff

Point", Biometrical Journal, Vol. 47, No. 4, pp. 458-472, Aug. 2005. https://doi.org/10.1002/bimj.200410135.

저자소개

윤 성 안 (Sung-An Yoon)

2021년 2월 : 순천향대학교 전기공학과(공학사)
2023년 3월 : 순천향대학교 전기통신시스템공학과(공학석사)
2023년 3월 ~ 현재 : 순천향대학교 전기통신시스템공학과 박사과정
관심 분야 : 컴퓨터 비전, 영상처리, 딥 러닝

이 은 성 (Eunsung Lee)

2005년 2월 : 건국대학교 항공우주공학과(공학박사)
2005년 3월 ~ 2006년 2월 : 건국대학교 산업기술연구원 박사후연구원
2006년 2월 ~ 2007년 3월 : UCLA 기계항공공학과 박사후연구원
2007년 7월 ~ 현재 : 한국항공우주연구원 책임연구원
관심분야 : 위성항법, 보강항법, 항법응용, 고장검출

조 정 호 (Jeongho Cho)

2004년 12월 : Univ. of Florida 컴퓨터및전기공학과(공학박사)
2005년 1월 ~ 2006년 2월 : Univ. of Florida 의용공학과 박사후연구원
2006년 5월 ~ 2007년 12월 : 삼성전자 책임연구원
2007년 12월 ~ 2014년 3월 : 한국항공우주연구원 선임연구원
2017년 3월 ~ 현재 : 순천향대학교 전기공학과 부교수
관심분야 : 시스템 FDE, 적응신호처리, 기계학습