MÔ HÌNH PHÁT HIỆN HÀNH VI BẠO LỰC ĐA TẦNG SỬ DỤNG MẠNG NƠRON TÍCH CHẬP VÀ MẠNG BỘ NHỚ DÀI-NGẮN HẠN

Nguyễn Mạnh Dũng, Vũ Hoài Nam, Phạm Đức Cường, Nguyễn Việt Hưng
Học Viện Công Nghệ Bưu Chính Viễn Thông

Tóm Tắt: Nhận diện hành vi là chủ đề nghiên cứu đầy thách thức của lĩnh vực thị giác máy tính với rất nhiều các ứng dụng hữu ích trong thực tế trong đó bao gồm phát hiện hành vi bạo lực. Phát hiện sớm hành vi bạo lực giúp chúng ta kịp thời có những hành động, có thể ngăn chặn hay chí ít cũng có thể giảm thiểu thiệt hại do tình trạng bạo lực gây ra. Trong bài báo này, chúng tôi trình bày phương pháp phát hiện hành vi bạo lực đa tầng, ở giai đoạn đầu, những nhóm đối tượng có nguy cơ xảy ra bạo lực cao được phát hiện bằng phương pháp sử dụng YOLO và thuật toán theo dõi đối tượng Deep SORT. Tiếp theo các đặc trưng 2D của nhóm đối tượng được trích xuất bằng mạng nơ-ron tích chập (CNN) ở giai đoạn thứ hai. Cuối cùng những đặc trưng này được sử dụng làm đầu vào cho mạng bộ nhớ dài ngắn (LSTM) ở giai đoạn cuối để xác định xem nhóm đối tượng có hành vi bạo lực hay chỉ là nhóm đối tượng có hành vi bình thường. Các kết quả thực nghiệm cho thấy, so với các nghiên cứu trước đó, phương pháp được đề xuất không chỉ hiệu quả để phát hiện hành vi bạo lực mà còn giảm số lượng phát hiện sai, hiệu suất tốt phù hợp để ứng dụng trong thực tế.

Từ khóa: Mạng nơ ron tích chập, phát hiện hành vi bạo lực, YOLO, mạng bộ nhớ dài-ngắn hạn.

I. MỞ ĐẦU

Bạo lực luôn là một vấn đề được quan tâm hàng đầu trong xã hội nào không chỉ riêng Việt Nam mà trên toàn thế giới. Hành vi bạo lực gây ra rất nhiều hệ lụy cho xã hội, gây tổn hại cả về sức khỏe, tinh thần, tài sản và đôi khi là cả tính mạng của con người. Đã có nhiều biện pháp được đưa ra nhưng tình trạng bạo lực vẫn thường xuyên xảy ra, không có dấu hiệu thuyên giảm.

Nếu có một cơ chế giám sát hiệu quả, kịp thời phát hiện và cảnh báo hành vi bạo lực, chúng ta hoàn toàn có thể ngăn chặn hay chí ít là giảm thiểu tối đa những thiệt hại do bạo lực gây ra.

Trong những năm gần đây hệ thống camera giám sát theo dõi CCTV phát triển rất nhanh và được cài đặt ở khắp mọi nơi, từ bệnh viện, trường học, đường phố cho đến những nơi công cộng khác. Một trong những chức năng quan trọng, đang thu hút được sự quan tâm của giới nghiên cứu là khả năng tăng cường giám sát, theo dõi hành vi bạo lực. Các hệ thống hiện tại đa phần chỉ có chức năng thu nhận và lưu trữ hình ảnh còn công việc giám sát chủ yếu vẫn dựa vào sức người, do đó hiệu quả giám sát vẫn còn rất thấp với chi phí vận hành cao.

Tích hợp trí tuệ nhân tạo AI vào hệ thống camera, giúp nâng cao hiệu quả điều hành giám sát là một trong những xu hướng nghiên cứu mới nhất hiện nay.

Phương pháp học máy (Machine Learning) là các phương pháp tự động xây dựng một mô hình toán học bằng cách sử dụng dữ liệu mẫu, còn được gọi là dữ liệu đào tạo, có khả năng tự học hỏi dựa trên dữ liệu đưa vào mà không cần phải được lập trình cụ thể. Học máy đã được phát triển từ những năm 1940, tuy nhiên kết quả thực sự không được ấn tượng. Nguyên nhân chủ yếu là khó khăn trong việc thu thập dữ liệu và hạn chế của tài nguyên máy tính. Cho đến cuối thập niên, khi mà phần cứng máy tính đã trở nên mạnh hơn cùng với sự phát triển của internet đã giúp cho quá trình thu thập dữ liệu trở nên dễ dàng và thúc đẩy sự phát triển của học máy.

Gần đây, một nhánh của học máy là học sâu (Deep Learning) đã nổi lên thành một phương pháp học máy tốt nhất. Học sâu bao gồm tập các kỹ thuật học máy mạnh sử dụng mạng nơ-ron nhiều lớp, nhờ đó đã đạt được nhiều kết quả tốt trong những bài toán thị giác máy tính. Thay đổi quan trọng này cũng giúp giải quyết những bài toán khó còn tồn tại như phát hiện hành vi bạo lực.

Qua bài báo này, chúng tôi đề xuất một phương pháp ứng dụng học sâu phát hiện hành vi bạo lực đa tầng sử dụng YOLO và CNN-LSTM. Phần còn lại của bài báo được bố cục như sau. Phần II giới thiệu các nghiên cứu liên quan. Phần III mô tả phương pháp được đề xuất. Phần IV báo cáo kết quả thực nghiệm. Hướng nghiên cứu trong tương lai và thảo luận được trình bày trong Phần V.

II. NGHIÊN CỨU LIÊN QUAN

Trong tài liệu, nhiều phương pháp đã được đề xuất cho bài toán phát hiện hành vi bạo lực [1, 2, 3]. Các phương pháp này tập trung vào sử dụng học máy và phân tích hình ảnh.

Phương pháp sử dụng học máy đạt kết quả tốt có thể kể đến Fast fight detection [1]. Phương pháp cho rằng, trong một đoạn video bạo lực, vùng điểm ảnh chuyển động có hình dạng và vị trí đặc biệt. Đầu tiên, sự khác biệt giữa những khung hình liên tiếp được tính toán và lấy giá trị tuyệt đối. Tiếp theo, ảnh kết quả được nhị phân hoá, tạo ra những vùng chuyển động. K vùng chuyển động lớn nhất được lựa chọn để xử lý tiếp. Cuối cùng, để phân loại K vùng chuyển động này, các thông số: tâm, chu vi, diện tích

Tác giả liên hệ: Nguyễn Mạnh Dũng,
Email: dungnm@ptit.edu.vn
Đến tòa soạn: 10/2021, chỉnh sửa: 11/2021, chấp nhận đăng: 12/2021.

và khoảng cách giữa chúng được tính toán. Kết quả thử nghiệm trên các tập dữ liệu Movie, Hockey Fight và UCF-101 cho thấy phương pháp có hiệu quả tốt và có thể ứng dụng trong thế giới thực. Tuy nhiên, phương pháp này không hiệu quả khi phân loại những video có chuyển động liên tục.

Học sâu là một tập hợp con của học máy, tập trung chủ yếu vào sử dụng mạng nơ-ron nhiều lớp. Học sâu đã đạt được độ chính xác và hiệu quả cao trong rất nhiều bài toán thị giác máy tính so với học máy truyền thống, trong đó phải kể đến phân loại hình ảnh.

Mạng nơ-ron tích chập CNN (Convolutional Neural Network) [4] là một trong những kiến trúc được sử dụng rộng rãi nhất được dùng để phân loại ảnh. CNN ra đời dựa vào việc mô phỏng một phần cách thức hoạt động của não bộ con người - sử dụng những đặc trưng từ không gian để phân loại một bức ảnh. CNN sử dụng rất nhiều bộ lọc có khả năng học hỏi để tự động trích xuất đặc trưng từ hình ảnh, vì vậy CNN có thể "nhìn được" những đặc trưng quan trọng mà trích xuất đặc trưng thủ công khó có thể phát hiện. Trong quá trình huấn luyện những đặc trưng quan trọng sẽ được giữ lại trong khi những đặc trưng không tốt sẽ được loại bỏ khỏi hệ thống.

HINH 1. HÌNH ẢNH TỪ TẬP DỮ LIỆU HOCKEY

Mặc dù đã gặt hái được những thành công nhất định trong nhiệm vụ phân loại ảnh, tuy nhiên CNN lại không hiệu quả với bài toán phân loại hành vi. Nguyên nhân chính do hành động là một chuỗi hình ảnh liên tiếp, nên nếu chỉ sử dụng một hình ảnh đơn lẻ thì khó có thể đưa ra được dự đoán chính xác. Ví dụ với hình ảnh từ tập dữ liệu Hockey ở Hình 1, chúng ta không thể phân biệt được đây là hành vi bạo lực hay chỉ là một hoạt động thể thao thông thường.

Khác với CNN, mạng bộ nhớ dài ngắn LSTM (Long short term memory network) [5] được tạo ra với ý tưởng bắt chước suy nghĩ của con người – một nhược điểm mà mạng nơ-ron truyền thống chưa thể làm được. Con người không bắt đầu suy nghĩ của họ từ đầu tại tất cả các thời điểm, ví dụ như để phân loại một hình ảnh trong bộ phim, con người sẽ sử dụng đến những hình ảnh trước nữa chứ không sử dụng duy nhất hình ảnh hiện tại như CNN. LSTM có dạng một chuỗi các mô-đun lặp đi lặp lại của mạng nơ-ron để mô phỏng lại cách suy nghĩ của não người. Ở đó mỗi mô-đun bao gồm 4 tầng mạng nơ-ron khác nhau và tương tác với nhau một cách đặc biệt. Nhờ vậy, LSTM có thể ghi nhớ thông tin trong thời gian dài, phù hợp để học tập với chuỗi hình ảnh trong bài toán phân loại hành động.

HÌNH 2. HÌNH ẢNH TỪ TẬP DỮ LIỆU PTIT

Nhiều nghiên cứu cho thấy, học sâu có thể áp dụng khá tốt cho bài toán phát hiện hành vi bạo lực [6, 7, 8, 9]. Trong đó, phương pháp đem lại chính xác và hiệu quả nhất là phương pháp sử dụng kết hợp CNN và LSTM [9]. Đầu tiên, các khung hình liên tiếp được đưa vào một CNN để trích xuất đặc trưng. Sau đó những đặc trưng này được đưa vào Bidirectional LSTM [10] để phân loại những khung hình liên tiếp đó là bạo lực hay không. Phương pháp được thử nghiệm trên tập dữ liệu Hockey, Peliculas và Collected Surveillance Camera (bộ dữ liệu được nhóm tác giả thu thập) cho kết quả rất tốt và có thể sử dụng để phân loại những video có chuyển động liên tục. Dù vậy, phương pháp đạt độ chính xác không cao khi phân loại những video có cảnh bạo lực chỉ chiếm phần nhỏ so với khung hình. Ví dụ ở hình ảnh số 2 từ tập dữ liệu PTIT, cảnh bạo lực trong video chỉ chiếm khoảng một phần diện tích nhỏ trong khung hình, khi đó thuật toán sẽ hoạt động không tốt.

Nguyên nhân là vì khi sử dụng toàn ảnh để trích chọn đặc trưng thì phần đặc trưng mô tả hành vi bạo lực không thực sự nổi bật so với các đối tượng khác. Để khắc phục nhược điểm này, chúng tôi đề xuất phương pháp nhận diện hành vi bạo lực đa tầng sử dụng CNN-LSTM. Mô hình này cho phép chúng tôi tập trung hơn vào vị trí xảy ra hành vi bạo lực, từ đó cho kết quả có độ chính xác cao hơn.

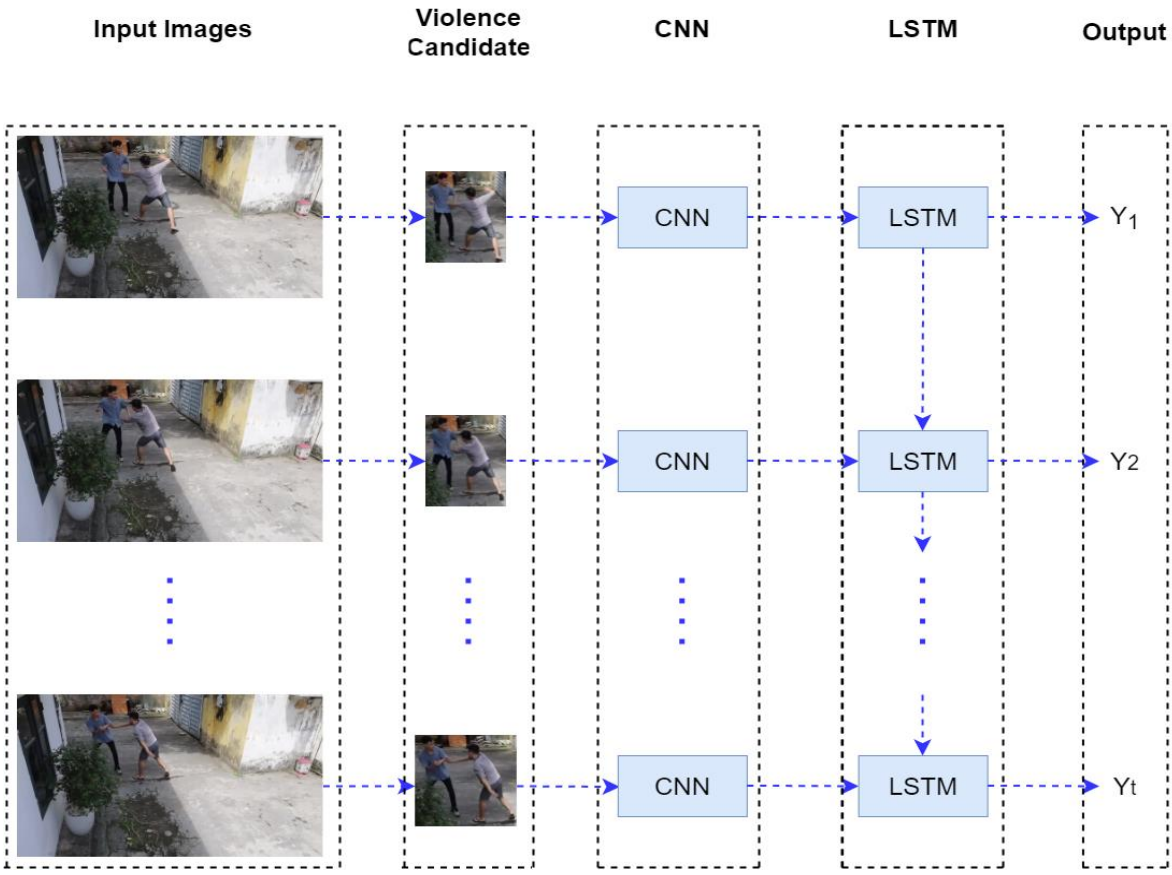
III. PHƯƠNG PHÁP ĐỀ XUẤT

Phương pháp phát hiện hành vi bạo lực chúng tôi đề xuất được minh hoạ trong Hình 3. Cách tiếp cận này được chia thành ba giai đoạn chính. Ở giai đoạn đầu, những nhóm người có khả năng bạo lực cao sẽ được phát hiện và khoanh vùng bằng YOLO [11], kết hợp thuật toán theo dõi đối tượng Deep SORT [12]. Trong giai đoạn tiếp theo, hình ảnh nhóm người có nguy cơ bạo lực cao từ các khung hình sẽ được đưa vào CNN để trích xuất đặc trưng. Và giai đoạn cuối cùng, những đặc trưng này sẽ được đưa vào LSTM để phân loại và quyết định xem nhóm người có hành vi bạo lực thực sự hay chỉ là hành động bình thường.

A. PHÁT HIỆN NHÓM NGƯỜI NGUY CƠ BẠO LỰC CAO

Bước đầu tiên, chúng tôi khoanh vùng những nhóm đối tượng có nguy cơ bạo lực cao. Chúng tôi đưa ra một luật đơn giản để chọn những nhóm này đó là nhóm những người đứng cạnh nhau.

Chúng tôi sử dụng YOLOv4 [13] – là một trong những phương pháp phát hiện đối tượng được sử dụng rộng rãi và tốt nhất hiện nay đã được kiểm chứng trên tập dữ liệu MS COCO.



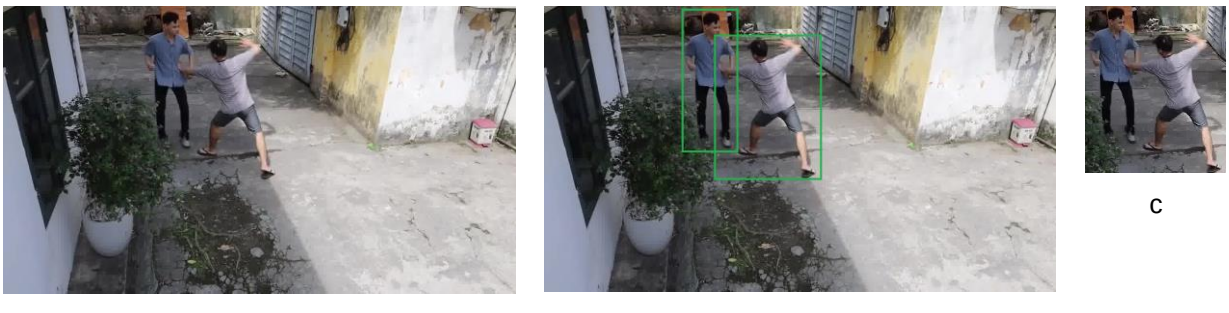
HÌNH 3. KIẾN TRÚC TỔNG QUAN MÔ HÌNH PHÁT HIỆN BẠO LỰC ĐA TẦNG KẾT HỢP CNN-LSTM

Các đối tượng sau khi được phát hiện bởi thuật toán YOLO sẽ được theo dõi và liên kết qua một chuỗi các khung hình bởi thuật toán Deep SORT [12], từ đó có thể tìm ra được nhóm các đối tượng gần nhau. Đây chính là nhóm đối tượng có nguy cơ xảy ra hành vi bạo lực. Kết quả quá trình được minh hoạ ở Hình 4.

B. CNN-LSTM

Không phải tất cả các nhóm đối tượng đứng cạnh nhau đều có thể xảy ra hành vi bạo lực, để đi đến quyết định cuối cùng chúng ta cần tiến hành thêm một bước xử lý nhằm phân biệt đâu là hành vi bạo lực thật đâu là những nhóm không có hành vi bạo lực.

Hành vi bạo lực là một chuỗi hành động, vì vậy chúng ta cần quan sát chuỗi các khung hình liên tiếp để có thể đưa ra được dự đoán cuối cùng.



HÌNH 4. VÍ DỤ PHÁT HIỆN NHÓM ĐỐI TƯỢNG NGUY CƠ BẠO LỰC CAO

Như đã đề cập trước đó, mô hình thích hợp nhất để phân loại dữ liệu dạng chuỗi các khung hình liên tiếp đó là kiến trúc kết hợp CNN-LSTM.

Kiến trúc CNN-LSTM sử dụng CNN để trích xuất đặc trưng 2D của ảnh đầu vào, sau đó kết hợp cùng LSTM để phân tích liên kết về mặt thời gian của dữ liệu trước khi đưa ra dự đoán cuối cùng.

Để lựa chọn mô hình CNN, chúng tôi tiến hành thử nghiệm một số mô hình đang được ứng dụng nhiều hiện nay trên tập dữ liệu ImageNet. Kết quả ở Bảng 1 cho thấy Resnet18 [14], đem lại sự cân bằng tốt nhất giữa độ chính xác và độ phức tạp của mô hình, vì vậy Resnet18 đã được lựa chọn sử dụng cho việc trích xuất đặc trưng 2D của ảnh. Chúng tôi thay đổi layer cuối cùng để thu được vector đặc trưng 256 chiều thay vì 1000 như ở phiên bản gốc.

Vector đặc trưng 256 chiều này sẽ được sử dụng làm dữ liệu đầu vào của mạng LSTM phân loại hành vi bạo lực.

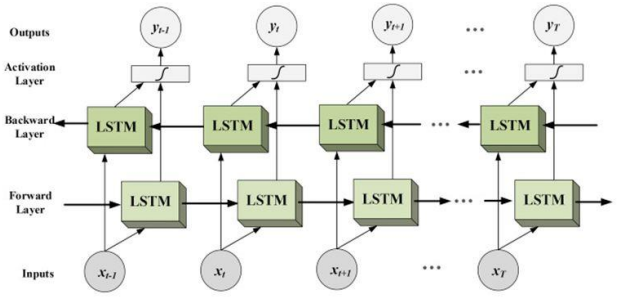
Các đặc trưng của những vùng nguy bạo lực cao từ các khung hình liên tiếp được đưa vào LSTM để trích xuất những đặc trưng về không gian và thời gian trước khi đưa vào bộ phân loại Softmax đưa ra quyết định cuối cùng. Chúng tôi cũng sử dụng Bidirectional LSTM thay cho LSTM thông thường để tăng tính liên kết giữa các đầu vào. Kiến trúc của một Bi-LSTM được minh hoạ ở Hình 5. Bi-LSTM không chỉ lưu trữ thông tin từ quá khứ mà còn lưu trữ cả thông tin đến từ tương lai,

BẢNG 1. KẾT QUẢ THỬ NGHIỆM MỘT SỐ MÔ HÌNH TRÊN TẬP DỮ LIỆU IMAGENET

Model	Parameters	Accuracy(%)
MobileNet	4.2 M	70.6
Resnet18	11.4 M	80.7
Resnet34	21.5 M	82.4
Resnet50	23.9 M	85.8
InceptionNet	23.2 M	83.2
VGG16	138.4 M	80.5
VGG19	143.7 M	84.2

kiến trúc như vậy giúp mô hình dễ đưa ra dự đoán hơn khi mà chuỗi hành vi bạo lực được Bi-LSTM tiếp nhận thông tin từ cả hai chiều thời gian.

Chúng tôi cũng chọn kiến trúc LSTM hai tầng bởi vì qua thực nghiệm, so với một tầng LSTM thì kiến trúc LSTM hai tầng cho kết quả tốt hơn, trong khi nếu sử dụng



HÌNH 5. KIẾN TRÚC CỦA MỘT BI-LSTM

nhiều hơn hai tầng LSTM thì độ chính xác tăng không đáng kể nhưng thời gian xử lý lại tăng lên nhiều.

IV. KẾT QUẢ THỰC NGHIỆM

A. TẬP DỮ LIỆU

Để đánh giá độ chính xác cũng như hiệu quả hoạt động của thuật toán, chúng tôi tiến hành các thực nghiệm trên ba tập dữ liệu Hockey Fight, Peliculas và PTIT được thống kê trên Bảng 2.

BẢNG 2. THỐNG KÊ CÁC TẬP DỮ LIỆU

Tập dữ liệu	# violence	# non-violence
Hockey Fight	500	500
Peliculas	100	100
PTIT	120	90

1) Tập dữ liệu Hockey Fight : Tập dữ liệu chứa cảnh bạo lực và không từ trò chơi khúc côn cầu trên băng. Có tổng cộng 1000 video, trong đó 500 mẫu là bạo lực và 500 mẫu là không bạo lực. Tất cả video có độ dài 2 giây, kích thước khung hình giữa các video là giống nhau và cảnh bạo lực chiếm phần lớn khung hình. Các video có chung nền và có chuyển động nền.

HÌNH 6. MỘT SỐ VÍ DỤ MINH HỌA TRONG TẬP DỮ LIỆU ĐÁNH GIÁ

2) Tập dữ liệu Peliculas : Tập dữ liệu bao gồm các phân cảnh bạo lực và không từ những bộ phim Hollywood, trò chơi

bóng đá và các sự kiện khác. Có tổng cộng 200 video tất cả. 100 trong đó là video bạo lực và 100 còn lại là video không bạo lực. Độ dài video là 2 giây, kích thước khung hình giữa các video không giống nhau toàn bộ và cảnh bạo lực chiếm phần lớn khung hình. Môi trường và con người trong video cũng khác nhau. Những video này cũng có chuyển động nền.

3) Tập dữ liệu PTIT: Đây là tập dữ liệu do chúng tôi thu thập để phục vụ cho nghiện cứu tại Học viện Công nghệ Bưu chính Viễn thông. Tập dữ liệu có tổng cộng 210 video, trong đó 110 video là bạo lực và 90 video là không bạo lực. Những video này có chung kích thước khung hình nhưng độ dài khác nhau, được quay với các bối cảnh khác nhau và khoảng cách tới camera khác nhau từ gần đến xa. Hình số 6 minh họa một số hình ảnh mô tả hành vi bạo lực được trích xuất từ các tập dữ liệu.

B. KẾT QUẢ

Chúng tôi tiến hành thử nghiệm bằng ngôn ngữ Python và thư viện học sâu PyTorch với cấu hình máy tính như sau:

- OS : Windows 10
- CPU : I9-10900K
- RAM : 32GB
- GPU: GEFORCE RTX 2070 SUPER

Thực nghiệm được tiến hành trên cả 3 tập dữ liệu Hockey Fight, Peliculas và PTIT với hai mô hình CNN là Resnet18 và VGG16 [15], hai mô hình LSTM là LSTM truyền thống và Bi-LSTM. Để xem xét sự ảnh hưởng của số lượng timesteps (số lượng khung hình LSTM sử dụng để dự đoán) đến độ chính xác, chúng tôi thử nghiệm với 10 và 15 timesteps. Tập dữ liệu được chia ra 80% cho huấn luyện và 20% cho kiểm tra. Kết quả thực nghiệm được tính toán bằng độ chính xác như trong Bảng 3. Thời gian chạy của Resnet18 với 10 timesteps và 15 timesteps lần lượt là 80ms và 95ms, hoàn toàn phù hợp với những ứng dụng trong thời gian thực.

Kết quả thực nghiệm cho thấy phương pháp chúng tôi phát triển (Fight Region Candidate + Resnet18 + 2 Bi-LSTM) cho kết quả tốt nhất. Phương pháp này đạt độ chính xác cao hơn so với các thuật toán chỉ sử dụng CNN-LSTM đơn thuần mà không có bước tiền xử lý để xác định vùng khả nghi. Trong khi nếu thay Resnet18 bằng VGG16 thì độ chính xác gần như không thay đổi trong khi độ phức tạp của mạng CNN tăng lên rất nhiều.

Thực nghiệm cũng cho thấy mạng LSTM 15 timesteps Bi-LSTM, cho kết quả tốt hơn so với 10 timesteps mà vẫn đảm bảo yêu cầu thời gian thực.

V. HƯỚNG NGHIÊN CỨU TƯƠNG LAI VÀ THẢO LUẬN

Tự động phát hiện hành vi bạo lực là rất quan trọng để kịp thời can thiệp, ngăn chặn và cảnh báo. Từ đó có thể giảm thiểu được thiệt hại cả về sức khỏe, vật chất, lẫn tinh thần cho con người.

Bài báo đã đưa ra phương pháp phát hiện hành vi bạo lực có hiệu quả cao, bằng việc kết hợp tiền xử lý phát hiện

Bảng 3. Kết quả thực nghiệm

	Number of time step	Hockey Fight			Peliculas			PTIT		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Resnet18 + 2 LSTM	10	0.94	0.95	0.94	0.86	0.89	0.87	0.82	0.84	0.83
	15	0.96	0.96	0.96	0.88	0.92	0.9	0.83	0.87	0.85
Resnet18 + 2 Bi-LSTM	10	0.96	0.97	0.96	0.89	0.91	0.9	0.84	0.86	0.85
	15	0.97	0.98	0.97	0.9	0.95	0.92	0.87	0.88	0.87
Fight Region Candidate + Resnet18 + 2 LSTM	10	0.95	0.96	0.95	0.9	0.9	0.9	0.92	0.92	0.92
	15	0.97	0.97	0.97	0.93	0.97	0.95	0.93	0.95	0.94
Fight Region Candidate + Resnet18 + 2 Bi-LSTM	**10**	**1**	**1**	**1**	**0.92**	**0.93**	**0.92**	**0.92**	**0.92**	**0.92**
	15	**1**	**1**	**1**	**0.96**	**0.99**	**0.97**	**0.97**	**0.99**	**0.98**
VGG16 + 2 LSTM	10	0.94	0.94	0.94	0.84	0.86	0.85	0.81	0.82	0.81
	15	0.96	0.96	0.96	0.85	0.89	0.87	0.85	0.86	0.85
VGG16 + 2 Bi-LSTM	10	0.96	0.97	0.96	0.88	0.92	0.9	0.83	0.84	0.83
	15	0.97	0.97	0.97	0.92	0.93	0.92	0.87	0.88	0.87
Fight Region Candidate + VGG16 + 2 LSTM	10	0.95	0.96	0.95	0.88	0.92	0.9	0.89	0.9	0.9
	15	0.97	0.98	0.97	0.94	0.96	0.95	0.92	0.96	0.94
Fight Region Candidate + VGG16 + 2 Bi-LSTM	10	1	1	1	0.92	0.93	0.92	0.92	0.92	0.92
	15	1	1	1	0.96	0.99	0.97	0.95	0.97	0.96

vùng khả nghi và sử dụng mô hình kết hợp CNN-LSTM để phân tích hành vi bạo lực trong cả không gian và thời gian.

Các kết quả thực nghiệm cũng cho thấy phương pháp của chúng tôi đủ nhanh, độ chính xác cao và hoàn toàn phù hợp cho những hệ thống yêu cầu xử lý thời gian thực.

Tuy nhiên phương pháp vẫn còn nhiều điểm hạn chế, như tập dữ liệu vẫn chưa đủ lớn để có thể bao quát được tất cả các trường hợp có thể xảy ra trên thực tế. Chưa được kiểm thử trong nhiều bối cảnh môi trường khác nhau.

Công việc dự kiến trong thời gian tới của nhóm dự án là tiếp tục xây dựng bộ dữ liệu đầy đủ hơn nhằm nâng cao độ chính xác của thuật toán.

Ngoài ra chúng tôi cũng dự định xây dựng một mô hình dạng end-to-end vừa có khả năng xác định vùng khả nghi đồng thời phân loại hành vi bạo lực mà không cần thêm bước tiền xử lý.

TÀI LIỆU THAM KHẢO

[1] I. S. Gracia, O. D. Suarez, G. B. Garcia, and T.-K. Kim, “Fast fight detection,” PLoS ONE, vol. 10, no. 4, Apr. 2015, Art. no. e0120448.

[2] P. C. Ribeiro, R. Audigier, and Q. C. Pham, “RIMOC, a feature to discriminate unstructured motions: Application to violence detection for video-surveillance,” Comput. Vis. Image Understand., vol. 144, pp. 121–143, Mar. 2016.

[3] E. Y. Fu, H. Va Leong, G. Ngai, and S. Chan, “Automatic fight detection in surveillance videos,” in Proc. 14th Int. Conf. Adv. Mobile Comput. Multi Media, Nov. 2016, pp. 225–234.

[4] [S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," in 2017 International Confe-rence on Engineering and Technology (ICET), 2017, pp. 1-6, doi: 10.1109/ICEng-Technol.2017.8308186.

[5] Ralf C. Staudemeyer and Eric Rothstein Morris, “Understanding LSTM - a tutorial into Long Short-Term Memory Recurrent Neural Networks”. arXiv, 2019.

[6] C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia, ‘‘Violence detection in video by using 3D convolutional neural networks,’’ in Proc. Int. Symp. Visual Comput., 2014, pp. 551–558.

[7] S. Sudhakaran and O. Lanz, ‘‘Learning to detect violent videos using convolutional long short-term memory,’’ in Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS), Aug./Sep. 2017, pp. 1–6.

[8] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, ‘‘Violence detection using spatiotemporal features with 3D convolutional neural network,’’ Sensors, vol. 19, no. 11, p. 2472, May 2019.

[9] Seymanur Akti, Gozde Ayse Tataroglu and Hazim Kemal Ekenel, “Vision-based Fight Detection from Surveillance Cameras”. IEEE, 2019.

[10] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," in IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673-2681, Nov. 1997, doi: 10.1109/78.650093.

[11] Joseph Redmon, Santosh Divvala, Ross Girshick and Ali Farhadi, “You Only Look Once: Unified, Real-Time Object Detection”. arXiv, 2016.

[12] Nicolai Wojke, Alex Bewley and Dietrich Paulus, “Simple Online and Realtime Tracking with a Deep Association Metric”. arXiv, 2017.

[13] Alexey Bochkovskiy, Chien-Yao Wang and Hong-Yuan Mark Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection”. arXiv, 2020.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, “Deep Residual Learning for Image Recognition”. arXiv, 2015.

[15] Karen Simonyan and Andrew Zisser-man, “Very Deep Convolutional Networks for Large-Scale Image Recognition”. arXiv, 2015.

MULTISTAGE REAL-TIME VIOLENCE DETECTION USING CONVOLUTIONAL NEURAL NETWORK AND LONG SHORT-TERM MEMORY

Abstract: Action detection is a challenging Computer Vision research topics. It has many practical applications in our lives and violence detection is one of the case that helps quickly prevent and reduce the human injury in a public places equipped with surveillance cameras such as on the streets, at the hospitals, schools or parks. In this study, we propose a detection method which taking the advantages of the convolutional neural network (CNN) and the long short-term memory network (LSTM). At the first stage, the high-risk group of violence is detected by using YOLO (You Only Look Once). CNN is then used to extract the features in stage 2, which will be directly used as input for LSTM at the last stage to predict the final class. The datasets we used in our experiments are Hockey Fight, Peliculas and a self-collected one, PTIT dataset. Experiment results of the proposed method has been compared to some prior works, showing that it is not only effective in detecting the violence but also reduces the number of false positive cases. Our method achieved high performance in detection and has high potential for real-time applications.

Keywords: Violence Detection; Convolutional Neural Network; Long Short-term Memory; YOLO; Hockey Fight; Peliculas.

Nguyễn Mạnh Dũng, tốt đại học chuyên ngành điên tử viễn thông, Đại học Back Khoa Hà Nội năm 2005. Tốt nghiệp Thạc sỹ chuyên ngành công nghệ thông tin, Đại học Quốc gia Kongju năm 2009. Và Tốt nghiệp tiến sỹ chuyên ngành công nghệ thông tin Đại học Quốc gia Kongju năm 2019. Hiện nay đang công tác và giảng dạy tại khoa kỹ thuật điện tử 1, Học Viện Công Nghệ Bưu Chính Viễn Thông. Lĩnh vực yêu thích bao gồm xử lý ảnh, thị giác máy tính, thuật toán và trí tuệ nhân tạo.

Vũ Hoài Nam, tốt nghiệp đại học chuyên ngành điện tử viễn thông, Đại học Bách Khoa Hà Nội năm 2013. Tốt nghiệp Thạc sỹ chuyên ngành Kỹ Sư Máy Tính, Đại học Quốc gia Wangju năm 2015. Hiện nay đang là nghiên cứu sinh chuyên ngành Khoa Học Máy Tính, Học Viện Công Nghệ Bưu Chính Viễn Thông. Lĩnh vực yêu thích bao gồm xử lý ảnh, thị giác máy tính, thuật toán và trí tuệ nhân tạo.

Phạm Đức Cường, tốt nghiệp đại học chuyên ngành Hệ Thống Thông Tin, Học Viện Công Nghệ Bưu Chính Viễn Thông, Hà Nội. Hiện nay đang công tác tại IVS, với vị trí kỹ sư nghiên cứu và phát triển các thuật toán xử lý, nhận dạng hình ảnh. Lĩnh vực yêu thích bao gồm xử lý ảnh, thị giác máy tính, học máy và trí tuệ nhân tạo.

Nguyễn Việt Hưng. Tốt nghiệp thạc sĩ năm 2009 tại ĐH Bách Khoa Grenoblem và bảo vệ luận án Tiến sỹ năm 2013 tại đại học Rennes 1, CH Pháp. Hiện công tác tại Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: Hệ thống thông tin thế hệ mới, trí tuệ nhân tạo, học máy.