Hệ Thống Nhận Dạng và Cảnh Báo Bạo Lực Trong Đám Đông

Nguyễn Anh Cương
Khoa Công nghệ Thông tin
Đại học Đại Nam
Hà Nội, Việt Nam

Vũ Khánh Hoàn
Khoa Công nghệ Thông tin
Đại học Đại Nam
Hà Nội, Việt Nam

Nông Trung Hiếu
Khoa Công nghệ Thông tin
Đại học Đại Nam
Hà Nội, Việt Nam

***Tóm tắt nội dung*—Bài báo này đề xuất một hệ thống nhận dạng và phân loại hành vi bạo lực trong đám đông từ dữ liệu video theo thời gian thực. Phương pháp tiếp cận của chúng tôi sử dụng Mạng Nơ-ron Tích Chập 3D (3D CNN) để trích xuất đặc trưng không gian-thời gian từ chuỗi video, cho phép nhận diện chính xác các sự cố bạo lực. Bằng cách kết hợp IoT với cảm biến thông minh, hệ thống thu thập và xử lý dữ liệu thời gian thực để cải thiện hiệu quả của cơ chế cảnh báo bạo lực. Kết quả thực nghiệm trên các bộ dữ liệu công khai chứng minh rằng mô hình 3D CNN của chúng tôi phân biệt hiệu quả giữa các cảnh bạo lực và không bạo lực. Hệ thống đề xuất có tiềm năng ứng dụng đáng kể trong nhiều lĩnh vực như giám sát an ninh, kiểm soát đám đông và hỗ trợ cơ quan thực thi pháp luật.**

***Index Terms*—nhận dạng bạo lực, mạng CNN 3D, đặc trưng không gian-thời gian, giám sát đám đông, IoT, giám sát thời gian thực**

I. Giới thiệu

Hệ thống giám sát video ngày càng trở nên quan trọng để đảm bảo an toàn và an ninh công cộng trong môi trường đông đúc. Tuy nhiên, các phương pháp giám sát truyền thống dựa vào con người gặp phải nhiều thách thức đáng kể, bao gồm sự mệt mỏi khi tập trung chú ý và không có khả năng theo dõi liên tục nhiều nguồn video cùng một lúc. Hạn chế này làm nổi bật nhu cầu về các hệ thống tự động có khả năng phát hiện và cảnh báo các sự cố bạo lực theo thời gian thực.

Bài báo này giới thiệu một phương pháp tiếp cận mới cho việc phát hiện bạo lực trong đám đông sử dụng Mạng Nơ-ron Tích Chập 3D. Hệ thống của chúng tôi nhằm mục đích nhận diện và phân loại hành vi bạo lực trong môi trường đám đông từ dữ liệu video theo thời gian thực. Tiềm năng ứng dụng của hệ thống bao gồm nhiều lĩnh vực như giám sát an ninh, kiểm soát đám đông và hỗ trợ các cơ quan thực thi pháp luật.

II. Các nghiên cứu liên quan

Phát hiện bạo lực trong video là một lĩnh vực nghiên cứu năng động trong thị giác máy tính. Các phương pháp ban đầu dựa vào các đặc trưng được thiết kế thủ công như quỹ đạo chuyển động, biểu đồ gradient hướng (HOG) và luồng quang học. Tuy nhiên, các phương pháp này thường gặp khó khăn với các cảnh phức tạp và điều kiện môi trường thay đổi.

Những tiến bộ gần đây trong học sâu đã dẫn đến những cải tiến đáng kể trong hệ thống phát hiện bạo lực. Các mô hình CNN 2D áp dụng cho từng khung hình riêng lẻ đã cho thấy kết quả đầy hứa hẹn nhưng không nắm bắt được động lực thời gian quan trọng để hiểu các hành động bạo lực. Mạng nơ-ron hồi quy (RNN) và mạng bộ nhớ dài-ngắn hạn (LSTM) đã được sử dụng để mô hình hóa các phụ thuộc thời gian, nhưng chúng thường yêu cầu các bước trích xuất đặc trưng riêng biệt.

CNN 3D, mở rộng phép tích chập đến chiều thời gian, đã nổi lên như một phương pháp tiếp cận mạnh mẽ cho các tác vụ nhận dạng hành động. Bằng cách xử lý chuỗi các khung hình đồng thời, CNN 3D có thể học các đặc trưng không gian-thời gian trực tiếp từ dữ liệu video thô, làm cho chúng đặc biệt phù hợp cho các ứng dụng phát hiện bạo lực.

Hình 1. Mô hình 3D CNN cho nhận dạng hành vi bạo lực.

III. Phương pháp đề xuất

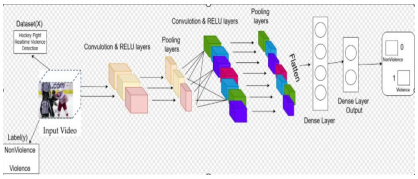
Hệ thống đề xuất của chúng tôi bao gồm bốn thành phần chính, được mô tả dưới đây:

A. Tiền xử lý dữ liệu

Bước đầu tiên liên quan đến việc tiền xử lý dữ liệu video để cải thiện độ chính xác của mô hình. Điều này bao gồm:

- Lọc nhiễu để loại bỏ các yếu tố không mong muốn có thể ảnh hưởng đến hiệu suất phát hiện
- Chuẩn hóa khung hình để tiêu chuẩn hóa giá trị pixel và đảm bảo đầu vào nhất quán cho mạng nơ-ron
- Chuyển đổi video thành định dạng tensor, biến đổi các khung hình liên tiếp thành biểu diễn có cấu trúc phù hợp cho xử lý CNN 3D

Dữ liệu video được chuyển đổi thành một chuỗi các khung hình liên tiếp và được biểu diễn dưới dạng tensor với cấu trúc: [batch_size, number_of_frames, height, width, channels].



Hình 2. Sơ đồ 3D CNN

B. Trích xuất đặc trưng không gian-thời gian

Chúng tôi sử dụng kiến trúc CNN 3D để trích xuất các đặc trưng không gian-thời gian có ý nghĩa từ dữ liệu video đã được tiền xử lý. Không giống như CNN 2D truyền thống hoạt động trên từng khung hình riêng lẻ, CNN 3D áp dụng kernel 3D (thường là 3×3×3) để tích chập trên cả chiều không gian và chiều thời gian. Phương pháp này cho phép mô hình học các mẫu dựa trên sự thay đổi của các khung hình theo thời gian, điều này là cần thiết để phân biệt các hành động bạo lực khỏi hành vi bình thường.

CNN 3D trích xuất các đặc trưng phân cấp thông qua nhiều lớp tích chập, nắm bắt cả các mẫu chuyển động cấp thấp và thông tin ngữ nghĩa cấp cao về các hoạt động diễn ra trong video.

C. Mô hình phân loại

Các đặc trưng không gian-thời gian được trích xuất được đưa vào một mô-đun phân loại xác định liệu một đoạn video có chứa hành vi bạo lực hay không. Mô hình của chúng tôi được đào tạo để phân loại video thành hai lớp: bạo lực và không bạo lực. Việc phân loại được thực hiện bằng cách sử dụng các lớp kết nối đầy đủ xử lý biểu diễn đặc trưng và xuất ra điểm xác suất cho mỗi lớp.

D. Tích hợp IoT và hệ thống cảnh báo

Để nâng cao tính thực tiễn của hệ thống phát hiện bạo lực, chúng tôi tích hợp công nghệ IoT với cảm biến thông minh để thu thập và xử lý dữ liệu thời gian thực. Sự tích hợp này cho phép:

- Giám sát liên tục nhiều vị trí cùng một lúc
- Cảnh báo ngay lập tức khi phát hiện các sự cố bạo lực
- Khả năng điện toán biên để giảm độ trễ trong các tình huống quan trọng

Khi hệ thống xác định một sự kiện bạo lực với độ tin cậy cao, nó kích hoạt cơ chế cảnh báo có thể thông báo cho nhân viên an ninh hoặc cơ quan thực thi pháp luật, cho phép can thiệp kịp thời.

IV. Thực nghiệm và kết quả

A. Bộ dữ liệu

Chúng tôi đã đánh giá phương pháp đề xuất trên hai bộ dữ liệu công khai:

1) Bộ dữ liệu Hockey Fight: Bộ dữ liệu này bao gồm các video trận đấu khúc côn cầu trên băng chứa cả các sự cố bạo lực (đánh nhau giữa các cầu thủ) và các hành động không bạo lực. Bộ dữ liệu bao gồm:

- 2 lớp: bạo lực và không bạo lực
- Khoảng 30 khung hình cho mỗi đoạn video
- 800 video cho việc huấn luyện
- 200 video cho việc kiểm tra

2) Bộ dữ liệu Real Life Violence Situation: Bộ dữ liệu này chứa các video được quay trong môi trường thế giới thực như đường phố, trung tâm mua sắm và không gian công cộng. Nó bao gồm:

- 2 lớp: bạo lực (đánh nhau, xô xát) và không bạo lực (hoạt động đám đông bình thường)
- Khoảng 50 khung hình cho mỗi đoạn video
- 1600 video cho việc huấn luyện
- 400 video cho việc kiểm tra

B. Chi tiết triển khai

Mô hình CNN 3D của chúng tôi được triển khai với cấu hình sau:

- Các lớp tích chập 3D với kernel 3×3×3 để nắm bắt các đặc trưng không gian-thời gian
- Chuẩn hóa batch sau mỗi lớp tích chập để ổn định việc huấn luyện
- Các lớp gộp cực đại để giảm kích thước không gian và trích xuất các đặc trưng nổi bật
- Regularization dropout để ngăn chặn overfitting
- Các lớp kết nối đầy đủ cho phân loại cuối cùng

Mô hình được huấn luyện sử dụng tối ưu hóa Adam với tốc độ học 0.001 và mất mát entropy chéo phân loại. Việc huấn luyện được tiến hành trong 50 epochs với kích thước batch là 16.

C. Đánh giá hiệu suất

Chúng tôi đánh giá hiệu suất của mô hình sử dụng các chỉ số phân loại tiêu chuẩn, bao gồm:

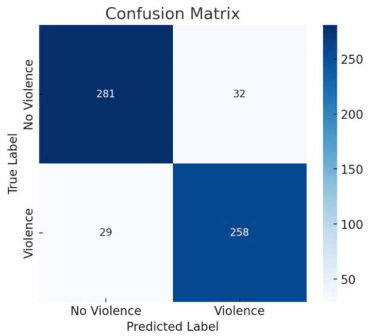
- Độ chính xác (Accuracy): tỷ lệ mẫu được phân loại chính xác
- Độ chính xác (Precision): tỷ lệ dự đoán dương tính thật so với tất cả dự đoán dương tính
- Độ nhạy (Recall): tỷ lệ dự đoán dương tính thật so với tất cả dương tính thực tế
- Điểm F1: trung bình điều hòa của độ chính xác và độ nhạy
- Ma trận nhầm lẫn: trực quan hóa hiệu suất phân loại của mô hình

D. Phân tích kết quả

Các thực nghiệm của chúng tôi đã chứng minh rằng mô hình CNN 3D nắm bắt hiệu quả động lực không gian-thời gian của các hành vi bạo lực, đạt độ chính xác cao trên cả hai bộ dữ liệu. Phân tích ma trận nhầm lẫn cho thấy mô hình phân biệt thành công giữa các cảnh bạo lực và không bạo lực với tỷ lệ dương tính giả và âm tính giả tối thiểu.

Việc tích hợp công nghệ IoT còn nâng cao hiệu suất của hệ thống bằng cách cho phép xử lý thời gian thực và giảm độ trễ

phản hồi. Sự kết hợp giữa kỹ thuật học sâu tiên tiến và cơ sở hạ tầng IoT tạo ra một hệ thống phát hiện bạo lực mạnh mẽ, phù hợp để triển khai trong các môi trường quan trọng về an ninh.



Hình 3. Ma Trân nhâm lân

V. Kết luận và hướng phát triển tương lai

A. Kết luận

Bài báo này đã trình bày một hệ thống phát hiện và cảnh báo bạo lực sử dụng kiến trúc CNN 3D để nhận diện hành vi bạo lực trong đám đông. Kết quả của chúng tôi chứng minh rằng CNN 3D cải thiện đáng kể độ chính xác trong nhận diện hành vi bằng cách nắm bắt hiệu quả các đặc trưng không gian-thời gian từ dữ liệu video. Việc tích hợp IoT và điện toán biên nâng cao khả năng phản hồi của hệ thống, làm cho nó trở thành một công cụ có giá trị cho việc giám sát an ninh và giảm thiểu rủi ro bạo lực.

B. Hướng phát triển tương lai

Tiến tới, chúng tôi lên kế hoạch tập trung vào một số cải tiến và mở rộng cho hệ thống hiện tại:

- Tối ưu hóa mô hình AI để tăng tốc độ xử lý và giảm tài nguyên tính toán
- Mở rộng khả năng phân loại để nhận diện các loại hành vi bạo lực cụ thể
- Triển khai hệ thống trong các ứng dụng thực tế như giám sát giao thông và sự kiện đông người
- Khám phá kỹ thuật học chuyển giao để cải thiện hiệu suất trên dữ liệu huấn luyện hạn chế
- Phát triển các cơ chế bảo vệ quyền riêng tư cho việc triển khai công khai

Lời cảm ơn

Chúng tôi xin cảm ơn Trường Đại học Đại Nam đã cung cấp cơ sở nghiên cứu và hỗ trợ trong suốt dự án này. Chúng tôi cũng ghi nhận công của các tác giả đã tạo ra các bộ dữ liệu công khai được sử dụng trong các thực nghiệm của chúng tôi.

Tài liệu

[1] S. Ji, W. Xu, M. Yang, và K. Yu, "3D convolutional neural networks for human action recognition,"IEEE Transactions on Pattern Analysis and Machine Intelligence, tập 35, số 1, tr. 221-231, 2013.

[2] D. Tran, L. Bourdev, R. Fergus, L. Torresani, và M. Paluri, "Learning spatiotemporal features with 3D convolutional networks,"trong Proceedings of the IEEE International Conference on Computer Vision, 2015, tr. 4489-4497.

[3] E. Bermejo, O. Deniz, G. Bueno, và R. Sukthankar, "Violence detection in video using computer vision techniques,"trong Computer Analysis of Images and Patterns, 2011, tr. 332-339.

[4] T. Hassner, Y. Itcher, và O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior,"trong IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012, tr. 1-6.

[5] P. Zhou, Q. Ding, H. Luo, và X. Hou, "Violence detection in surveillance video using low-level features,"PLOS ONE, tập 13, số 10, 2018.

[6] Y. Gao, H. Liu, X. Sun, C. Wang, và Y. Liu, "Violence detection using oriented violent flows,"Image and Vision Computing, tập 48, tr. 37-41, 2016.

[7] A. Datta, M. Shah, và N. Da Vitoria Lobo, "Person-on-person violence detection in video data,"trong Proceedings of the 16th International Conference on Pattern Recognition, 2002, tr. 433-438.