



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

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

(11)



1-0042185

(51)<sup>2020.01</sup> G06K 9/00

(13) B

(21) 1-2021-06111

(22) 30/09/2021

(45) 25/12/2024 441

(43) 25/01/2022 406

(73) TẬP ĐOÀN CÔNG NGHIỆP - VIỄN THÔNG QUÂN ĐỘI (VN)

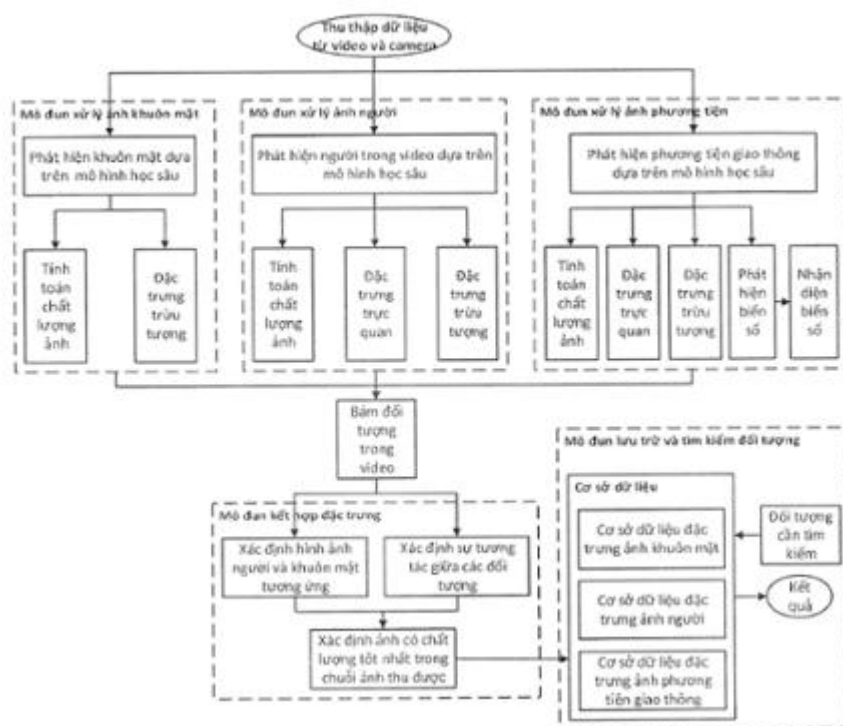
Lô D26 Khu đô thị mới Cầu Giấy, phường Yên Hoà, quận Cầu Giấy, thành phố Hà Nội

(72) Lê Văn Bằng (VN).

(74) Công ty TNHH NACILAW (NACILAW)

(54) PHƯƠNG PHÁP TÌM KIẾM ĐỐI TƯỢNG TRÊN VIDEO DỰA TRÊN ĐẶC TRƯNG PHỨC HỢP CỦA ẢNH

(57) Sáng chế này đề cập đến phương pháp tìm kiếm đối tượng trên video dựa trên đặc trưng phức hợp của ảnh. Trong đó bao gồm mô đun thu thập dữ liệu video từ các camera giám sát; mô đun phát hiện và trích xuất đặc trưng khuôn mặt; mô đun phát hiện và trích xuất đặc trưng hình dáng người; mô đun phát hiện và trích xuất đặc trưng của phương tiện giao thông; mô đun tính toán và kết hợp đặc trưng của các đối tượng trong video và mô đun lưu trữ dữ liệu và tìm kiếm đối tượng. Quá trình thực hiện bao gồm đặc trưng hóa các đối tượng là người và phương tiện trong video, đồng thời có thể tìm kiếm đối tượng theo các đặc trưng trực quan, trừu tượng và sự tương tác với nhau trong video. Sử dụng hệ thống và phương pháp này để tìm kiếm đối tượng trong video giám sát có tính thông dụng cao, độ chính xác tin cậy và khả năng ứng dụng rộng rãi.



### **Lĩnh vực kỹ thuật được đề cập**

Sáng chế này đề cập đến phương pháp tìm kiếm đối tượng trên video dựa trên đặc trưng phức hợp của ảnh. Cụ thể, phương pháp tìm kiếm đối tượng trên video dựa trên đặc trưng phức hợp của ảnh được ứng dụng trong lĩnh vực trí tuệ nhân tạo và thị giác máy tính.

### **Tình trạng kỹ thuật của sáng chế**

Trong những năm gần đây, sự phát triển của trí tuệ nhân tạo đã tạo tiền đề phát triển nhiều ứng dụng trên nền tảng công nghệ thị giác máy, xử lý ảnh. Hiện nay, nhu cầu sử dụng hệ thống camera giám sát bảo đảm an toàn, an ninh tại các khu vực công cộng ngày càng cao. Trong đó, đối tượng được giám sát chính là người và phương tiện giao thông, ngoài ra với lượng dữ liệu video càng ngày càng lớn, yêu cầu đối với hiệu quả tìm kiếm, độ chính xác và tính hữu dụng đối tượng trong video ngày càng cao.

Các hệ thống giám sát qua video thông thường chỉ xử lý các bài toán đơn lẻ, chỉ tập trung vào các yếu tố nhất định, ví dụ như khuôn mặt, biển số xe. Tuy nhiên, loại hình cũng như số lượng đối tượng trong video là rất nhiều, đơn thuần theo dõi, truy vấn theo khuôn mặt hoặc biển số xe đối với phương tiện giao thông sẽ không tận dụng được nguồn thông tin phong phú từ video, tỉ suất sử dụng thông tin thấp và mang tính đặc thù. Khi chỉ sử dụng các yếu tố đơn lẻ thì vô hình chung đã bỏ sót nhiều chi tiết khác trong video, và cả sự tương tác giữa các đối tượng, làm giảm tính hữu dụng của các hệ thống giám sát an ninh, tìm kiếm đối tượng trong thực tế.

Các hệ thống được sử dụng hiện nay còn chưa tận dụng được tính tương quan thông tin trong chuỗi ảnh của video, không có cơ chế cập nhật lại đặc trưng của đối tượng trong chuỗi thời gian đối tượng xuất hiện trong video. Từ đó dẫn đến khó khăn trong quá trình tìm kiếm đối tượng trong video, độ chính xác thấp, thời gian tìm kiếm dài do phải truy vấn trên toàn bộ thời gian mà đối tượng có thể xuất hiện.

### **Bản chất kỹ thuật của sáng chế**

Sáng chế khắc phục các vấn đề kỹ thuật còn tồn tại, như nêu trên, bằng việc đề xuất phương pháp tìm kiếm đối tượng diện rộng dựa trên đặc trưng phức hợp. Sáng chế có mục đích cung cấp mô hình và phương pháp tìm kiếm người và phương tiện trên diện rộng với nhiều camera giám sát khác nhau. Trong đó sử dụng các mô hình học sâu để trích xuất các đặc trưng khuôn mặt, hình dáng người, quần áo, và phương tiện đi kèm để phân tích, bắt bám các đối tượng xuất hiện trong video từ camera giám sát, với tính thông dụng cao, bảo đảm độ chính xác và ổn định.

Nhằm thực hiện mục đích trên, sáng chế phương pháp tìm kiếm đối tượng trên video dựa trên đặc trưng phức hợp của ảnh được thực hiện trên các mô đun như sau: mô đun thu thập dữ liệu video từ các camera giám sát; mô đun phát hiện và trích xuất đặc trưng khuôn mặt; mô đun phát hiện và trích xuất đặc trưng hình dáng người; mô đun phát hiện và trích xuất đặc trưng của phương tiện giao thông; mô đun tính toán và kết hợp đặc trưng từ mô đun thứ hai, thứ ba và thứ tư; mô đun lưu trữ dữ liệu và tìm kiếm đối tượng.

Cụ thể, phương pháp tìm kiếm đối tượng trên video dựa trên đặc trưng phức hợp của ảnh bao gồm các bước thực hiện như sau:

Bước 1: mô đun thu thập dữ liệu thu thập dữ liệu video; dữ liệu ở đây là chuỗi hình ảnh màu đỏ-xanh lục-xanh nước biển (Red-Green-Blue: RGB) theo trục thời gian từ các camera giám sát khác nhau.

Bước 2: mô đun phát hiện và trích xuất đặc trưng khuôn mặt phát hiện khuôn mặt trong hình ảnh, và trích xuất các đặc trưng của khuôn mặt dựa trên các mô hình học sâu.

Bước 3: mô đun phát hiện và trích xuất đặc trưng hình dáng người phát hiện người trong hình ảnh, và trích xuất các đặc trưng trực quan và trừu tượng của hình dáng người dựa trên các mô hình học sâu.

Bước 4: mô đun phát hiện và trích xuất đặc trưng phương tiện giao thông phát hiện phương tiện giao thông trong hình ảnh, và trích xuất các đặc trưng trực quan và trừu tượng của phương tiện giao thông dựa trên các mô hình học sâu.

Bước 5: mô đun tính toán và kết hợp đặc trưng khuôn mặt, hình dáng người, phương tiện giao thông sử dụng, cấu trúc thành đặc trưng phức hợp, mô tả sự tương tác giữa người và phương tiện giao thông được sử dụng tại các thời điểm và địa điểm nhất định.

Bước 6: mô đun lưu trữ dữ liệu và tìm kiếm đối tượng lưu trữ các đặc trưng dạng véc-tơ và dữ liệu hình ảnh của đối tượng, cơ sở dữ liệu này là căn cứ để thực hiện tìm kiếm đối tượng dựa trên đặc trưng phức hợp, bao gồm: đặc trưng khuôn mặt, đặc trưng hình dáng người, đặc trưng phương tiện giao thông được sử dụng.

### **Mô tả vắn tắt các hình vẽ**

Hình 1 là hình vẽ thể hiện kiến trúc hệ thống và phương pháp thực hiện tìm kiếm đối tượng dựa trên đặc trưng phức hợp của ảnh;

Hình 2 là hình vẽ thể hiện quá trình xác định tương tác giữa các đối tượng trong video;

Hình 3 là hình vẽ thể hiện quá trình cập nhật, thay thế đặc trưng của đối tượng trong video dựa trên cơ sở đối chiếu so sánh chất lượng của ảnh.

### Mô tả chi tiết các phương án thực hiện sáng chế

Nhằm mô tả một cách rõ ràng cách thực hiện sáng chế, các bước được mô tả cùng với hình vẽ kèm theo dưới đây. Như mô tả trong Hình 1, các bước thực hiện phương pháp được mô tả cụ thể như sau:

Bước 1: mô đun thu thập dữ liệu thu thập dữ liệu video; dữ liệu video bao gồm các video đã được lưu trữ dạng .mp4, .m4p, .webm, .flv, .avi, .svi và dữ liệu được truyền trực tiếp theo thời gian thực từ các camera giám sát theo đường dẫn rtsp, được mã hóa chuẩn H264-265;

Tại bước này, mô đun thu thập dữ liệu sẽ tiến hành thu thập dữ liệu video, dữ liệu ở đây là chuỗi hình ảnh màu đỏ-xanh lục-xanh nước biển (Red-Green-Blue: RGB) theo thực thời gian từ các camera giám sát khác nhau.

Bước 2: mô đun phát hiện và trích xuất đặc trưng khuôn mặt phát hiện khuôn mặt trong hình ảnh, và trích xuất các đặc trưng của khuôn mặt dựa trên các mô hình học sâu.

Tại bước này, dựa trên dữ liệu video được thu thập từ bước 1, đặc trưng khuôn mặt được trích xuất theo các bước:

B2.1: phát hiện khuôn mặt trong video dựa trên mô hình học sâu.

Cụ thể, hình ảnh thu thập từ video được đưa vào mô hình học sâu để định vị khuôn mặt có trong khung hình, mỗi khuôn mặt có trong ảnh được đánh dấu về vị trí theo hệ tọa độ ảnh, bao gồm tọa độ trung tâm của khuôn mặt, tỉ lệ chiều dài/chiều rộng của ảnh khuôn mặt, và chiều rộng của khuôn mặt.

B2.2: đánh giá chất lượng ảnh của các ảnh khuôn mặt được cắt ra từ video bằng mô hình hồi quy.

Cụ thể, mô hình đánh giá chất lượng ảnh được xây dựng dựa trên mô hình học máy hồi quy tuyến tính, huấn luyện trên dữ liệu khuôn mặt quy mô trên 100000 ảnh khuôn mặt với các dải chất lượng khác nhau được gán nhãn chỉ số chất lượng ảnh từ 0 đến 1 bằng mắt thường, trong đó giá trị càng lớn chất lượng ảnh càng tốt. Các ảnh to, sáng, rõ nét, không mờ, không nhiễu, không bị mất màu, hướng khuôn mặt thẳng với hướng tạo ảnh của camera được coi là ảnh chất lượng tốt, ngược lại là ảnh chất lượng xấu. Đối với ảnh khuôn mặt phát hiện được trong ảnh, được đưa vào mô hình hồi quy đã huấn luyện và tính toán chỉ số chất lượng ảnh, chất lượng ảnh của khuôn mặt tại thời điểm  $t$  là  $Q_{face}^t$ .

B2.3: ảnh khuôn mặt cắt ra từ video được đưa vào mô hình học sâu để trích xuất đặc trưng;

Tại bước này, ảnh khuôn mặt cắt ra từ video được đưa vào mô hình học sâu để trích xuất đặc trưng, đặc trưng khuôn mặt là một véc-tơ có độ dài 512 chiều. Thông tin về đặc trưng trù tượng của khuôn mặt tại thời điểm  $t$ :  $F_{face}^t$ . Đặc trưng của khuôn mặt tại thời điểm  $t$  được tối ưu bởi chuỗi đặc trưng trong khoảng thời gian  $T$  được tính như sau:

$$\bar{F}_{face}^t = w^t \times F_{face}^t + \sum_{\tau=1}^T w^{t-\tau} \times \bar{F}_{face}^{t-\tau}$$

Trong đó  $w^t = Q_{face}^t$ , là trọng số tự thích nghi đối với từng dạng chất lượng ảnh, thuộc phạm vi từ 0~1, tốt nhất tại giá trị bằng 1.

Bước 3: mô đun phát hiện và trích xuất đặc trưng hình dáng người phát hiện người trong hình ảnh, và trích xuất các đặc trưng trực quan và trù tượng của hình dáng người dựa trên các mô hình học sâu.

Tiếp theo, dựa trên dữ liệu video được thu thập từ bước 1, đặc trưng trực quan và trù tượng của hình dáng người trong video được trích xuất theo các bước:

B3.1: phát hiện người trong video dựa trên mô hình học sâu.

Cụ thể, hình ảnh thu thập từ video được đưa vào mô hình học sâu để định vị người có trong khung hình, mỗi người có trong ảnh được đánh dấu về vị trí theo hệ tọa độ ảnh, bao gồm tọa độ trung tâm của người, tỉ lệ chiều dài/chiều rộng của ảnh khuôn mặt, và chiều rộng của người.

B3.2: đánh giá chất lượng ảnh của các ảnh người được cắt ra từ video bằng mô hình hồi quy.

Cụ thể, mô hình đánh giá chất lượng ảnh được xây dựng dựa trên mô hình học máy hồi quy tuyến tính, huấn luyện trên dữ liệu khuôn mặt quy mô trên 100000 ảnh hình người với các dải chất lượng khác nhau được gắn nhãn chỉ số chất lượng ảnh từ 0 đến 1 bằng mắt thường, trong đó giá trị càng lớn chất lượng ảnh càng tốt. Các ảnh to, sáng, rõ nét, không mờ, không nhiễu, không bị mất màu, hướng người thẳng với hướng tạo ảnh của camera được coi là ảnh chất lượng tốt, ngược lại là ảnh chất lượng xấu. Đối với ảnh người phát hiện được trong ảnh, được đưa vào mô hình hồi quy đã huấn luyện và tính toán chỉ số chất lượng ảnh, chất lượng ảnh của hình người tại thời điểm  $t$  là  $Q_{person}^t$ .

B3.3: ảnh người cắt ra từ video được đưa vào mô hình học sâu để trích xuất đặc trưng trực quan;

Tại bước này, tiến hành đưa ảnh người được cắt ra từ video vào mô hình học sâu để trích xuất đặc trưng trực quan. Mô hình này bao gồm hai bộ phận:

Thứ nhất, bộ phận phân loại các đặc trưng bao gồm độ tuổi, giới tính, hướng so với camera.

Thứ hai, bộ phận bóc tách các vị trí mang mặc của người và phụ kiện kèm theo trong ảnh, bao gồm mang mặc thân trên, mang mặc thân dưới, mũ, khăn, kính, khẩu trang, túi xách, ba lô điện thoại, giày dép.

Ảnh người được cắt ra từ video được đưa vào mô hình học sâu đã huấn luyện, tính toán độ tuổi; giới tính sử dụng bộ phận mô hình thứ nhất; đồng thời phân tích đặc điểm mang mặc thân trên, đặc điểm mang mặc thân dưới, đặc điểm tóc, đặc điểm giày dép, đặc điểm các phụ kiện kèm theo bao gồm mũ, kính, khăn, túi xách, ba lô, điện thoại từ bộ phận mô hình thứ hai.

B3.4: ảnh người cắt ra từ video được đưa vào mô hình học sâu để trích xuất đặc trưng trừu tượng sử dụng mô hình học sâu.

Thông tin về đặc trưng trừu tượng của người tại thời điểm  $t$  là  $F_{person}^t$ , trong đó  $D$  là số chiều của đặc trưng. Đặc trưng của khuôn mặt tại thời điểm  $t$  được tối ưu bởi chuỗi đặc trưng trong khoảng thời gian  $T$  được tính như sau:

$$\bar{F}_{person}^t = w^t \times F_{person}^t + \sum_{\tau=1}^T w^{t-\tau} \times \bar{F}_{person}^{t-\tau}$$

Trong đó  $w^t = Q_{person}^t$ , là trọng số tự thích nghi đối với từng dạng chất lượng ảnh, thuộc phạm vi từ 0~1, tốt nhất tại giá trị bằng 1.

Bước 4: mô đun phát hiện và trích xuất đặc trưng phương tiện giao thông phát hiện phương tiện giao thông trong hình ảnh, và trích xuất các đặc trưng trực quan và trừu tượng của phương tiện giao thông dựa trên các mô hình học sâu.

Tiếp theo, dựa trên dữ liệu video được thu thập từ bước 1, đặc trưng trực quan và trừu tượng của phương tiện giao thông trong video được trích xuất theo các bước:

B4.1: phát hiện phương tiện giao thông trong video dựa trên mô hình học sâu.

Cụ thể, hình ảnh thu thập từ video được đưa vào mô hình học sâu để định vị phương tiện có trong khung hình, mỗi phương tiện có trong ảnh được đánh dấu về vị trí theo hệ tọa độ ảnh, bao gồm tọa độ trung tâm của phương tiện, tỉ lệ chiều dài/chiều rộng của ảnh khuôn mặt, và chiều rộng của phương tiện.

B4.2: đánh giá chất lượng ảnh của các ảnh phương tiện giao thông được cắt ra từ video bằng mô hình hồi quy.

Cụ thể, mô hình đánh giá chất lượng ảnh được xây dựng dựa trên mô hình học máy hồi quy tuyến tính, huấn luyện trên dữ liệu khuôn mặt quy mô trên 100000 ảnh hình phương tiện với các dải chất lượng khác nhau được gán nhãn chỉ số chất lượng ảnh từ 0 đến 1 bằng mắt thường, trong đó giá trị càng lớn chất lượng ảnh càng tốt. Các ảnh

to, sáng, rõ nét, không mờ, không nhiễu, không bị mất màu, bao quát được phương tiện được coi là ảnh chất lượng tốt, ngược lại là ảnh chất lượng xấu. Đối với ảnh phương tiện phát hiện được trong ảnh, được đưa vào mô hình hồi quy đã huấn luyện và tính toán chỉ số chất lượng ảnh, chất lượng ảnh của phương tiện tại thời điểm  $t$  là  $Q_{vehicle}^t$ .

B4.3: ảnh phương tiện giao thông cắt ra từ video được đưa vào mô hình học sâu để trích xuất đặc trưng trực quan, bao gồm: loại xe, nhãn hàng phương tiện, màu xe.

B4.4: ảnh phương tiện giao thông cắt ra từ video được đưa vào mô hình học sâu để trích xuất đặc trưng trừu tượng. Thông tin về đặc trưng trừu tượng của phương tiện giao thông tại thời điểm  $t$   $F_{vehicle}^t$ , là một véc-tơ  $D$  chiều. Đặc trưng của khuôn mặt tại thời điểm  $t$  được tối ưu bởi chuỗi đặc trưng trong khoảng thời gian  $T$  được tính như sau:

$$\bar{F}_{vehicle}^t = w^t \times F_{vehicle}^t + \sum_{\tau=1}^T w^{t-\tau} \times \bar{F}_{vehicle}^{t-\tau}$$

Trong đó  $w^t = Q_{vehicle}^t$ , là trọng số tự thích nghi đối với từng dạng chất lượng ảnh, thuộc phạm vi từ 0~1, tốt nhất tại giá trị bằng 1.

B4.5: ảnh phương tiện giao thông cắt ra từ video được đưa vào mô hình học sâu để đọc biển số xe  $LPR_{vehicle}$ .

Bước 5: mô đun tính toán và kết hợp đặc trưng khuôn mặt, hình dáng người, phương tiện giao thông sử dụng, cấu trúc thành đặc trưng phức hợp, mô tả sự tương tác giữa người và phương tiện giao thông được sử dụng tại các thời điểm và địa điểm nhất định.

Tiếp theo, dựa trên các đặc trưng tính toán từ các bước 2, bước 3, bước 4, mối liên hệ giữa các đặc trưng thành phần và cấu thành đặc trưng phức hợp này được xác định như sau:

B5.1: sử dụng thông tin từ quá trình phát hiện đối tượng, thông qua cách tính khoảng cách Mahalanobis và độ tương quan đặc trưng bám theo các đối tượng trong video, bao gồm khuôn mặt, người và phương tiện giao thông. Quá trình này nhằm bám đối tượng bao gồm khuôn mặt, người và phương tiện xuất hiện trong video, mỗi đối tượng được gắn với một mã bám, mã này không đổi đối với cùng một đối tượng khi di chuyển giữa các vị trí khác nhau trong ảnh.

Cụ thể cách thực hiện như sau, các đối tượng bao gồm khuôn mặt, người và phương tiện giao thông trong ảnh đã được xác định tọa độ trung tâm, tỉ lệ chiều dài/chiều rộng, và chiều rộng đối tượng như mô tả trong bước B2.1, B3.1 và B4.1, lần lượt đối với khuôn mặt, người và phương tiện giao thông. Nay khởi tạo bổ sung giá trị vận tốc trong chuỗi ảnh của video từ ba phần tử nêu trên, tạo thành véc-tơ mô tả vị trí không gian và thời gian của đối tượng  $B^t = (u, v, \gamma, h, \dot{x}, \dot{y}, \dot{\gamma}, \dot{h})$ . Hai đối tượng tại hai thời điểm

liên tục  $t-1$  và  $t$  là  $O(t)$  và  $O(t-1)$  được coi là cùng một mã bám đối tượng nếu thỏa mãn:

+ Khoảng cách Mahalanobis của hai véc-tơ mô tả vị trí không gian, thời gian nhỏ hơn ngưỡng  $\Gamma_0$  đã được thiết lập sẵn.

+ Hai véc-tơ đặc trưng  $F_{O(t)}$  và  $F_{O(t-1)}$  có độ dài bằng nhau và có khoảng cách Euclid nhỏ hơn ngưỡng  $\Gamma_0$  đã được thiết lập sẵn.

B5.2: khuôn mặt có mã bám đối tượng  $i$  và người có mã bám đối tượng  $j$ , tại thời điểm  $t$  được coi là cùng một người nếu thỏa mãn công thức dưới đây, đây là công thức lần đầu được nghiên cứu và sử dụng trong sáng chế này:

$$\left\{ \begin{array}{l} Mahalanobis \left( B'_{face, tid=i}, B'_{person, tid=j} \right) \leq \Gamma_1 \\ \left| B'_{person, tid=j} - B'_{face, tid=i} \cup B'_{person, tid=j} \right| \leq \Gamma_2 \\ \left| \left( v_{face} - \frac{h_{face}}{2} \right)_{B'(tid=i)} - \left( v_{person} - \frac{h_{person}}{2} \right)_{B'(tid=j)} \right| \leq \Gamma_3 \end{array} \right.$$

Trong đó,

+ Mahalanobis(.) là khoảng cách Mahalanobis

+  $B'_{person, tid=j}$  và  $B'_{face, tid=i}$  lần lượt là thông tin phát hiện đối tượng từ B2.1 và B3.1 của người và khuôn mặt trong video với mã bám đối tượng là  $i$  và  $j$ .

+  $\Gamma_1, \Gamma_2, \Gamma_3$  lần lượt là ngưỡng được thiết lập để phán đoán khuôn mặt và hình người có phải cùng thuộc một đối tượng trong video hay không.

B5.3: sự tương tác giữa các đối tượng là người, phương tiện giao thông trong video được thực hiện như mô tả trong Hình 2. Trong đó, các đối tượng được coi là có tương tác với nhau khi di chuyển trên cùng một hướng, có vận tốc tương đồng và khoảng cách giữa vị trí trung tâm các đối tượng trong phạm vi cho phép, cụ thể được định nghĩa bởi:

$$Mahalanobis \left( B'_{tid=i}, B'_{tid=j} \right) \leq \Gamma_4$$

Trong đó  $\Gamma_4$  là ngưỡng để xác định các đối tượng có sự tương tác.

Bước 6: mô đun lưu trữ dữ liệu và tìm kiếm đối tượng lưu trữ các đặc trưng dạng véc-tơ và dữ liệu hình ảnh của đối tượng; cơ sở dữ liệu này là căn cứ để thực hiện tìm

kiểm đối tượng dựa trên đặc trưng phức hợp, bao gồm: đặc trưng khuôn mặt, đặc trưng hình dáng người, đặc trưng phương tiện giao thông được sử dụng.

Đặc trưng phức hợp của người, phương tiện giao thông xác định từ bước 5 được lưu trữ và tìm kiếm theo các bước như sau:

B6.1: chọn lọc đặc trưng tốt của đối tượng và lưu trữ vào trong cơ sở dữ liệu. Cụ thể, các dữ liệu về phát hiện và bám đối tượng được tính ở B5.1 được lưu trong khoảng thời gian T. Trong thời gian đó, đặc trưng tiêu biểu nhất của đối tượng được xác định bởi thời điểm t có chất lượng ảnh tương ứng tốt nhất, và danh sách đối tượng khác có sự tương tác với đối tượng đó trong video. Quá trình thực hiện như mô tả trong Hình 3.

Trong quá trình khởi tạo và xây dựng cơ sở dữ liệu, cập nhật các đặc trưng trực quan và trau tượng vào cơ sở dữ liệu khi có đối tượng mới là người, phương tiện xuất hiện trong video. Các đối tượng là người, phương tiện xuất hiện trong video có đặc trưng được lưu trữ theo cấu trúc dữ liệu như sau:

+ Loại hình đối tượng: là người hoặc phương tiện, được xác định ở bước B2.1, B3.1 và B4.1

+ Mã bám đối tượng: là các số nguyên dương sắp xếp theo thứ tự lớn dần theo thời gian xuất hiện trong video của đối tượng, được xác định ở bước B5.1

+ Vị trí xuất hiện: là tọa độ trung tâm của đối tượng theo tọa độ ảnh, được xác định ở bước B5.1.

+ Thời điểm xuất hiện: là thời gian trong video đã được lưu trữ và theo thời gian thực đối với video xử lý trực tiếp từ camera.

+ Đặc trưng trực quan của đối tượng, được xác định ở bước B3.3 và B4.3.

+ Đặc trưng trau tượng của đối tượng, được xác định ở bước B3.4 và B4.4.

+ Nếu là người, lưu trữ đặc trưng trau tượng của khuôn mặt, được xác định ở bước B2.3.

+ Nếu là phương tiện giao thông, số biển số xe được lưu trữ bởi một chuỗi ký tự gồm chữ cái và chữ số, được xác định ở bước B4.5.

+ Danh sách các đối tượng có sự tương tác, được xác định ở bước B5.3.

B6.2: tìm kiếm đối tượng theo đặc trưng trực quan và đặc trưng trau tượng; cụ thể bước này được thực hiện theo các bước như sau:

B6.2.1: tìm kiếm người, phương tiện trong video sử dụng các đặc trưng trực quan thông qua phương pháp truy vấn cơ sở dữ liệu dựa trên so khớp phần tử tìm kiếm.

Cụ thể, căn cứ vào đặc trưng trực quan của đối tượng là người và phương tiện để truy vấn và tìm kiếm trong cơ sở dữ liệu được lưu trữ ở bước B6.1, từ đó liệt kê ra danh sách các đối tượng khớp với yêu cầu tìm kiếm.

B6.2.2: tìm kiếm phương tiện dựa trên thông tin về biển số xe, thông qua phương pháp so khớp các ký tự trên biển số xe trong cơ sở dữ liệu.

Cụ thể, căn cứ vào chuỗi ký tự biển số xe cần tìm kiếm, so sánh với từng biển số xe được lưu trữ trong cơ sở dữ liệu từ bước B6.1, chỉ số so khớp được tính bởi số ký tự giống nhau trên tổng số ký tự của biển số xe cần tìm. Từ đó, liệt kê ra danh sách một số lượng đã định nghĩa trước các biển số xe khớp và các thông tin đi kèm, được sắp xếp theo thứ tự chỉ số so khớp giảm dần.

B6.2.3: tìm kiếm khuôn mặt, người và phương tiện trong video sử dụng các đặc trưng trù tượng thông qua phương pháp tính khoảng cách Euclid đối với các đặc trưng được lưu trữ trong cơ sở dữ liệu, đặc trưng khớp nhất có khoảng cách thấp nhất.

Cụ thể, thực hiện bước B2.1, B2.3 để phát hiện và trích xuất đặc trưng trù tượng đối với ảnh khuôn mặt, bước B3.1, B3.4 để phát hiện và trích xuất đặc trưng trù tượng đối với ảnh người, bước B4.1, B4.4 để phát hiện và trích xuất đặc trưng trù tượng ảnh phương tiện giao thông. Sau đó, các đặc trưng trù tượng được so khớp với toàn bộ dữ liệu của đối tượng tương ứng trong cơ sở dữ liệu lưu trữ ở bước B6.1, chỉ số so khớp được tính bằng khoảng cách Euclid giữa đặc trưng đối tượng cần tìm kiếm và đặc trưng lưu trữ trong cơ sở dữ liệu. Từ đó, liệt kê ra danh sách một số lượng đã định nghĩa trước các đối tượng khớp và thông tin đi kèm, được sắp xếp theo thứ tự chỉ số so khớp tăng dần.

B6.2.4: tìm kiếm theo sự tương tác giữa các đối tượng trong video, các đối tượng được xác định là có tương tác với các đối tượng được tìm kiếm theo các bước B.6.2.1, B.6.2.2, B6,2,3 sử dụng công thức ở bước B.5.3 sẽ được coi là kết quả cho quá trình tìm kiếm.

### **Hiệu quả có thể đạt được**

Sáng chế này là cơ sở để tạo nên một hệ thống tìm kiếm đối tượng hoàn chỉnh, thể hiện được sự tương tác giữa các đối tượng khác nhau trong video. Từ đó tối ưu, hoàn thiện khả năng truy vấn dữ liệu video tại các điểm giám sát, tăng hiệu quả, hiệu suất và độ chính xác trong quá trình tìm kiếm đối tượng và sự kiện trong video trong các giải pháp sử dụng nền tảng công nghệ trí tuệ nhân tạo và thị giác máy tính.

### Yêu cầu bảo hộ

1. Phương pháp tìm kiếm đối tượng trên video dựa trên đặc trưng phức hợp của ảnh bao gồm:

bước 1: mô đun thu thập dữ liệu thu thập dữ liệu video; dữ liệu ở đây là chuỗi hình ảnh màu đỏ-xanh lục-xanh nước biển (red-green-blue: RGB) theo trục thời gian từ các camera giám sát khác nhau;

bước 2: mô đun phát hiện và trích xuất đặc trưng khuôn mặt phát hiện khuôn mặt trong hình ảnh, và trích xuất các đặc trưng của khuôn mặt dựa trên các mô hình học sâu; bước này được thực hiện như sau:

phát hiện khuôn mặt trong video dựa trên mô hình học sâu; cụ thể, hình ảnh thu thập từ video được đưa vào mô hình học sâu để định vị khuôn mặt có trong khung hình, mỗi khuôn mặt có trong ảnh được đánh dấu về vị trí theo hệ tọa độ ảnh, bao gồm tọa độ trung tâm của khuôn mặt, tỉ lệ chiều dài/chiều rộng của ảnh khuôn mặt, và chiều rộng của khuôn mặt;

đánh giá chất lượng ảnh của các ảnh khuôn mặt được cắt ra từ video bằng mô hình hồi quy; cụ thể, mô hình đánh giá chất lượng ảnh được xây dựng dựa trên mô hình học máy hồi quy tuyến tính, huấn luyện trên dữ liệu khuôn mặt quy mô trên 100000 ảnh khuôn mặt với các dải chất lượng khác nhau được gán nhãn chỉ số chất lượng ảnh từ 0 đến 1 bằng mắt thường, trong đó giá trị càng lớn chất lượng ảnh càng tốt; các ảnh to, sáng, rõ nét, không mờ, không nhiễu, không bị mất màu, hướng khuôn mặt thẳng với hướng tạo ảnh của camera được coi là ảnh chất lượng tốt, ngược lại là ảnh chất lượng xấu; đối với ảnh khuôn mặt phát hiện được trong ảnh, được đưa vào mô hình hồi quy đã huấn luyện và tính toán chỉ số chất lượng ảnh, chất lượng ảnh của khuôn mặt tại thời điểm  $t$  là  $Q_{face}^t$ ;

ảnh khuôn mặt cắt ra từ video được đưa vào mô hình học sâu để trích xuất đặc trưng; tại bước này, ảnh khuôn mặt cắt ra từ video được đưa vào mô hình học sâu để trích xuất đặc trưng, đặc trưng khuôn mặt là một véc-tơ có độ dài 512 chiều; thông tin về đặc trưng trừu tượng của khuôn mặt tại thời điểm  $t$ :  $F_{face}^t$ ; đặc trưng của khuôn mặt tại thời điểm  $t$  được tối ưu bởi chuỗi đặc trưng trong khoảng thời gian  $T$  được tính như sau:

$$\bar{F}_{face}^t = w^t \times F_{face}^t + \sum_{\tau=1}^T w^{t-\tau} \times \bar{F}_{face}^{t-\tau}$$

trong đó  $w^t = Q_{face}^t$ , là trọng số tự thích nghi đối với từng dạng chất lượng ảnh, thuộc phạm vi từ 0~1, tốt nhất tại giá trị bằng 1;

bước 3: mô đun phát hiện và trích xuất đặc trưng hình dáng người phát hiện người trong hình ảnh, và trích xuất các đặc trưng trực quan và trừu tượng của hình dáng người dựa trên các mô hình học sâu; bước này được thực hiện như sau:

phát hiện người trong video dựa trên mô hình học sâu; cụ thể, hình ảnh thu thập từ video được đưa vào mô hình học sâu để định vị người có trong khung hình, mỗi người có trong ảnh được đánh dấu về vị trí theo hệ tọa độ ảnh, bao gồm tọa độ trung tâm của người, tỉ lệ chiều dài/chiều rộng của ảnh khuôn mặt, và chiều rộng của người;

đánh giá chất lượng ảnh của các ảnh người được cắt ra từ video bằng mô hình hồi quy; cụ thể, mô hình đánh giá chất lượng ảnh được xây dựng dựa trên mô hình học máy hồi quy tuyến tính, huấn luyện trên dữ liệu khuôn mặt quy mô trên 100000 ảnh hình người với các dải chất lượng khác nhau được gán nhãn chỉ số chất lượng ảnh từ 0 đến 1 bằng mắt thường, trong đó giá trị càng lớn chất lượng ảnh càng tốt; các ảnh to, sáng, rõ nét, không mờ, không nhiễu, không bị mất màu, hướng người thẳng với hướng tạo ảnh của camera được coi là ảnh chất lượng tốt, ngược lại là ảnh chất lượng xấu; đối với ảnh người phát hiện được trong ảnh, được đưa vào mô hình hồi quy đã huấn luyện và tính toán chỉ số chất lượng ảnh, chất lượng ảnh của hình người tại thời điểm  $t$  là  $Q_{person}^t$ ;

ảnh người cắt ra từ video được đưa vào mô hình học sâu để trích xuất đặc trưng trực quan; tại bước này, tiến hành đưa ảnh người được cắt ra từ video vào mô hình học sâu để trích xuất đặc trưng trực quan; mô hình này bao gồm hai bộ phận:

thứ nhất, bộ phận phân loại các đặc trưng bao gồm độ tuổi, giới tính, hướng so với camera;

thứ hai, bộ phận bóc tách các vị trí mang mặc của người và phụ kiện kèm theo trong ảnh, bao gồm mang mặc thân trên, mang mặc thân dưới, mũ, khăn, kính, khẩu trang, túi xách, ba lô điện thoại, giày dép;

ảnh người được cắt ra từ video được đưa vào mô hình học sâu đã huấn luyện, tính toán độ tuổi; giới tính sử dụng bộ phận mô hình thứ nhất; đồng thời phân tích đặc điểm mang mặc thân trên, đặc điểm mang mặc thân dưới, đặc điểm tóc, đặc điểm giày dép, đặc điểm các phụ kiện kèm theo bao gồm mũ, kính, khăn, túi xách, ba lô, điện thoại từ bộ phận mô hình thứ hai;

ảnh người cắt ra từ video được đưa vào mô hình học sâu để trích xuất đặc trưng trừu tượng sử dụng mô hình học sâu;

thông tin về đặc trưng trừu tượng của người tại thời điểm  $t$  là  $F_{person}^t$ , trong đó  $D$  là số chiều của đặc trưng; đặc trưng của khuôn mặt tại thời điểm  $t$  được tối ưu bởi chuỗi đặc trưng trong khoảng thời gian  $T$  được tính như sau:

$$\bar{F}_{person}^t = w^t \times F_{person}^t + \sum_{\tau=1}^T w^{t-\tau} \times \bar{F}_{person}^{t-\tau}$$

trong đó  $w^t = Q_{person}^t$ , là trọng số tự thích nghi đối với từng dạng chất lượng ảnh, thuộc phạm vi từ 0~1, tốt nhất tại giá trị bằng 1;

bước 4: mô đun phát hiện và trích xuất đặc trưng phương tiện giao thông phát hiện phương tiện giao thông trong hình ảnh, và trích xuất các đặc trưng trực quan và trừu tượng của phương tiện giao thông dựa trên các mô hình học sâu; cụ thể bước này được thực hiện như sau:

phát hiện phương tiện giao thông trong video dựa trên mô hình học sâu; cụ thể, hình ảnh thu thập từ video được đưa vào mô hình học sâu để định vị phương tiện có trong khung hình, mỗi phương tiện có trong ảnh được đánh dấu về vị trí theo hệ tọa độ ảnh, bao gồm tọa độ trung tâm của phương tiện, tỉ lệ chiều dài/chiều rộng của ảnh khuôn mặt, và chiều rộng của phương tiện;

đánh giá chất lượng ảnh của các ảnh phương tiện giao thông được cắt ra từ video bằng mô hình hồi quy; cụ thể, mô hình đánh giá chất lượng ảnh được xây dựng dựa trên mô hình học máy hồi quy tuyến tính, huấn luyện trên dữ liệu khuôn mặt quy mô trên 100000 ảnh hình phương tiện với các dải chất lượng khác nhau được gắn nhãn chỉ số chất lượng ảnh từ 0 đến 1 bằng mắt thường, trong đó giá trị càng lớn chất lượng ảnh càng tốt; các ảnh to, sáng, rõ nét, không mờ, không nhiễu, không bị mất màu, bao quát được phương tiện được coi là ảnh chất lượng tốt, ngược lại là ảnh chất lượng xấu; đối với ảnh phương tiện phát hiện được trong ảnh, được đưa vào mô hình hồi quy đã huấn luyện và tính toán chỉ số chất lượng ảnh, chất lượng ảnh của phương tiện tại thời điểm  $t$  là  $Q_{vehicle}^t$ ;

ảnh phương tiện giao thông cắt ra từ video được đưa vào mô hình học sâu để trích xuất đặc trưng trực quan, bao gồm: loại xe, nhãn hàng phương tiện, màu xe;

ảnh phương tiện giao thông cắt ra từ video được đưa vào mô hình học sâu để trích xuất đặc trưng trừu tượng; thông tin về đặc trưng trừu tượng của phương tiện giao thông tại thời điểm  $t$   $F_{vehicle}^t$ , là một véc-tơ  $D$  chiều; đặc trưng của khuôn mặt tại thời điểm  $t$  được tối ưu bởi chuỗi đặc trưng trong khoảng thời gian  $T$  được tính như sau:

$$\bar{F}_{vehicle}^t = w^t \times F_{vehicle}^t + \sum_{\tau=1}^T w^{t-\tau} \times \bar{F}_{vehicle}^{t-\tau}$$

trong đó  $w^t = Q_{vehicle}^t$ , là trọng số tự thích nghi đối với từng dạng chất lượng ảnh, thuộc phạm vi từ 0~1, tốt nhất tại giá trị bằng 1;

bước 5: mô đun tính toán và kết hợp đặc trưng khuôn mặt, hình dáng người, phương tiện giao thông sử dụng, cấu trúc thành đặc trưng phức hợp, mô tả sự tương tác giữa người và phương tiện giao thông được sử dụng tại các thời điểm và địa điểm nhất định; cụ thể bước này được thực hiện như sau:

sử dụng thông tin từ quá trình phát hiện đối tượng, thông qua cách tính khoảng cách Mahalanobis và độ tương quan đặc trưng bám theo các đối tượng trong video, bao gồm khuôn mặt, người và phương tiện giao thông; quá trình này nhằm bám đối tượng bao gồm khuôn mặt, người và phương tiện xuất hiện trong video, mỗi đối tượng được gắn với một mã bám, mã này không đổi đối với cùng một đối tượng khi di chuyển giữa các vị trí khác nhau trong ảnh; cụ thể cách thực hiện như sau, các đối tượng bao gồm khuôn mặt, người và phương tiện giao thông trong ảnh đã được xác định tọa độ trung tâm, tỉ lệ chiều dài/chiều rộng, và chiều rộng đối tượng như mô tả trong bước 2, bước 3 và bước 4, lần lượt đối với khuôn mặt, người và phương tiện giao thông; nay khởi tạo bổ sung giá trị vận tốc trong chuỗi ảnh của video từ ba phần tử nêu trên, tạo thành véc-tơ mô tả vị trí không gian và thời gian của đối tượng  $B^t = (u, v, \gamma, h, \dot{x}, \dot{y}, \dot{\gamma}, \dot{h})$ ; hai đối tượng tại hai thời điểm liên tục  $t-1$  và  $t$  là  $O(t)$  và  $O(t-1)$  được coi là cùng một mã bám đối tượng nếu thỏa mãn:

khoảng cách Mahalanobis của hai véc-tơ mô tả vị trí không gian, thời gian nhỏ hơn ngưỡng  $\Gamma_0$  đã được thiết lập sẵn;

hai véc-tơ đặc trưng  $F_{O(t)}$  và  $F_{O(t-1)}$  có độ dài bằng nhau và có khoảng cách Euclid nhỏ hơn ngưỡng  $\Gamma'_0$  đã được thiết lập sẵn;

khuôn mặt có mã bám đối tượng  $i$  và người có mã bám đối tượng  $j$ , tại thời điểm  $t$  được coi là cùng một người nếu thỏa mãn công thức dưới đây, đây là công thức lần đầu được nghiên cứu và sử dụng trong sáng chế này:

$$\left\{ \begin{array}{l} Mahalanobis \left( B'_{face, tid=i}, B'_{person, tid=j} \right) \leq \Gamma_1 \\ \left| B'_{person, tid=j} - B'_{face, tid=i} \cup B'_{person, tid=j} \right| \leq \Gamma_2 \\ \left| \left( v_{face} - \frac{h_{face}}{2} \right)_{B'(tid=i)} - \left( v_{person} - \frac{h_{person}}{2} \right)_{B'(tid=j)} \right| \leq \Gamma_3 \end{array} \right.$$

trong đó,

mahalanobis(.) là khoảng cách mahalanobis;

$B'_{person, tid=j}$  và  $B'_{face, tid=i}$  lần lượt là thông tin phát hiện đối tượng từ bước

2 và bước 3 của người và khuôn mặt trong video với mã bám đối tượng là i và j;

$\Gamma_1, \Gamma_2, \Gamma_3$  lần lượt là ngưỡng được thiết lập để phán đoán khuôn mặt và hình người có phải cùng thuộc một đối tượng trong video hay không;

sự tương tác giữa các đối tượng là người, phương tiện giao thông trong video; trong đó, các đối tượng được coi là có tương tác với nhau khi di chuyển trên cùng một hướng, có vận tốc tương đồng và khoảng cách giữa vị trí trung tâm các đối tượng trong phạm vi cho phép, cụ thể được định nghĩa bởi:

$$Mahalanobis \left( B'_{tid=i}, B'_{tid=j} \right) \leq \Gamma_4$$

trong đó  $\Gamma_4$  là ngưỡng để xác định các đối tượng có sự tương tác;

bước 6: mô đun lưu trữ dữ liệu và tìm kiếm đối tượng lưu trữ các đặc trưng dạng véc-tơ và dữ liệu hình ảnh của đối tượng, cơ sở dữ liệu này là căn cứ để thực hiện tìm kiếm đối tượng dựa trên đặc trưng phức hợp, bao gồm: đặc trưng khuôn mặt, đặc trưng hình dáng người, đặc trưng phương tiện giao thông được sử dụng; cụ thể bước này được thực hiện như sau:

các dữ liệu về phát hiện và bám đối tượng được tính ở bước 5 được lưu trong khoảng thời gian T; trong thời gian đó, đặc trưng tiêu biểu nhất của đối tượng được xác định bởi thời điểm t có chất lượng ảnh tương ứng tốt nhất, và danh sách đối tượng khác có sự tương tác với đối tượng đó trong video; trong quá trình khởi tạo và xây dựng cơ sở dữ liệu, cập nhật các đặc trưng trực quan

và trừu tượng vào cơ sở dữ liệu khi có đối tượng mới là người, phương tiện xuất hiện trong video;

tìm kiếm đối tượng theo đặc trưng trực quan và đặc trưng trừu tượng; cụ thể:

tìm kiếm người, phương tiện trong video sử dụng các đặc trưng trực quan thông qua phương pháp truy vấn cơ sở dữ liệu dựa trên so khớp phần tử tìm kiếm;

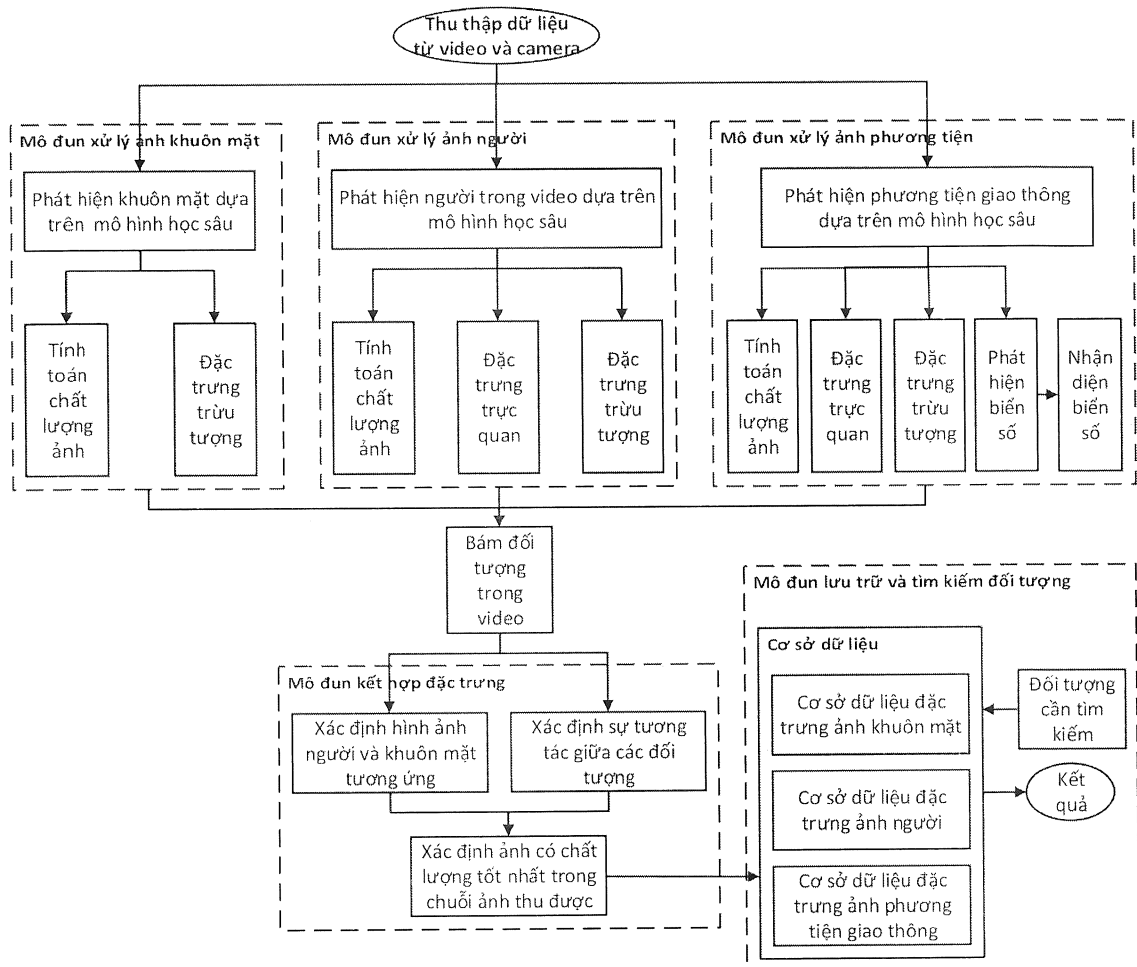
tìm kiếm phương tiện dựa trên thông tin về biển số xe, thông qua phương pháp so khớp các ký tự trên biển số xe trong cơ sở dữ liệu;

tìm kiếm khuôn mặt và người trong video sử dụng các đặc trưng trừu tượng thông qua phương pháp tính khoảng cách euclid đối với các đặc trưng được lưu trữ trong cơ sở dữ liệu, đặc trưng khớp nhất có khoảng cách thấp nhất;

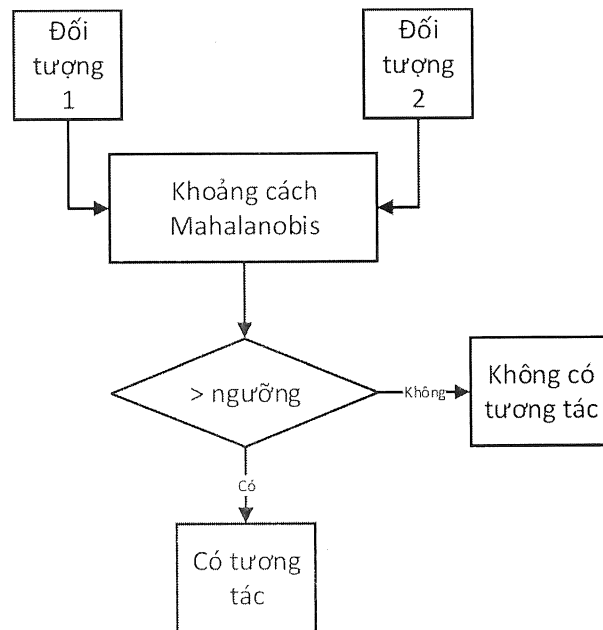
tìm kiếm theo sự tương tác giữa các đối tượng trong video, các đối tượng được xác định là có tương tác với các đối tượng được tìm kiếm theo các bước tìm kiếm vừa được trình bày ở trên sử dụng công thức

$$Mahalanobis\left(\begin{matrix} B^i, B^j \\ tid=i \quad tid=j \end{matrix}\right) \leq \Gamma_4 \text{ ở bước 5 sẽ được coi là kết quả cho quá trình}$$

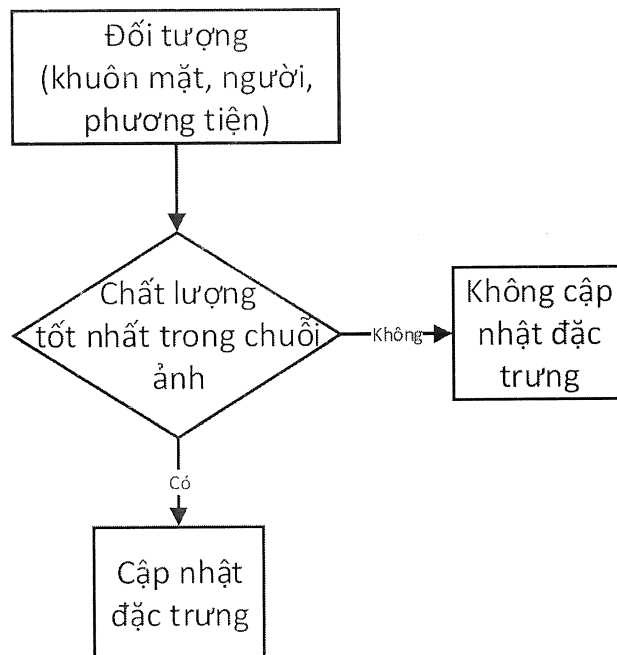
tìm kiếm.



Hình 1



Hình 2



Hình 3