

VIDEO SUMMARIZATION USING BACKGROUND SUBTRACTION TECHNIQUES

Ngo Huu Huy^{1*}, Le Hung Linh¹, Nguyen Duy Minh¹, Ngo Thi Thu Hang²

¹TNU – University of Information and Communication Technology

²Kim Dong Secondary School - Ha Long City, Quang Ninh

ARTICLE INFO	ABSTRACT
Received: 24/02/2022	Multimedia information systems have been massively and diversely used in research and practical applications. Among them, video data is one of the most common data types. However, the management and use of videos have faced difficulties, such as organizing storage or finding events in a video. Therefore, this study presents an efficient and simple method based on the background subtraction technique for video summarization. First, the input video is used to extract consecutive frames. These frames are then preprocessed, such as converting to grayscale images and image smoothing. The background subtraction technique is used to detect movement in the current frame relative to the previous frame. If this frame has motion detection, it will be saved for the output video. We also propose a video summarization algorithm. Experimental results demonstrate the effectiveness of this method, especially for video surveillance.
Revised: 25/4/2022	
Published: 11/5/2022	
KEYWORDS	
Background subtraction	
Motion detection	
Motion tracking	
Video summarization	
Video surveillance	

TÓM TẮT VIDEO SỬ DỤNG KỸ THUẬT TRỪ NỀN

Ngô Hữu Huy^{1*}, Lê Hùng Linh¹, Nguyễn Duy Minh¹, Ngô Thị Thu Hằng²

¹Trường Đại học Công nghệ Thông tin và Truyền thông - ĐHTT Thái Nguyên

²Trường THCS Kim Đồng - Thành phố Hạ Long, Quảng Ninh

THÔNG TIN BÀI BÁO	TÓM TẮT
Ngày nhận bài: 24/02/2022	Hệ thống thông tin đa phương tiện đã được sử dụng rộng rãi và đa dạng trong các nghiên cứu và ứng dụng thực tế. Trong đó, dữ liệu video là một trong những kiểu dữ liệu phổ biến nhất. Tuy nhiên, việc quản lý và sử dụng dữ liệu video gặp nhiều vấn đề như việc tổ chức lưu trữ, hay tìm kiếm sự kiện trong một video. Do đó, nghiên cứu này sẽ trình bày một phương pháp tóm tắt video hiệu quả và đơn giản, sử dụng kỹ thuật trừ nền. Đầu tiên, video đầu vào được sử dụng để trích xuất các khung ảnh liên tiếp. Sau đó, các khung ảnh này sẽ được tiền xử lý, như chuyển sang ảnh đa mức xám và làm mịn ảnh. Kỹ thuật trừ nền được sử dụng để phát hiện chuyển động trong khung ảnh hiện tại so với khung ảnh ngay trước đó. Nếu khung ảnh có phát hiện chuyển động thì nó sẽ được lưu lại cho video đầu ra. Chúng tôi cũng đề xuất một thuật toán tóm tắt video. Các kết quả thực nghiệm cho thấy tính hiệu quả của phương pháp này, đặc biệt đối với các video giám sát.
Ngày hoàn thiện: 25/4/2022	
Ngày đăng: 11/5/2022	
TỪ KHÓA	
Kỹ thuật trừ nền	
Phát hiện chuyển động	
Theo dõi chuyển động	
Tóm tắt video	
Video giám sát	

DOI: <https://doi.org/10.34238/tnu-jst.5582>

* Corresponding author. Email: nh Huy@ictu.edu.vn

1. Giới thiệu

Với sự phát triển rất mạnh mẽ của công nghệ số ngày nay, cùng với đó là sự bùng nổ của các mạng xã hội trong những năm gần đây. Việc tìm kiếm và chia sẻ thông tin của người dùng ngày càng trở nên phổ biến, đặc biệt là các nhu cầu tìm kiếm, truy cập và chia sẻ các dữ liệu đa phương tiện như: âm thanh, hình ảnh, video. Do đó, tóm tắt dữ liệu đa phương tiện mà đặc biệt là tóm tắt video là một chủ đề nghiên cứu rất được quan tâm.

Với sự tiến bộ trong công nghệ video kỹ thuật số, giám sát video đang có một vai trò quan trọng trong việc đảm bảo trật tự và an toàn xã hội. Các hệ thống giám sát được triển khai trong một loạt các ứng dụng và rộng rãi ở khắp nơi như: ở nhà riêng, trường học, cơ quan công sở, nhà máy sản xuất, nơi công cộng. Từ camera giám sát đơn hoặc nhiều camera, một lượng lớn dữ liệu video được tạo ra, lưu trữ và xử lý cho mục đích bảo mật. Tuy nhiên, do video giám sát thường có các đoạn trùng lặp nên việc xem toàn bộ video là một quá trình rất tẻ nhạt và tốn nhiều thời gian đối với các nhà phân tích video. Để khắc phục hạn chế này thì video tóm tắt đã được sử dụng [1].

Bên cạnh đó, với việc phần cứng lưu trữ được nâng cấp và tốc độ internet ngày càng nhanh, việc quay video ngày càng rẻ và tiện lợi hơn. Tuy nhiên, có một lượng lớn thông tin không hiệu quả trong nội dung video cũng được lưu trữ. Do đó, việc tóm tắt video là một vấn đề cấp bách cần được giải quyết, không những có thể tiết kiệm tài nguyên lưu trữ mà còn tiết kiệm thời gian cho những người duyệt video [2].

Trong những năm gần đây, tóm tắt video đã trở thành một lĩnh vực nghiên cứu đầy hứa hẹn và thu hút rất nhiều nghiên cứu. Tóm tắt video là tóm tắt thông tin ngắn gọn về video đã quay và nó cung cấp cho người dùng bản tóm tắt trực quan tổng hợp và hữu ích về chuỗi video. Các kỹ thuật tóm tắt video dựa trên các tính năng [3]-[5], phân cụm [6], sự kiện [7], lựa chọn cảnh quay [8], mạng nơ-ron [9]-[13]. Tuy nhiên, những phương pháp này thường yêu cầu tính toán phức tạp và đặc biệt có những phương pháp yêu cầu về phần cứng phải có hiệu năng cao.

Do đó, nghiên cứu này trình bày một phương pháp tóm tắt video hiệu quả và đơn giản, sử dụng kỹ thuật trừ nền. Điều này hỗ trợ việc tổ chức lưu trữ, hay tìm kiếm sự kiện trong một video được thực hiện dễ dàng hơn, tiết kiệm thời gian và chi phí. Nội dung còn lại của bài nghiên cứu này sẽ được bố cục như sau: Mục 2 mô tả chi tiết phương pháp tóm tắt video. Mục 3 trình bày các kết quả thực nghiệm. Cuối cùng, Mục 4 là phần kết luận và định hướng phát triển nghiên cứu.

2. Phương pháp tóm tắt video

2.1. Cấu trúc hệ thống

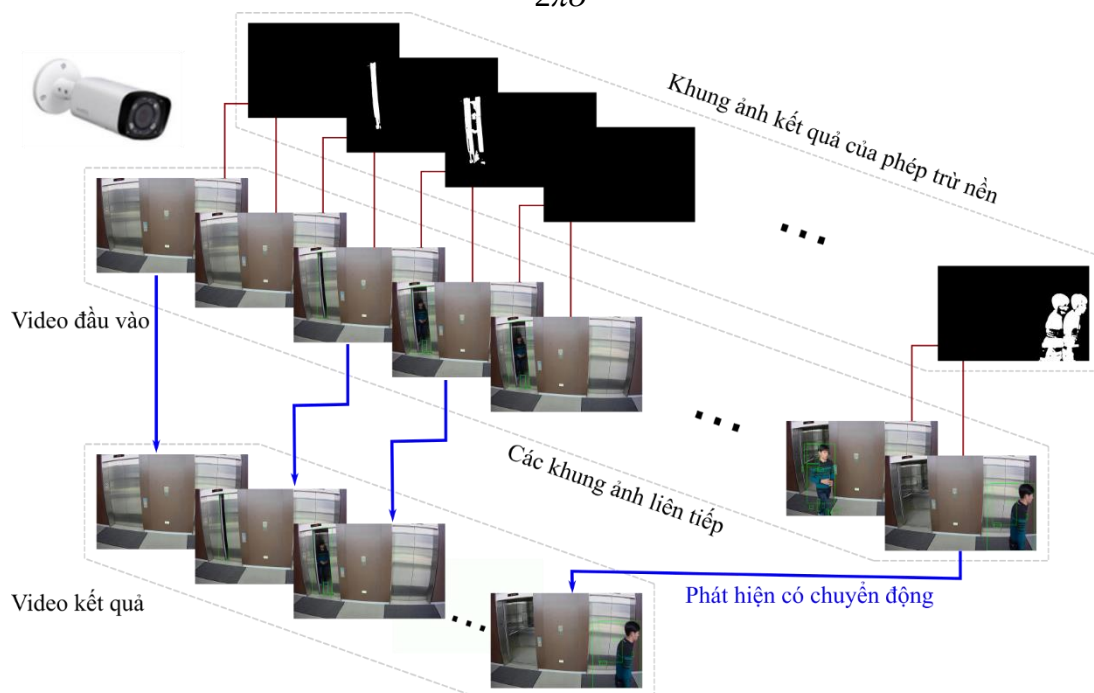
Tổng quan của phương pháp này được thể hiện như trong Hình 1. Dữ liệu đầu vào là các video giám sát và các video này thường có nền ít thay đổi. Đầu tiên, chúng tôi thực hiện trích xuất các khung ảnh liên tiếp từ video đầu vào để thực hiện xử lý trên các khung ảnh riêng biệt đó. Các khung ảnh này sau đó được tiền xử lý như chuyển sang ảnh đa mức xám và làm mịn ảnh. Để xác định khung ảnh nào sẽ được lưu lại cho video đầu ra, chúng tôi cần xác định được khung ảnh chính. Trong nghiên cứu này, khung ảnh chính là khung ảnh có phát hiện chuyển động so với khung ảnh ngay trước đó. Kỹ thuật trừ nền được sử dụng để phát hiện chuyển động. Cuối cùng, video đầu ra bao gồm những khung chính đã được lưu lại.

2.2. Kỹ thuật trừ nền

Trong kỹ thuật trừ nền, màu sắc không ảnh hưởng đến việc phát hiện chuyển động, do đó các khung ảnh màu sẽ được chuyển sang ảnh đa mức xám. Mặt khác, những khung ảnh này có thể chứa các loại nhiễu khác nhau và hai khung ảnh liên tiếp rất khó để giống nhau 100%. Vì vậy, chúng tôi đã làm mịn ảnh giúp giảm nhiễu và phương pháp lọc Gauss được áp dụng trong nghiên cứu này. Trong phương pháp lọc Gauss này, ảnh đầu vào sẽ được nhân chập với một ma trận lọc Gauss và cộng các kết quả lại với nhau để tạo thành giá trị điểm ảnh đầu ra. Giá trị mỗi điểm ảnh

sẽ phụ thuộc nhiều vào các điểm ảnh ở gần hơn là các điểm ảnh ở xa. Hàm Gauss được sử dụng để xác định trọng số của sự phụ thuộc này. Ma trận lọc Gauss [14] hai chiều có thể được xác định như Công thức (1). Trong đó, (x, y) là kích thước của ma trận lọc Gauss và σ là độ lệch chuẩn.

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (1)$$



Hình 1. Tổng quan về phương pháp

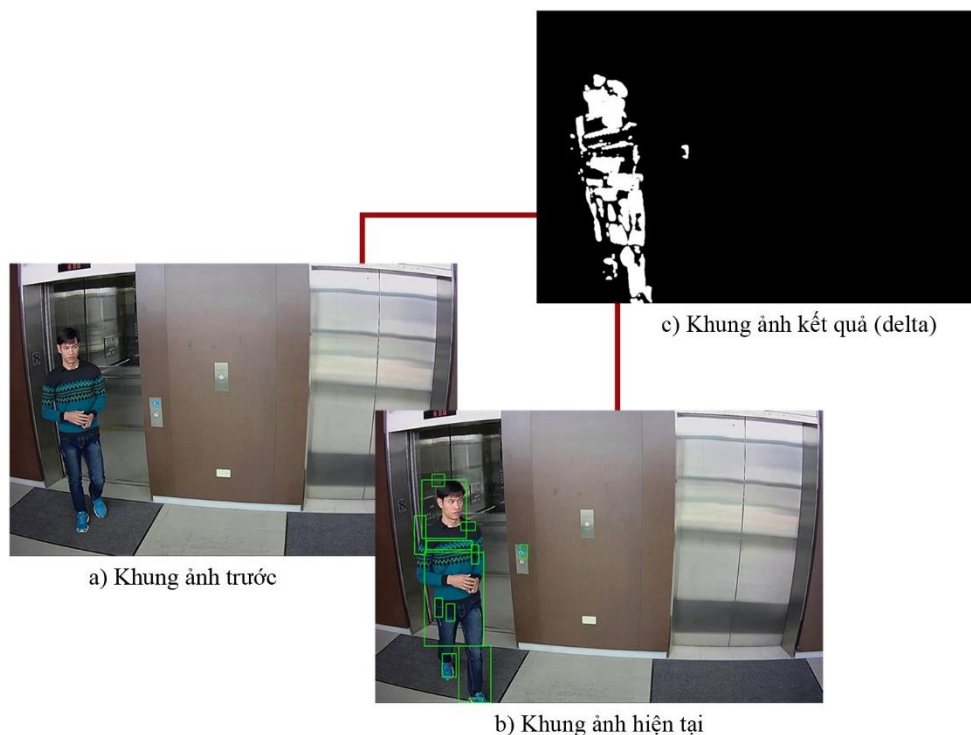
Sau khi thực hiện tiền xử lý, các khung ảnh này được sử dụng để tính toán sự khác biệt giữa khung ảnh hiện tại và khung ảnh ngay trước đó trong luồng video đầu vào. Tính toán sự khác biệt giữa hai khung ảnh là một phép trừ đơn giản như Công thức (2). Trong đó, khung ảnh kết quả (*delta_frame*) lấy giá trị tuyệt đối của sự chênh lệch cường độ điểm ảnh tương ứng của 2 khung ảnh. Hình 2 thể hiện ví dụ minh họa khung ảnh kết quả (*delta_frame*) cho thấy sự khác biệt giữa khung ảnh hiện tại và khung ảnh ngay trước đó. Trong ảnh kết quả cho thấy nền của khung ảnh có màu đen và các vùng có màu trắng là phát hiện có chuyển động. Vì vậy, căn cứ vào hình ảnh kết quả, chúng tôi có thể dễ dàng phát hiện có chuyển động trong khung ảnh hiện tại và khung ảnh này sẽ được lưu lại cho video đầu ra.

$$\text{delta_frame} = |\text{current_frame} - \text{previous_frame}| \quad (2)$$

2.3. Thuật toán tóm tắt video

Chúng tôi đã đề xuất Thuật toán 1 để trình bày phương pháp tóm tắt video dựa trên kỹ thuật trừ nền. Các bước thực hiện chi tiết như sau:

- Dòng (1, 2) thể hiện đầu vào và đầu ra của thuật toán. Đầu vào của thuật toán là video v_1 và giá trị kích thước vùng tối thiểu min_area . Giá trị này để đánh giá vùng chuyển động trong ảnh đủ lớn. Đầu ra của thuật toán là video đã được tóm tắt v_2 .
- Dòng (3-5) thực hiện khởi tạo các biến và chương trình thực hiện duyệt từng khung ảnh trong video đầu vào v_1 để xử lý.
- Dòng (6-8) đọc giá trị từng khung ảnh đầu vào. Sau đó khung ảnh màu này được chuyển sang ảnh đa mức xám. Chương trình sẽ thực hiện làm mịn ảnh đa mức xám này.



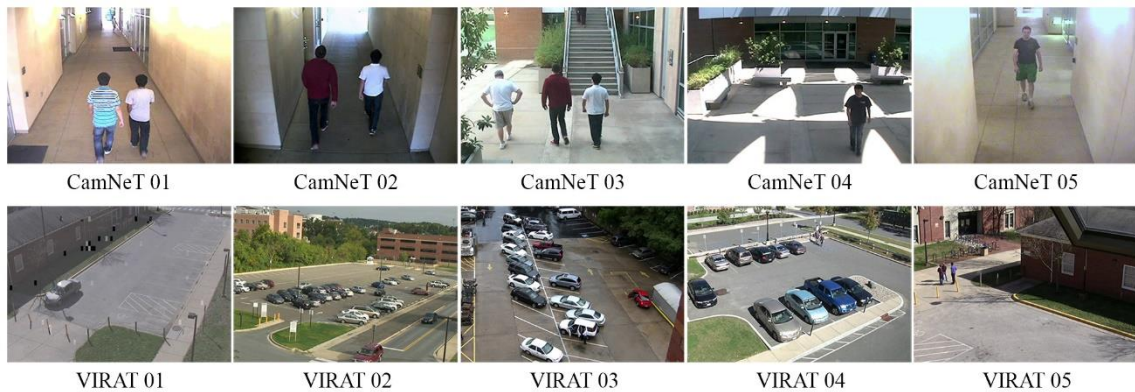
Hình 2. Minh họa khung ảnh kết quả

• Dòng (9-11) thực hiện gán khung ảnh hiện tại là khung ảnh sau khi được tiền xử lý. Tiếp theo, chương trình sẽ thực hiện phép trừ nền và thu được khung ảnh kết quả. Sau đó, khung ảnh kết quả sẽ được dùng để phát hiện chuyển động dựa vào việc phát hiện các vùng trắng sáng trên ảnh đồ.

• Dòng (12-16) duyệt tất cả các vùng trắng sáng (vùng chuyển động) trong khung ảnh kết quả. Nếu tồn tại một vùng sáng đủ lớn (kích thước vùng lớn hơn giá trị kích thước tối thiểu) thì thực hiện lưu lại khung ảnh hiện tại vào video đầu ra v_2 và khung ảnh trước sẽ được gán là khung ảnh hiện tại. Cuối cùng, chương trình trả về kết quả là video đã được tóm tắt v_2 .

Thuật toán 1. Tóm tắt video dựa trên kỹ thuật trừ nền

1. **Đầu vào:** Video đầu vào (v_1), Kích thước vùng tối thiểu (min_area)
 2. **Đầu ra:** Video được tóm tắt (v_2)
 3. Thiết lập giá trị min_area
 4. Thiết lập giá trị khung ảnh trước $previous_frame$
 5. Duyệt từng khung ảnh trong video đầu vào (v_1):
 6. Đọc giá trị khung ảnh ($frame$)
 7. Chuyển khung ảnh màu sang ảnh đa mức xám ($gray_frame$)
 8. Làm mịn ảnh đa mức xám bằng bộ lọc Gauss ($gray_frame$)
 9. Gán giá trị khung ảnh hiện tại $current_frame = gray_frame$
 10. Thực hiện phép trừ nền $delta_frame = |current_frame - previous_frame|$
 11. Phát hiện vùng chuyển động ($find_contours$) trên khung ảnh kết quả ($delta_frame$)
 12. Duyệt các vùng chuyển động ($find_contours$):
 13. Nếu vùng chuyển động $> min_area$:
 14. Lưu khung ảnh hiện tại ($current_frame$) vào video đầu ra (v_2)
 15. Gán giá trị khung ảnh trước $previous_frame = current_frame$
 16. Trả kết quả video tóm tắt (v_2)
-



Hình 3. Ảnh chụp trong các cơ sở dữ liệu thực nghiệm

3. Kết quả thực nghiệm

3.1. Cơ sở dữ liệu thử nghiệm

Cơ sở dữ liệu CamNeT [15] là một cơ sở dữ liệu được thu thập từ mạng lưới các máy quay, với mục đích theo dõi. Số lượng máy quay là từ 5 cho đến 8 máy được đặt ở trong tòa nhà và ngoài trời tại Trường Đại học California, Riverside, Mỹ. Cơ sở dữ liệu này gồm có 6 kịch bản và mỗi kịch bản được thiết kế trong các tình huống khác nhau như ánh sáng thay đổi, địa hình phức tạp, cảnh nhóm người đông đúc. Trong nghiên cứu này, chúng tôi sử dụng 5 video trong cơ sở dữ liệu CamNeT để tiến hành thực nghiệm. Hình 3 là một ví dụ cảnh chụp trong các video. Thông tin chi tiết của 5 video này được thể hiện trong Bảng 1.

Cơ sở dữ liệu VIRAT [16] được thu thập trong các cảnh tự nhiên, hoạt động hàng ngày của mọi người trong nhiều bối cảnh khác nhau, các cảnh nền cũng có nhiều thay đổi. Dữ liệu được thu thập ở nhiều nơi, thu thập các loại hoạt động khác nhau của con người, tương tác giữa người và phương tiện. Để tiến hành thực nghiệm, chúng tôi sử dụng 5 video trong cơ sở dữ liệu VIRAT. Một ví dụ cảnh chụp trong các video được thể hiện như trong Hình 3. Thông tin chi tiết của 5 video này được thể hiện trong Bảng 1.

Bảng 1. Thông tin chi tiết các cơ sở dữ liệu thực nghiệm

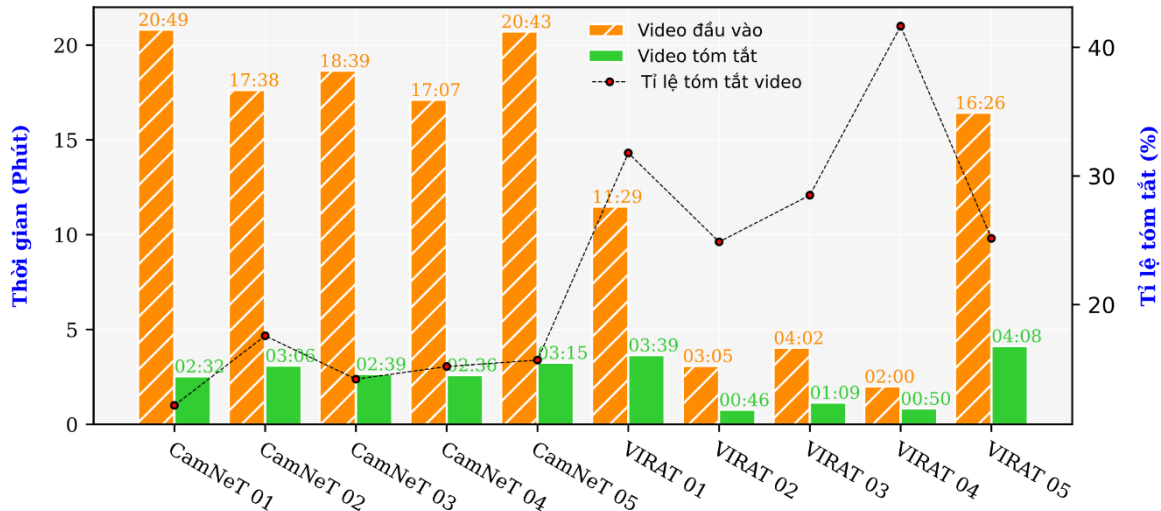
Stt	Các video	Độ dài video (Phút)	Mô tả
1	CamNeT 01	20:49	Video giám sát ở hành lang
2	CamNeT 02	17:38	Video giám sát ở hành lang
3	CamNeT 03	18:39	Video giám sát ở cầu thang
4	CamNeT 04	17:07	Video giám sát ở ngoài trời
5	CamNeT 05	20:43	Video giám sát ở hành lang
6	VIRAT 01	11:29	Video giám sát ở ngoài trời
7	VIRAT 02	03:05	Video giám sát ở bãi đậu xe
8	VIRAT 03	04:02	Video giám sát ở bãi đậu xe
9	VIRAT 04	02:00	Video giám sát ở bãi đậu xe
10	VIRAT 05	16:26	Video giám sát ở ngoài trời

3.2. Đánh giá kết quả tóm tắt video

Trong phần này, chúng tôi thực hiện đánh giá kết quả tóm tắt video. Chúng tôi thực hiện trên 10 video thực nghiệm đã được mô tả ở trên, với tham số kích thước vùng tối thiểu $min_area = 500$ pixel. Hình 4 cho thấy kết quả tóm tắt các video. Chúng tôi thấy rằng việc tóm tắt video là rất hiệu quả. Độ dài thời gian của các video tóm tắt đã giảm đi đáng kể. Để xác định tính hiệu quả, chúng tôi xem xét tỉ lệ tóm tắt được định nghĩa như Công thức (3). Trong đó, α là tỉ lệ tóm tắt, $time_{in}$ là độ dài thời gian của video đầu vào và $time_{out}$ là độ dài thời gian của video tóm tắt.

$$\alpha = \frac{time_{out}}{time_{in}} \times 100\% \quad (3)$$

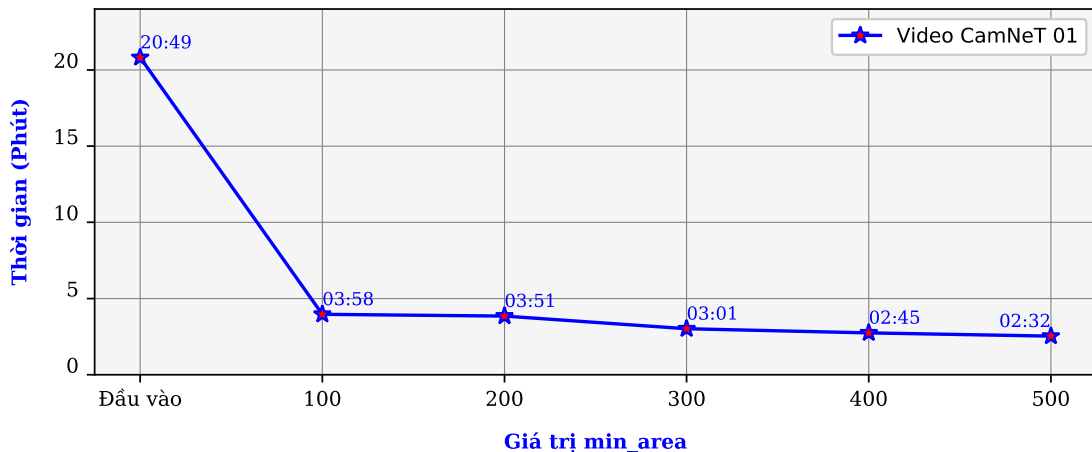
Tỉ lệ tóm tắt càng thấp cho thấy hiệu quả tóm tắt càng tốt. Chúng ta có thể thấy video CamNeT 01 có hiệu quả tóm tắt tốt nhất. Với độ dài thời gian video đầu vào và video tóm tắt tương ứng là 20 phút 49 giây và 02 phút 32 giây, tỉ lệ tóm tắt của video này là 12,17%. Ngược lại, video VIRAT 04 có hiệu quả kém nhất, với tỉ lệ tóm tắt là 41,65%. Tỉ lệ tóm tắt trung bình của 10 video là 22,68%. Kết quả này cho thấy tính hiệu quả rõ ràng của phương pháp với những video giám sát này.



Cơ sở dữ liệu

Hình 4. Kết quả tóm tắt video (Kích thước vùng tối thiểu min_area = 500 pixel)

3.3. Phân tích kết quả dựa trên kích thước vùng tối thiểu (min_area)



Hình 5. Kết quả tóm tắt video CamNeT 01 với các kích thước vùng tối thiểu khác nhau

Kích thước vùng tối thiểu có ảnh hưởng trực tiếp đến kết quả tóm tắt video và giá trị kích thước vùng tối thiểu này được xem xét trong các tình huống, bối cảnh khác nhau: như những nơi có nhiều chuyển động nhỏ có thể bỏ qua, hay những nơi yêu cầu kiểm tra nghiêm ngặt các chuyển động. Trong phần này, chúng tôi thực hiện phân tích kết quả tóm tắt video dựa trên kích thước vùng tối thiểu. Chúng tôi thực hiện trên video CamNeT 01, với tham số kích thước vùng

tối thiểu (min_area) từ 100 pixel đến 500 pixel. Hình 5 cho thấy các kết quả tóm tắt video. Kích thước vùng tối thiểu càng tăng thì video được tóm tắt càng nhiều. Độ dài thời gian video đầu vào là 20 phút 49 giây, kết quả video tóm tắt ứng với kích thước vùng tối thiểu 100 pixel và 500 pixel là 3 phút 58 giây và 2 phút 32 giây. Do đó, việc xác định giá trị kích thước vùng tối thiểu phù hợp với yêu cầu của từng trường hợp là rất quan trọng và nó ảnh hưởng đến hiệu quả của việc tóm tắt video.

4. Kết luận

Nghiên cứu này đã trình bày một phương pháp tóm tắt video hiệu quả và đơn giản, sử dụng kỹ thuật trừ nền. Do đó, nó hỗ trợ việc tổ chức lưu trữ hay tìm kiếm sự kiện trong một video được thực hiện dễ dàng hơn, tiết kiệm thời gian và chi phí. Trong nghiên cứu này, chúng tôi cũng đã phân tích sự ảnh hưởng của kích thước vùng tối thiểu đến việc tóm tắt. Bên cạnh đó, các kết quả thực nghiệm cho thấy tính hiệu quả của phương pháp này, đặc biệt đối với các video giám sát, với tỉ lệ tóm tắt trung bình đạt được là 22,68%. Trong thời gian tới, chúng tôi dự kiến kết hợp với một số phương pháp thông minh sử dụng trí tuệ nhân tạo, để thực hiện tóm tắt video theo sự kiện, ví dụ như chỉ thực hiện lưu lại video khi phát hiện có người hay phương tiện di chuyển.

Lời cảm ơn

Nghiên cứu này là sản phẩm của đề tài cấp cơ sở có mã số T2022-07-01, được tài trợ bởi kinh phí của Trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên.

TÀI LIỆU THAM KHẢO/ REFERENCES

- [1] A. S. Murugan, K. S. Devi, A. Sivaranjani, and P. Srinivasan, "A Study on Various Methods Used for Video Summarization and Moving Object Detection for Video Surveillance Applications," *Multimedia Tools and Applications*, vol. 77, no. 18, pp. 23273-23290, 2018.
- [2] H. Wei, B. Ni, Y. Yan, H. Yu, X. Yang, and C. Yao, "Video Summarization via Semantic Attended Networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 01, pp. 216-223, 2018.
- [3] B. A. Plummer, M. Brown, and S. Lazebnik, "Enhancing Video Summarization via Vision-Language Embedding," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 5781-5789.
- [4] Y. Yuan, T. Mei, P. Cui, and W. Zhu, "Video Summarization by Learning Deep Side Semantic Embedding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 226-237, 2019.
- [5] S. Zhang, Y. Zhu, and A. K. Roy-Chowdhury, "Context-Aware Surveillance Video Summarization," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5469-5478, 2016.
- [6] S. K. Kuanar, K. B. Ranga, and A. S. Chowdhury, "Multi-View Video Summarization Using Bipartite Matching Constrained Optimum-Path Forest Clustering," *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1166-1173, 2015.
- [7] O. Elharrouss, N. Al-Maadeed, and S. Al-Maadeed, "Video Summarization Based on Motion Detection for Surveillance Systems," in *Proc. 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, Tangier, Morocco, 2019, pp. 366-371.
- [8] A. Kanehira, L. Van Gool, Y. Ushiku, and T. Harada, "Viewpoint-Aware Video Summarization," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 7435-7444.
- [9] T. -J. Fu, S. -H. Tai, and H. -T. Chen, "Attentive and Adversarial Learning for Video Summarization," in *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 2019, pp. 1579-1587.
- [10] X. He, Y. Hua, T. Song, Z. Zhang, Z. Xue, R. Ma, N. Robertson, and H. Guan, "Unsupervised Video Summarization with Attentive Conditional Generative Adversarial Networks," in *Proc. 27th ACM International Conference on Multimedia*, New York, NY, USA, 2019, pp. 2296-2304.

-
- [11] Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video Summarization with Attention-Based Encoder-Decoder Networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1709-1717, 2020.
 - [12] M. Rochan and Y. Wang, "Video Summarization by Learning from Unpaired Data," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 7902-7911.
 - [13] M. Rochan, L. Ye, and Y. Wang, "Video Summarization Using Fully Convolutional Sequence Networks," in *Proc. European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 347-363.
 - [14] A. Khumaidi, E. M. Yuniarno, and M. H. Purnomo, "Welding Defect Classification Based on Convolution Neural Network (CNN) and Gaussian Kernel," in *Proc. International Seminar on Intelligent Technology and Its Applications (ISITIA)*, Surabaya, Indonesia, 2017, pp. 261-265.
 - [15] S. Zhang, E. Staudt, T. Faltemier, and A. K. Roy-Chowdhury, "A Camera Network Tracking (CamNeT) Dataset and Performance Baseline," in *Proc. IEEE Winter Conference on Applications of Computer Vision*, Waikoloa, HI, USA, 2015, pp. 365-372.
 - [16] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C. -C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, and M. Shah, "A Large-Scale Benchmark Dataset for Event Recognition in Surveillance Video," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, CO, USA, 2011, pp. 3153-3160.