(12) BẢN MÔ TẢ SÁNG CHẾ THUỘC BẰNG ĐỘC QUYỀN SÁNG CHẾ

(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)**
CỤC SỞ HỮU TRÍ TUỆ

(11) **1-0053135**

(51)[2022.01] **G06N 3/04**

(13) **B**

(21) 1-2023-08939

(22) 14/12/2023

(45) 25/11/2025 452

(43) 26/02/2024 431

(73) TẬP ĐOÀN CÔNG NGHIỆP - VIỄN THÔNG QUÂN ĐỘI (VN)
Lô D26 Khu đô thị mới Cầu Giấy, phường Yên Hoà, quận Cầu Giấy, thành phố Hà Nội

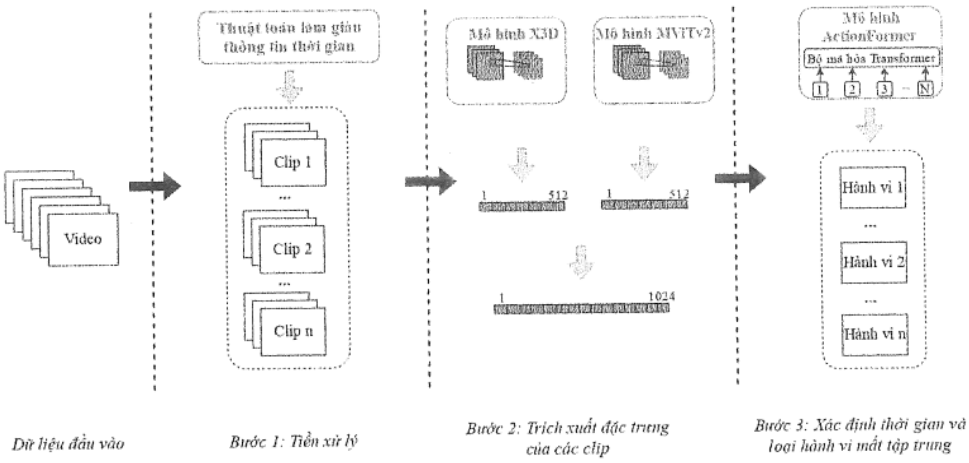
(72) LÊ HUY DƯƠNG (VN); TRẦN MẠNH TÙNG (VN); VŨ MINH QUÂN (VN).

(74) Công ty TNHH NACILAW (NACILAW)

(54) PHƯƠNG PHÁP PHÁT HIỆN CÁC HÀNH VI MẤT TẬP TRUNG CỦA NGƯỜI LÁI XE Ô TÔ TỪ VIDEO SỬ DỤNG CÔNG NGHỆ TRÍ TUỆ NHÂN TẠO

(21) 1-2023-08939

(57) Sáng chế này đề xuất phương pháp phát hiện các hành vi mất tập trung của người lái xe ô tô từ video sử dụng công nghệ trí tuệ nhân tạo. Phương pháp này gồm các bước: bước 1: tiền xử lý dữ liệu; bước 2: trích xuất đặc trưng của các clip ngắn sử dụng mô hình kết hợp của mạng nơ-ron tích chập ba chiều và mô hình học sâu kiến trúc Transformer; bước 3: xác định thời gian xảy ra hành vi và phân loại các hành vi mất tập trung của người lái xe ô tô trên video tổng thể.



Hình 1

Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến phương pháp phát hiện các hành vi mất tập trung của người lái xe ô tô từ video sử dụng công nghệ trí tuệ nhân tạo. Cụ thể, phương pháp đề cập trong sáng chế được ứng dụng trong việc theo dõi hành vi của người lái xe ô tô qua camera giám sát trên xe, hoặc trong hệ thống hỗ trợ lái xe nâng cao (Advanced driver assistance systems - ADAS) để đưa ra những cảnh báo trong tình huống nguy hiểm.

Tình trạng kỹ thuật của sáng chế

Các hành vi mất tập trung khi lái xe ô tô xảy ra khi người tài xế không chú ý và tập trung vào điều khiển phương tiện giao thông, thay vào đó thực hiện các hành vi khác như nhắn tin, nghe điện thoại, ăn uống, ngáp, ... Sự mất tập trung này có thể dẫn đến các va chạm, thậm chí là tai nạn giao thông, ảnh hưởng không chỉ đến bản thân người lái xe mà còn những người tham gia giao thông khác. Vì vậy việc phát hiện và cảnh báo các hành vi mất tập trung là cần thiết để giảm thiểu các tai nạn giao thông.

Hiện nay, để giảm thiểu các nguy cơ gây ra tai nạn giao thông do sự mất tập trung của tài xế, một số quy định về việc bắt buộc lắp đặt camera giám sát trên xe ô tô kinh doanh vận tải đã được ban hành. Với số lượng xe ngày càng lớn, đồng nghĩa với việc số lượng camera giám sát lớn, yêu cầu cần phải có công nghệ giám sát tự động và nhanh chóng. Ngoài ra, các hệ thống hỗ trợ lái xe nâng cao (ADAS) cũng mới được nghiên cứu gần đây nhằm hỗ trợ và nâng cao trải nhiệm lái xe cho các tài xế. Hệ thống ADAS thường được gắn nhiều camera trong xe, xung quanh xe cùng các hệ thống cảm biến khác để theo dõi sự di chuyển của xe và hành vi của tài xế. Do vậy, sáng chế này đề cập đến một phương pháp sử dụng công nghệ trí tuệ nhân tạo hỗ trợ việc chủ động phát hiện, sàng lọc các hành vi bất thường của người lái xe qua video để đưa ra cảnh báo tự động giúp giảm thiểu các rủi ro, tai nạn giao thông do các hành vi mất tập trung gây ra.

Bản chất kỹ thuật của sáng chế

Mục đích của sáng chế là đề xuất phương pháp phát hiện các hành vi mất tập trung của người lái xe ô tô từ video một cách chính xác và nhanh chóng, phục vụ việc giám sát

và cảnh báo các tình huống nguy hiểm khi lái xe.

Để đạt được mục đích trên, phương pháp phát hiện các hành vi mất tập trung của người lái xe ô tô từ video sử dụng công nghệ trí tuệ nhân tạo bao gồm các bước:

Bước 1: tiền xử lý dữ liệu; bước này được thực hiện với mục đích xử lý dữ liệu video truyền về từ các camera giám sát thành dữ liệu đầu vào phù hợp với mô hình học sâu (*deep learning*) ở bước 2, bao gồm việc làm giàu thông tin về chiều thời gian, cắt video thành các clip ngắn, giảm kích thước khung hình và chuẩn hóa.

Bước 2: trích xuất đặc trưng của các clip ngắn; bước này thực hiện dựa trên sự kết hợp của mạng nơ-ron tích chập ba chiều (*three-dimensional convolution neural network – 3DCNN*) và mô hình học sâu kiến trúc Transformer để trích chọn và học các đặc trưng không-thời gian (*spatial-temporal features*) của các clip đã qua tiền xử lý ở bước 1.

Bước 3: xác định thời gian xảy ra hành vi và phân loại các hành vi mất tập trung của người lái xe; bước này nhận đầu vào là đặc trưng của các clip ở bước 2, dựa trên một mô hình học sâu định vị thời gian của hành động để dự đoán thời gian bắt đầu và kết thúc hành vi, đồng thời phân loại hành vi mất tập trung của người lái xe ô tô.

Mô tả vắn tắt các hình vẽ

Hình 1: là hình ảnh mô tả các bước trong phương pháp phát hiện các hành vi mất tập trung của người lái xe ô tô từ video sử dụng công nghệ trí tuệ nhân tạo;

Hình 2: là hình ảnh mô tả thuật toán xếp chồng khung hình để làm giàu thông tin thời gian trong bước 1 tiền xử lý dữ liệu.

Mô tả chi tiết sáng chế

Sáng chế được mô tả chi tiết dưới đây có dựa vào các hình vẽ kèm theo, các hình vẽ này nhằm mục đích minh họa các phương án của sáng chế mà không làm giới hạn phạm vi bảo hộ sáng chế.

Tuy nhiên, cũng cần nói thêm rằng trong sáng chế này, một số thuật ngữ như: "Transformer", "Expand 3D", "X3D", "Improved Multiscale Vision Transformer", "MViTv2", "ActionFormer", "Adam", "Focal loss", "DIoU loss", "Cross-Entropy" là các danh từ riêng (tên của các thuật toán, mô hình, hàm số, v.v) không có từ tiếng Việt tương ứng và do đó trong sáng chế này sẽ sử dụng những từ chuyên biệt dạng chuẩn quốc tế.

Cụ thể, các bước của phương pháp đề cập trong sáng chế được thực hiện như sau (tham chiếu Hình 1):

Bước 1: tiền xử lý dữ liệu;

Bước này được thực hiện với mục đích tiền xử lý thông tin dạng video truyền liên tục về từ camera giám sát gắn trên xe ô tô (hay còn gọi là luồng Real Time Streaming Protocol (RTSP)) để đưa vào mô hình học sâu ở bước 2. Video được đưa về cùng một tốc độ phát là ba mươi khung hình trên giây (*frame per second - FPS*). Xuất phát từ thực tế, hầu hết các camera giám sát là loại đơn sắc, tức là thu được ảnh đen trắng (*gray-scale*), chỉ có một kênh màu. Các mô hình học sâu thường nhận đầu vào là ảnh với ba kênh màu: RGB (Red - Green – Blue (đỏ - xanh lá cây – xanh dương)) hoặc HSV (Hue - Saturation - Value (vùng màu – độ bão hòa – độ sáng)). Mặt khác, vấn đề đặt ra là nhận diện hành động, tức là các chuyển động thay đổi theo thời gian, do đó yếu tố thông tin về thời gian quan trọng hơn so với thông tin về màu sắc. Từ hai khía cạnh đề cập trên, thuật toán xếp chồng khung hình được đề xuất trong sáng chế này để tận dụng tính toán theo chiều kênh màu đồng thời làm giàu thêm thông tin về thời gian trong các khung hình (tham chiếu Hình 2). Cụ thể, tại mỗi khung hình đơn sắc, thuật toán này xếp chồng khung hình hiện tại và hai khung hình tiếp theo theo chiều kênh màu để được một khung hình có ba kênh, chứa thông tin về các khoảng thời gian liên tiếp.

Các mô hình học sâu trong vấn đề về xử lý video thường tiếp cận với đầu vào là các clip ngắn, do đó video đầu vào được cắt thành các clip có độ dài 1 giây với bước nhảy cửa sổ trượt (*sliding window stride*) của khung hình là sáu. Từ đó, tính được rằng mỗi clip có ba mươi khung hình, và các clip được lấy liền nhau liên tiếp. Để giảm chi phí và tăng tốc độ tính toán và giảm bộ nhớ lưu trữ, bước này thực hiện lấy mẫu (*sampling*): năm khung hình sẽ được lấy mẫu từ ba mươi khung hình theo phân phối đều, mỗi khung hình được lấy cách nhau sáu đơn vị. Sau đó, các khung hình sẽ được giảm kích thước về 384×384 điểm ảnh (*pixel*) và chuẩn hóa (*normalization*) về phân phối chuẩn N(0, 1) để trở thành đầu vào của mô hình học sâu ở bước 2.

Bước 2: trích xuất đặc trưng của các clip ngắn;

Bước này thực hiện dựa trên sự kết hợp của mô hình mạng nơ-ron tích chập ba chiều (3DCNN) và mô hình học sâu kiến trúc Transformer để trích chọn và học các đặc

trưng không-thời gian (*spatial-temporal features*) của các clip đã qua tiền xử lý ở bước 1.

Sáng chế này đề xuất sử dụng kết hợp mô hình mạng nơ-ron tích chập là Expand 3D (X3D) và mô hình học sâu kiến trúc Transformer là Improved Multiscale Vision Transformer (MViTv2). Thứ nhất, mô hình X3D được phát triển để giải quyết các bài toán nhận diện hành động. Với kiến trúc mạng nơ-ron tích chập ba chiều, mô hình X3D có thể học các đặc trưng không-thời gian hiệu quả. Sử dụng chiến lược xây dựng kiến trúc tối ưu, mô hình này cho thấy kết quả tốt trong khi có độ phức tạp tính toán thấp. Thứ hai, mô hình MViTv2 với thế mạnh dựa trên kiến trúc Transformer, chứng tỏ được khả năng vượt trội trong các bài toán xử lý video. Mô hình sử dụng chiến lược phân cấp với các kích cỡ đa dạng để học các đặc trưng từ tổng quan đến chi tiết. Việc kết hợp hai mô hình X3D và MViTv2 với nhằm phục vụ cho việc trích xuất đặc trưng từ các clip có sự đa dạng về các đặc trưng cục bộ (*local features*) và toàn cục (*global features*). Tại bước này, hai mô hình X3D và MViTv2 trả về hai véc-tơ đặc trưng cho từng clip, mỗi véc-tơ có 512 chiều; hai véc-tơ sau đó được nối thành một véc-tơ đặc trưng 1024 chiều duy nhất.

Dựa trên nghiên cứu và thí nghiệm thực tế, hai mô hình hoạt động tốt khi được huấn luyện với bộ siêu tham số (*hyper-parameters*) như sau: số lượng vòng lặp huấn luyện (epoch) là 100, kích thước lô huấn luyện (*batch size*) là 32, hàm mất mát (loss function) là Cross-Entropy được tối ưu hóa bởi thuật toán tối ưu Adam với các tham số (*parameters*): tốc độ học khởi tạo (*learning rate*) là $5x10^{-4}$, các tốc độ phân rã *(exponential decay)* lần lượt là $\beta_1 = 0{,}9, \beta_2 = 0{,}999$. Tốc độ học sẽ giảm dần trong quá trình huấn luyện theo hàm cosin.

Bước 3: xác định thời gian xảy ra hành vi và phân loại các hành vi mất tập trung của người lái xe ô tô;

Bước này nhận đầu vào là đặc trưng của các clip ở bước 2, dựa trên một mô hình định vị thời gian của hành động ActionFormer để dự đoán thời gian bắt đầu và kết thúc hành vi, đồng thời phân loại hành vi mất tập trung của người lái xe ô tô. Mô hình ActionFormer được xây dựng dựa trên kiến trúc Transformer, nhận đầu vào các đặc trưng của clip như là các mã thông báo (*token*), để dự đoán thời gian bắt đầu và kết thúc hành động cũng như phân loại hành động. Các nhãn hành vi mất tập trung của người lái xe ô

tô bao gồm: "*Ăn*", "*Uống nước*", "*Ngáp*", "*Gọi điện thoại*", "*Nhắn tin*", "*Quay ra phía sau*", "*Cúi xuống nhặt đồ*", "*Nói chuyện với hành khách*", "*Đặt tay trên đầu*", "*Hát hoặc nhảy múa với âm nhạc*", và "*Lái xe bình thường*".

Dựa trên nghiên cứu và thí nghiệm thực tế, mô hình hoạt động tốt khi được huấn luyện với bộ siêu tham số (*hyper-parameters*) như sau: số lượng vòng lặp huấn luyện (*epoch*) là 50, kích thước lô huấn luyện (*batch size*) là 32, hàm mất mát (*loss function*) là Focal loss cho phân loại và DIoU loss cho xác định thời gian xảy ra hành vi, được tối ưu hóa bởi thuật toán tối ưu Adam với các tham số (*parameters*): tốc độ học khởi tạo (*learning rate*) là $1x10^{-4}$, các tốc độ phân rã *(exponential decay)* lần lượt là $\beta_1 = 0,9, \beta_2 = 0,999$. Tốc độ học sẽ giảm dần trong quá trình huấn luyện theo hàm cosin.

Cuối cùng, sau khi đi qua mô hình định vị thời gian ActionFormer, thu được các hành vi mất tập trung có các thông số: thời gian bắt đầu, thời gian kết thúc và loại hành vi cùng với xác suất (%) của hành vi đó. Nếu xác suất xảy ra hành vi mất tập trung của các dự đoán lớn hơn một ngưỡng nhất định thì dự đoán sẽ được tính là một hành vi mất tập trung. Trong sáng chế này, dựa trên nghiên cứu và thực nghiệm thực tế, ngưỡng tối ưu được lựa chọn là 25%.

Mặc dù các mô tả nêu trên chứa nhiều chi tiết cụ thể, chúng không được coi là giới hạn về phương án thực thi sáng chế, mà chỉ nhằm mục đích minh họa một số phương án thực thi được ưu tiên.

Yêu cầu bảo hộ

1. Phương pháp phát hiện các hành vi mất tập trung của người lái xe ô tô từ video sử dụng công nghệ trí tuệ nhân tạo bao gồm các bước:

bước 1: tiền xử lý dữ liệu;

tại bước này, dữ liệu video truyền về từ các camera giám sát trên xe ô tô được làm giàu thông tin về chiều thời gian, cắt video thành các clip ngắn, giảm kích thước khung hình và chuẩn hóa.

bước 2: trích xuất đặc trưng của các clip ngắn;

tại bước này, các clip đã qua tiền xử lý ở bước 1 được đưa qua mô hình kết hợp của mạng nơ-ron tích chập ba chiều (*three-dimensional convolution neural network – 3DCNN*) và mô hình học sâu kiến trúc Transformer để trích chọn và học các đặc trưng không-thời gian (*spatial-temporal features*).

bước 3: xác định thời gian xảy ra hành vi và phân loại các hành vi mất tập trung của người lái xe ô tô;

tại bước này, đặc trưng đã trích chọn được từ bước 2 đi qua một mô hình học sâu định vị thời gian của hành động để dự đoán thời gian bắt đầu và kết thúc hành vi, đồng thời phân loại hành vi mất tập trung của người lái xe ô tô.

2. Phương pháp phát hiện các hành vi mất tập trung của người lái xe ô tô từ video sử dụng công nghệ trí tuệ nhân tạo theo điểm 1, trong đó:

ở bước 1, luồng video truyền liên tục về từ camera giám sát gắn trên xe ô tô được tiền xử lý trước khi đưa vào mô hình học sâu ở bước 2; video được đưa về cùng một tốc độ phát là ba mươi khung hình trên giây (*frame per second - FPS*); thuật toán xếp chồng khung hình được đề xuất trong sáng chế này để tận dụng tính toán theo chiều kênh màu đồng thời làm giàu thêm thông tin về thời gian trong các khung hình; cụ thể, tại mỗi khung hình đơn sắc, thuật toán này xếp chồng khung hình hiện tại và hai khung hình tiếp theo theo chiều kênh màu để được một khung hình có ba kênh, chứa thông tin về các khoảng thời gian liên tiếp; để giảm chi phí và tăng tốc độ tính toán và giảm bộ nhớ lưu trữ, bước này thực hiện lấy mẫu (*sampling*): năm khung hình sẽ được lấy mẫu từ ba mươi khung hình theo phân phối đều, mỗi khung hình được lấy cách nhau sáu đơn vị; sau đó, các khung hình sẽ được giảm kích thước về 384×384 điểm ảnh (*pixel*) và chuẩn hóa

(*normalization*) về phân phối chuẩn N(0, 1) để trở thành đầu vào của mô hình học sâu ở bước 2.

3. Phương pháp phát hiện các hành vi mất tập trung của người lái xe ô tô từ video sử dụng công nghệ trí tuệ nhân tạo theo điểm 1, trong đó:

tại bước 2, nhằm trích xuất đặc trưng từ các clip có độ đa dạng về các đặc trưng cục bộ và toàn cục, sáng chế này đề xuất việc trích xuất đặc trưng từ sự kết hợp của hai mô hình expand 3D (X3D) và improved multiscale vision transformer (MViTv2); thứ nhất, mô hình X3D được phát triển để giải quyết các bài toán nhận diện hành động, với kiến trúc mạng nơ-ron tích chập ba chiều, mô hình X3D có thể học các đặc trưng không-thời gian hiệu quả; sử dụng chiến lược xây dựng kiến trúc tối ưu, mô hình này cho thấy kết quả tốt trong khi có độ phức tạp tính toán thấp; thứ hai, mô hình MViTv2 với thế mạnh dựa trên kiến trúc transformer, chứng tỏ được khả năng vượt trội trong các bài toán xử lý video; mô hình sử dụng chiến lược phân cấp với các kích cỡ đa dạng để học các feature từ tổng quan đến chi tiết; sáng chế này đề xuất việc kết hợp hai mô hình X3D và MViTv2 với nhau phục vụ cho việc trích xuất đặc trưng từ các clip có độ đa dạng về các đặc trưng cục bộ (*local features*) và toàn cục (*global features*); kết quả, hai mô hình X3Dvà MViTv2 trả về hai véc-tơ đặc trưng cho từng clip, mỗi véc-tơ có 512 chiều; hai véc-tơ sau đó được nối thành một véc-tơ đặc trưng 1024 chiều duy nhất.

4. Phương pháp phát hiện các hành vi mất tập trung của người lái xe ô tô từ video sử dụng công nghệ trí tuệ nhân tạo theo điểm 1, trong đó:

tại bước 2, dựa trên nghiên cứu và thí nghiệm thực tế, hai mô hình hai mô hình X3Dvà MViTv2 hoạt động tốt khi được huấn luyện với bộ siêu tham số (*hyper-parameters*) như sau: số lượng vòng lặp huấn luyện (epoch) là 100, kích thước lô huấn luyện (*batch size*) là 32, hàm mất mát (loss function) là cross-entropy được tối ưu hóa bởi thuật toán tối ưu adam với các tham số (*parameters*): tốc độ học khởi tạo (*learning rate*) là $5x10^{-4}$, các tốc độ phân rã *(exponential decay)* lần lượt là $\beta_1 = 0.9, \beta_2 = 0.999$; tốc độ học sẽ giảm dần trong quá trình huấn luyện theo hàm cosin.

5. Phương pháp phát hiện các hành vi mất tập trung của người lái xe ô tô từ video sử dụng công nghệ trí tuệ nhân tạo theo điểm 1, trong đó:

tại bước 3, nhận đầu vào là đặc trưng của các clip ở bước 2, dựa trên một mô hình

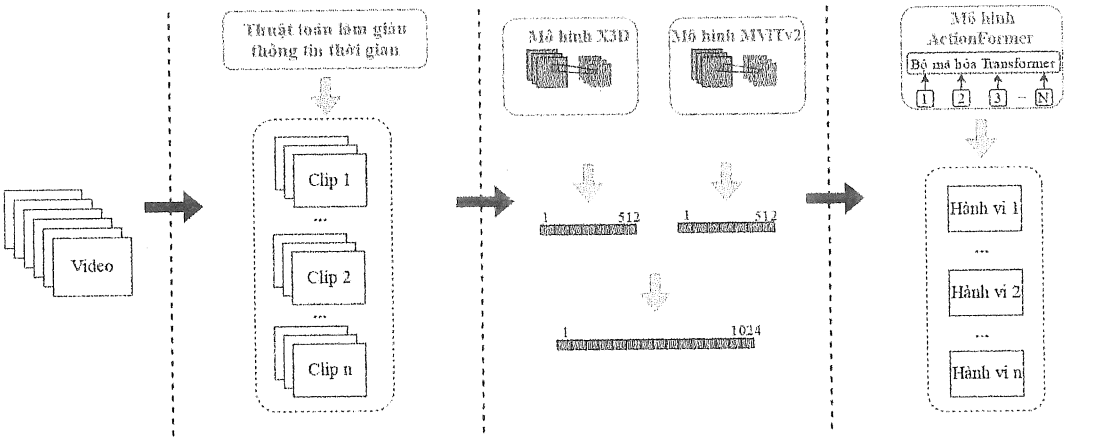
định vị thời gian của hành động actionformer để dự đoán thời gian bắt đầu và kết thúc hành vi, đồng thời phân loại hành vi mất tập trung của người lái xe ô tô; mô hình actionformer được xây dựng dựa trên kiến trúc transformer, nhận đầu vào các đặc trưng của clip như là các mã thông báo (*token*), để dự đoán thời gian bắt đầu và kết thúc hành động cũng như phân loại hành động; các loại hành vi mất tập trung của người lái xe ô tô bao gồm: "*ăn*", "*uống nước*", "*ngáp*", "*gọi điện thoại*", "*nhắn tin*", "*quay ra phía sau*", "*cúi xuống nhặt đồ*", "*nói chuyện với hành khách*", "*đặt tay trên đầu*", "*hát hoặc nhảy múa với âm nhạc*", và "*lái xe bình thường*".

6. Phương pháp phát hiện các hành vi mất tập trung của người lái xe ô tô từ video sử dụng công nghệ trí tuệ nhân tạo theo điểm 1, trong đó:

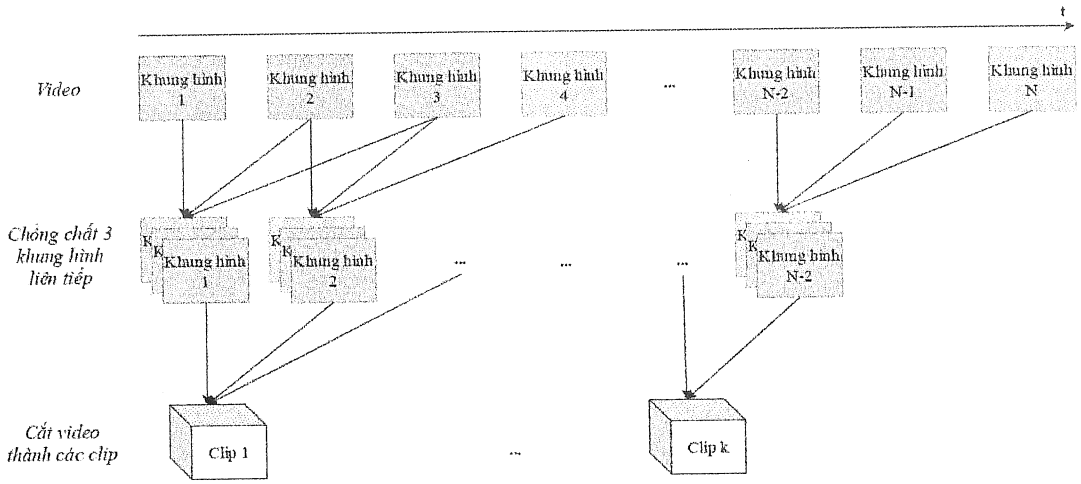
tại bước 3, dựa trên nghiên cứu và thí nghiệm thực tế, mô hình actionformer hoạt động tốt khi được huấn luyện với bộ siêu tham số (*hyper-parameters*) như sau: số lượng vòng lặp huấn luyện (*epoch*) là 50, kích thước lô huấn luyện (*batch size*) là 32, hàm mất mát (*loss function*) là focal loss cho phân loại hành vi và DIoU loss cho xác định thời gian xảy ra hành vi, được tối ưu hóa bởi thuật toán tối ưu adam với các tham số (*parameters*): tốc độ học khởi tạo (*learning rate*) là $1x10^{-4}$, các tốc độ phân rã *(exponential decay)* lần lượt là $\beta_1 = 0{,}9, \beta_2 = 0{,}999$; tốc độ học sẽ giảm dần trong quá trình huấn luyện theo hàm cosin.

7. Phương pháp phát hiện các hành vi mất tập trung của người lái xe ô tô từ video sử dụng công nghệ trí tuệ nhân tạo theo điểm 1, trong đó:

tại bước 3, sau khi đi qua mô hình định vị thời gian actionformer, thu được các hành vi mất tập trung có các thông số: thời gian bắt đầu, thời gian kết thúc và loại hành vi cùng với xác suất (%) của hành vi đó; dựa trên nghiên cứu và thực nghiệm thực tế, ngưỡng tối ưu được lựa chọn là 25%.


Video
Thuật toán làm giàu thông tin thời gian
Clip 1
Clip 2
Clip n
Mô hình X3D
Mô hình MVITv2
1
512
1024
Mô hình ActionFormer
Bộ mã hóa Transformer
Hành vi 1
Hành vi 2
Hành vi n
Dữ liệu đầu vào
Bước 1: Tiền xử lý
Bước 2: Trích xuất đặc trưng của các clip
Bước 3: Xác định thời gian và loại hành vi mất tập trung

Hình 1


t
Video
Khung hình 1
Khung hình 2
Khung hình 3
Khung hình 4
Khung hình N-2
Khung hình N-1
Khung hình N
Chồng chất 3 khung hình liên tiếp
K
Cắt video thành các clip
Clip 1
Clip k

Hình 2