Tạp chí Khoa học Trường Đại học Cần Thơ
Phần A: Khoa học Tự nhiên, Công nghệ và Môi trường
website: sj.ctu.edu.vn

DOI:10.22144/ctu.jsi.2017.001

ỨNG DỤNG KỸ THUẬT THEO DÕI ĐỐI TƯỢNG CHO BÀI TOÁN NHẬN DẠNG HÀNH VI CỦA KHÁCH HÀNG TRONG SIÊU THỊ

Trần Thị Hồng Ân, Phạm Nguyên Khang và Trần Minh Tân
Khoa Công nghệ Thông tin và Truyền thông, Trường Đại học Cần Thơ

Thông tin chung:
Ngày nhận bài: 15/09/2017
Ngày nhận bài sửa: 10/10/2017
Ngày duyệt đăng: 20/10/2017

Title:
Application of object tracking techniques in the analysist of activity customer in suppermarket

Từ khóa:
Nhận dạng hành vi, phân tích hoạt động, theo dõi đối tượng, video giám sát

Keywords:
Activity analysis, behavior recognition, object tracking, video surveillance

ABSTRACT

This paper presented a model using object tracking techniques to categorize the activities of customers in the supermarket. Then, the number of customers, who were interested in the booth, were determined to evaluate the display efficiency. With the image obtained from the surveillance camera, the system can identify most of the objects entering the observation area, tracking them to obtain the trajectory and time of observation. Trajectory was segmented, and representative coordinates were used, thus using a support vector learning algorithm to classify customer activity including booth attendance and drop-in options or other activities. Also, this article proposed the improvements of the speed of object tracking algorithms in the case of tracking multiple objects at the same time. Experimentally, it found that the proposed speed improvements were significantly effective, averaging 2.8 times higher than the original, while accuracy was not changed. Data for detecting was collected from internet sources and surveillance camera data located at a large supermarket in Soc Trang province.

TÓM TẮT

Chúng tôi trình bày mô hình sử dụng các kỹ thuật theo dõi đối tượng để phân loại hoạt động của khách hàng trong siêu thị; từ đó xác định số lượng khách hàng quan tâm đến gian hàng và đánh giá hiệu quả trưng bày. Với hình ảnh thu được từ camera giám sát, hệ thống có thể nhận dạng được hầu hết các đối tượng là người đi vào vùng quan sát, theo dõi họ để có được quỹ đạo đường đi và thời gian lưu lại vùng quan sát. Quỹ đạo được phân đoạn và lấy tọa độ đại diện, sau đó dùng giải thuật máy học véc-tơ hỗ trợ để phân loại hoạt động của khách hàng gồm có quan tâm đến gian hàng và ghé vào lựa chọn hoặc là các hoạt động còn lại. Ngoài ra, trong bài báo, chúng tôi đề xuất các cải tiến nhằm cải thiện tốc độ của giải thuật theo dõi đối tượng trong trường hợp theo dõi nhiều đối tượng cùng lúc. Qua thực nghiệm, chúng tôi nhận thấy các đề xuất cải thiện tốc độ có hiệu quả đáng kể, trung bình tăng 2,8 lần so với ban đầu, trong khi độ chính xác không thay đổi. Dữ liệu nhận dạng người và nhận dạng hoạt động của khách hàng ở siêu thị được thu thập từ nguồn internet và dữ liệu thu được của camera giám sát đặt tại một siêu thị lớn ở tỉnh Sóc Trăng.

Trích dẫn: Trần Thị Hồng Ân, Phạm Nguyên Khang và Trần Minh Tân, 2017. Ứng dụng kỹ thuật theo dõi đối tượng cho bài toán nhận dạng hành vi của khách hàng trong siêu thị. Tạp chí Khoa học Trường Đại học Cần Thơ. Số chuyên đề: Công nghệ thông tin: 1-9.

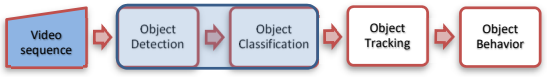
1 GIỚI THIỆU

Trong lý thuyết quản trị, việc nhận biết và đánh giá được hành vi khách hàng để đưa các chiến lược hiệu quả đóng vai trò quan trọng cho sự thành công trong kinh doanh. Trong đó, có một phần là đánh giá được hành vi, biểu hiện của khách hàng khi họ đến tham quan mua sắm tại các khu vực trưng bày, các gian hàng kinh doanh.

Để quyết định được một chiến thuật trưng bày, một thông tin không thể thiếu hỗ trợ cho người quản lý là phải đánh giá được hành vi của khách hàng qua các con số thống kê, tự động hóa việc nhận biết và đánh giá hành vi là một biện pháp hiện đại mang tính chính xác cao rất cần thiết trong trường hợp này.

Việc xây dựng hệ thống nhận dạng hành vi đối tượng thông qua việc áp dụng kỹ thuật theo dõi đối tượng qua camera xuất phát từ nhu cầu đó.

Hệ thống theo dõi đối tượng nhận vào các khung hình video thu nhận từ các camera, qua một số bước xử lý, phân tích và cuối cùng là đưa ra quỹ đạo đường đi của đối tượng theo thời gian làm cơ sở cho việc tiếp theo là nhận biết hành vi. Bắt đầu là quá trình phát hiện đối tượng chuyển động trong các khung hình. Sau đó các đối tượng này sẽ qua quá trình phân lớp để phân biệt các đối tượng thuộc lớp nào, sự vật nào như: con người, xe, mây bay, cây lắc lư …. Tiếp theo là quá trình xử lý để theo vết nhằm tìm ra đường chuyển động của đối tượng, từ đó có thể phân tích, nhận biết hành vi của đối tượng. Hình 1 giới thiệu mô hình theo dõi đối tượng chuyển động, trong đó mỗi bước là một lĩnh vực nghiên cứu rộng lớn.



Hình 1: Mô hình theo dõi đối tượng chuyển động (Ragland & Tharcis, 2014)

2 THEO DÕI ĐỐI TƯỢNG THEO GIẢI THUẬT CMT (CONSENSUS-BASED TRACKING AND MATCHING)

Giải thuật CMT gốc đã được G.Nebehay khởi tạo từ năm 2014 (Nebehay & Pflugfelder, 2014), sau đó vào năm 2015 CMT được cải tiến thành giải thuật CMT thích nghi (Nebehay & Pflugfelder, 2015). Cũng giống với CMT, CMT thích nghi sử dụng vùng khởi tạo b_0 ở khung hình đầu tiên của video. Trong vùng này rút ra được tập các keypoint $P_0 = \{x_1^0, \dots, x_m^0\}$. Một cặp so khớp m_i được định nghĩa là $m_i = (x_i^t, x_i^0)$, với x_i^t là vị trí mới của x_i^0 trong khung ảnh thứ t. Ở khung ảnh thứ t, ta cần phải xác định được một tập các cặp so khớp $\mathcal{L}_t = \{m_1, \dots, m_n\}$ biểu diễn cho đối tượng đang theo dõi càng chính xác càng tốt.

2.1 Tương ứng thích nghi tĩnh (static-adaptive correspondencs)

Mô hình hình dáng tĩnh dựa trên mô hình hình dáng khởi tạo của đối tượng được tạo nên từ bộ mô tả của tập các keypoint $P_0 = \{x_1^0, \dots, x_m^0\}$. Chúng ta gọi các cặp so khớp thu được từ mô hình này là sự tương ứng tĩnh. Trước hết, chúng ta sử dụng tìm kiếm toàn cục để thiết lập các cặp so khớp giữa các keypoint x_i^0 trong khung ảnh khởi tạo và x_i^t trong khung ảnh hiện tại bằng cách sử dụng một ngưỡng (threshold) và điều kiện dựa trên khoảng cách láng giềng gần nhất (Lowe, 2004) sử dụng khoảng cách d(.,.) giữa các mô tả của keypoint.

$$d(x_i^0, x_j^t) < \theta \wedge \frac{d(x_i^0, x_j^t)}{d(x_i^0, x_k^t)} < \gamma, j \neq k \qquad (1)$$

Ngoài ra, ta cần phải loại bỏ các keypoint ứng viên mà nó được so khớp với các keypoint nền trong khung ảnh đầu tiên. Mô hình tĩnh này rất hiệu quả và có thể xử lý được các trường hợp phát hiện lại các keypoint sau khi bị che khuất. Tuy nhiên, mô hình này không thể phát hiện thêm các keypoint mới cho đối tượng đang được theo dõi.

Ngược lại, mô hình thích nghi được cập nhật lại ở mỗi khung ảnh, bao gồm các vùng ảnh nhỏ xung quanh các keypoint $x_i^{t-1} \in \mathcal{L}_{t-1}$. Trong khi ở mô hình tĩnh, ta cần phải tìm kiếm sự tương ứng trên toàn bộ khung ảnh hiện tại, thì với mô hình thích nghi ta có thể giả sử thời gian giữa hai khung ảnh liên tiếp tương đối nhỏ. Vì thế, ta có thể thiết lập sự tương ứng một cách hiệu quả bằng cách sử dụng luồng quang học từ khung ảnh t – 1 đến khung ảnh t. Hợp của các keypoint được tìm kiếm bằng cách so khớp toàn cục và các keypoint có được do truy vết từ các keypoint trong khung ảnh t – 1 tạo thành tập các keypoint $\mathcal{L}_t^*$

2.2 Gom cụm các tương ứng

Ý tưởng chính là sử dụng độ đo D phản ánh sự khác nhau giữa các tương ứng m_i và m_j dựa trên khả năng tương thích hình học của chúng, điều này phản ánh trực tiếp sự biến dạng của đối tượng. Độ đo D được định nghĩa như sau:

$$D(m_i, m_j) = \left\| (x_i^t - Hx_i^0) - (x_j^t - Hx_j^0) \right\| \qquad (2)$$

Trong đó $\|\cdot\|$ là khoảng cách Eclide và H là phép biến đổi tương đồng được ước lượng từ $\mathcal{L}_t^*$.

D được sử dụng để phân hoạch $\mathcal{L}_t^*$ vào trong các tập con sử dụng giải thuật phân cụm phân cấp từ trên

xuống có sử dụng ngưỡng δ. Giả sử $\mathcal{L}_t^+$ là cụm lớn nhất chứa các keypoint liên quan đến đối tượng.

Trong đó, tỷ lệ s và α được tính theo công thức (3) và (4).

$$s = med\left(\left\{\frac{x_i^t - x_j^t}{x_i^0 - x_j^0}, i \neq j\right\}\right) \quad (3)$$

$$\alpha = med(\{atan2(x_i^0 - x_j^0) - atan2(x_i^t - x_j^t)\}) \quad (4)$$

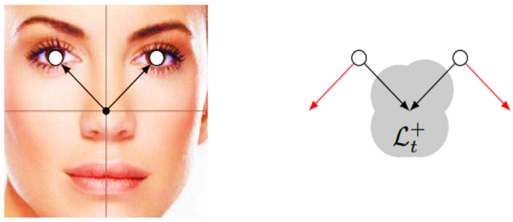
2.3 Tránh nhầm lẫn cho các tương ứng

Bộ mô tả hình dáng tương tự nhau trên nhiều phần của đối tượng hoặc nền dẫn đến vấn đề chính của việc so khớp dựa trên các bộ mô tả (Hình 2). Để loại bỏ các keypoint ứng viên gây nhầm lẫn, cần dựa vào sự khác nhau về mặt hình học trong tập $\mathcal{L}_t^+$ trong lần so khớp thứ hai.

Thay vì phải so khớp keypoint thứ i trong P_t với tất cả các keypoint tĩnh trong P_0, ta chỉ cần so khớp trong tập con.

$$P_0^i = \left\{x_j^0 \middle| \min_{m_k \in \mathcal{L}_t^+} D\left((x_j^0, x_i^t), m_k\right) < \delta\right\} \quad (5)$$

Tập này bao gồm các keypoint trong P_0 cách $\mathcal{L}_t^+$ một khoảng D nhỏ hơn δ. Các điều kiện so khớp trong lần hai tương tự lần một trong mục 2.1. $\mathcal{L}_t^+$ kết hợp với các keypoint đã được loại bỏ sự nhầm lẫn trong lần so khớp thứ hai sẽ hình thành tập $\mathcal{L}_t$ chứa tất cả các keypoint thuộc đối tượng đang theo dõi trong khung ảnh t.



Hình 2: Hình trái: các keypoint có bộ mô tả tương tự nhau rất khó so khớp nếu chỉ dựa trên bộ mô tả. Hình phải: tránh nhầm lẫn khi so khớp các keypoint bằng cách loại bỏ các ứng viên tương ứng dựa vào sự khác nhau về mặt hình học trong tập $\mathcal{L}_t^+$

2.4 Kết quả đầu ra của giải thuật

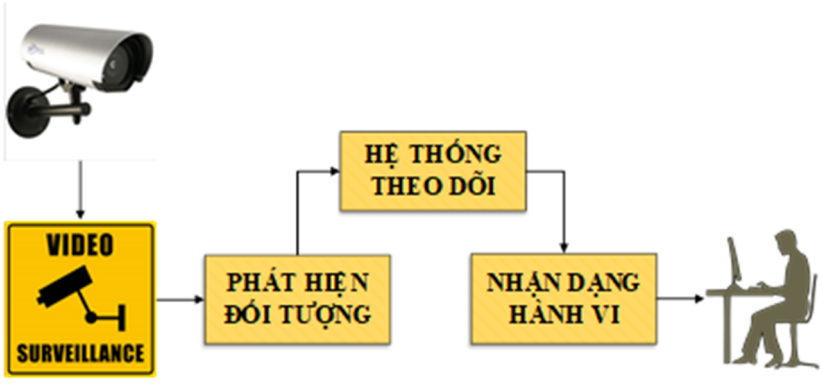
Tiêu chuẩn đầu ra của giải thuật theo dõi là một hình chữ nhật bao quanh đối tượng. Vì vậy, chúng ta tính toán tâm của đối tượng theo công thức:

$$\mu = \frac{1}{|\mathcal{L}_t|}\sum_{m_i \in \mathcal{L}_t}(x_i^t - Hx_i^0) \quad (6)$$

Với các thông số tâm đối tượng μ, tỷ lệ *s,* và góc quay α tư thế của đối tượng đang theo dõi được xác định tương tự như giải thuật CMT.

3 MÔ HÌNH HỆ THỐNG

Hình ảnh thu nhận từ camera giám sát sẽ được chuyển đến hệ thống phát hiện đối tượng đi vào vùng quan sát, thông tin đối tượng kích hoạt hệ thống khởi động bộ theo dõi đến khi đối tượng biến mất (bị che khuất) hoặc rời khỏi vùng quan sát, tiếp theo thông tin quỹ đạo thu được từ bước theo dõi sẽ chuyển qua giai đoạn nhận dạng hoạt động (hành vi) của khách hàng là có quan tâm đến gian hàng đang được trưng bày hoặc không.



Hình 3: Các bước chính trong mô hình theo dõi đối tượng và nhận dạng hành vi

Ở Hình 3, hệ thống gồm có các giai đoạn chính: nhận thông tin đầu vào từ video giám sát, phát hiện đối tượng (người trong siêu thị), theo dõi với giải thuật CMT, nhận dạng hành vi.

Trong giai đoạn tiếp nhận thông tin đầu vào hệ thống: dữ liệu video giám sát có thể được truyền trực tiếp từ camera hoặc dữ liệu video đã được ghi hình và lưu lại vào thiết bị lưu trữ.

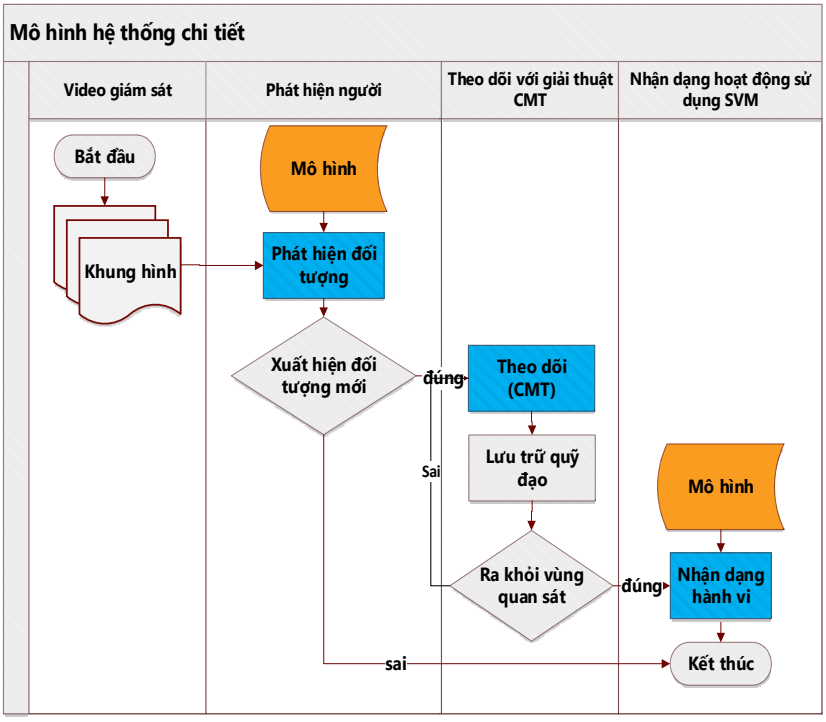
Hình 4 mô tả chi tiết các bước xử lý trong mô hình

3.1 Phát hiện người

Sau khi tiếp nhận dữ liệu video giám sát, cần xác định phạm vi khu vực quan sát, tiếp theo là giai đoạn nhận dạng người trong siêu thị nhờ vào model đã được xây dựng trước theo phương pháp Cascade Classifier (Viola & Jones, 2001).

Kết quả của bước này là một danh sách các khung hình chữ nhật bao quanh đối tượng phát hiện được trong khung hình hiện tại. Tuy nhiên, một số đối tượng sẽ không nằm trong phạm vi của khu vực cần quan sát hoặc là đối tượng đã được phát hiện từ khung hình trước.

Vấn đề đặt ra ở đây là "Làm sao phân biệt được đối tượng nào là đối tượng mới phát hiện lần đầu và đối tượng nào đã được phát hiện ở khung hình trước và đang được theo dõi ở khung hình hiện tại?"



Hình 4: Mô hình chi tiết hệ thống theo dõi đối tượng qua camera giám sát và nhận dạng hành vi

Để giải quyết vấn đề này, chúng tôi đề xuất một phương pháp dựa trên độ chồng lấp của khung bao đối tượng, độ chồng lấp được tính toán bằng diện tích chồng lấp của 2 hình chữ nhật: một là kết quả của đối tượng mới được phát hiện và một thể hiện cho đối tượng đang được theo dõi.

Giả sử ta có hai hình chữ nhật R_1 và R_2 với các thông số về tọa độ gốc, chiều rộng, chiều cao lần lượt là (R_1.x, R_1.y), R_1.w, R_1.h và (R_2.x, R_2.y), R_2.w, R_2.h. Độ chồng lấp $d(r1, r2)$ được tính như sau:

$$O.w = \max(0, \min(R_1.x, R_2.x) - \max(R_1.w, R_2.w))$$

$$O.h = \max(0, \min(R_1.y, R_2.y) - \max(R_1.h, R_2.h))$$

$$d(r1, r2) = \max(\frac{O.w * O.h}{R_1.w * R_1.h}, \frac{O.w * O.h}{R_2.w * R_2.h})$$

Độ chồng lấp sẽ có một ngưỡng để phân biệt, nếu vượt khỏi ngưỡng này thì xem như đối tượng đã được phát hiện trước đó và ngược lại đây là đối tượng mới. Để xác định được ngưỡng này, chúng ta cần xem xét dựa trên môi trường cài đặt hệ thống, ở những môi trường có mật độ xuất hiện các đối tượng dày đặc, đan xen lẫn nhau thì chúng ta sử dụng ngưỡng thấp (<50%), nếu như ở môi trường thông thoáng, các đối tượng trong ảnh ít bị che khuất thì cần đặt thông số ngưỡng cao (>50%) (Hình 5).

Phát hiện đối tượng trong ảnh một cách chính xác là một bài toán có lịch sử lâu đời, tuy nhiên các phương pháp nhận dạng chủ yếu dựa trên các mô tả đặc trưng cục bộ. Với cách này đòi hỏi khối lượng tính toán rất lớn dẫn đến giảm tốc độ thực thi toàn hệ thống. Trong khi đó, cách tính độ chồng lấp với vài bước tính toán đơn giản, giảm tối đa ảnh hưởng tới tốc độ thực thi toàn hệ thống. Vì vậy, lựa chọn phân ngưỡng độ chồng lấp là lựa chọn tối ưu cho

những giải pháp tích hợp nhiều kỹ thuật khác nhau, mà bản thân các kỹ thuật này có khối lượng tính toán lớn.

Hình 5: Mô tả độ chồng lấp (a: trên, b: dưới)

Hình 5a bên trái là khung hình đầu tiên của video, bộ phát hiện cho kết quả là 2 đối tượng được khoanh vùng màu đen (hình a, bên trái), do đối tượng không chồng lấp lên đối tượng đang theo dõi nên đây là đối tượng mới phát hiện. Ở Hình 5a bên phải, bộ phát hiện vẫn phát hiện được 2 đối tượng như khung hình 5a trái, nhưng do 2 đối tượng này chồng lấp khung hình màu trắng (kết quả của giải thuật theo dõi) nên đây không phải là đối tượng mới xuất hiện.

3.2 Theo dõi đối tượng

Sau khi phát hiện đối tượng mới xuất hiện đi vào vùng quan sát, việc khởi tạo bộ theo dõi được kích hoạt và thực hiện việc theo vết đối tượng ở khung hình tiếp theo. Việc theo dõi lúc này không đơn thuần là một đối tượng mà là nhiều đối tượng cùng lúc. Các bước chính được minh hoạ trong Hình 6.

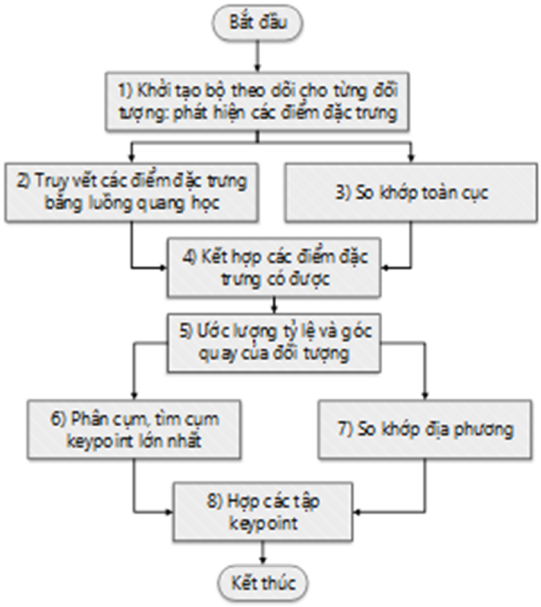
1) Khởi tạo bộ theo dõi cho từng đối tượng phát hiện và tính toán mô tả các đặc trưng cục bộ. Giải thuật CMT của tác giả G. Nebehay (Nebehay & Pflugfelder, 2014) sử dụng bộ phát hiện đặc trưng FAST (Rosten & Drummond, 2006) và bộ mô tả đặc trưng BRISK (Leutenegger *et al.*, 2011), nhưng trong hệ thống này chúng tôi sử dụng bộ phát hiện đặc trưng FAST và bộ mô tả đặc trưng ORB (Rublee *et al.*, 2011) do tốc độ xử lý nhanh hơn, xem mục 3.4.1.

2) Truy vết các đặc trưng cục bộ của từng đối tượng trong khung ảnh trước đó bằng phương pháp luồng quang học (optical flow (Lucas *et al.*, 1981)). Mục tiêu là xác định vị trí của các điểm đặc trưng đại diện cho đối tượng được xác định ở khung hình hiện tại, tập các điểm đặc trưng này gọi là active keyoints.

(3) So khớp toàn cục: mỗi đặc trưng trong khung ảnh hiện tại được so khớp với các đặc trưng của khung ảnh khởi tạo để tìm các cặp điểm tương đồng.

(4) Hợp các điểm đặc trưng thu được ở bước (2) và bước (3), các keypoint trùng nhau bị loại bỏ.

(5) Ước lượng tỷ lệ và góc quay của đối tượng ở khung ảnh hiện hành so với đối tượng xác định ở khung ảnh khởi tạo, hệ thống chỉ theo dõi người theo dáng đứng thẳng nên thông số góc quay được bỏ qua.



Hình 6: Các bước theo dõi đối tượng

(6) Tìm cụm lớn nhất bằng phương pháp consensus, có sử dụng giải thuật phân cụm phân cấp từ trên xuống.

(7) So khớp cục bộ: bước 6 tìm được cụm lớn nhất cũng có nghĩa là tìm được tâm đối tượng. Từ đó có thể loại bỏ một số keypoint trong tập P0 bằng cách tính khoảng cách từ các keypoint ban đầu trong P0 đến tâm đối tượng, nếu lớn hơn một ngưỡng nào đó (threshold_cutoff) thì keypoint đó bị loại. Như vậy, ở lần so khớp này, mỗi điểm đặc trưng trong khung ảnh hiện tại được so khớp (cục bộ) với các keypoint trong tập P0 đã loại bỏ đi một số điểm nào đó xa trung tâm đối tượng.

(8) Hợp các đặc trưng trong cụm lớn nhất tìm được ở bước (6) và các đặc trưng được so khớp lần hai ở bước (7) thành một tập duy nhất. Kết quả cho ra tập các đặc trưng theo dõi được của khung ảnh hiện hành $\mathcal{L}_t$. Kết quả này sẽ là các active keypoint được dùng để truy vết cho khung ảnh tiếp theo.

Với các active point là kết quả của bước (8), khung bao của đối tượng cần theo dõi (bounding box) được tính toán bằng cách tính khung bao của các đặc trưng kết hợp với tỷ lệ và góc quay so với khung bao của đối tượng trong khung ảnh khởi tạo.

Ở mỗi khung hình, bước theo dõi đối tượng bằng giải thuật CMT sẽ xác định được vị trí tâm đối tượng. Với nhiều khung hình liên tiếp, ta sẽ thu được một danh sách vị trí đối tượng tạo nên vĩ đạo đường đi từ khi đối tượng xuất hiện đến khi biến mất khỏi hệ thống.

3.3 Nhận dạng hành vi

Các phương pháp sử dụng trong việc nhận dạng hành vi được phân thành ba loại chính dựa trên công nghệ hiện tại (Lavee *et al*., 2009): gồm phương pháp qua mô hình nhận dạng mẫu (pattern recognition model), phương pháp dựa trên mô hình trạng thái (stage-base), phương pháp dựa trên mô hình ngữ nghĩa (sematic-base).

Trong ba phương pháp, phương pháp thứ nhất sử dụng các kỹ thuật phân loại cơ bản, không đòi hỏi phải có một phát minh đặc biệt giành cho nó. Điểm thuận lợi là các phương pháp đã được chứng minh là có cơ sở khoa học vững chắc và đã được sử dụng trong một thời gian dài. Tuy nhiên, điều bất lợi là việc nhận dạng các hành vi đã được định nghĩa trước trong giai đoạn phân loại, muốn bổ sung một hành vi mới thì phải thực hiện việc phân loại lại. Các phương pháp sử dụng như mô hình láng giềng gần, kỹ thuật Boosting, Support Vector Machine (SVM) và Neural Networks... Trong hệ thống này, chúng tôi sử dụng phương pháp máy học véc-tơ hỗ trợ (SVM).

Tuy nhiên, để áp dụng được máy học véc-tơ hỗ trợ trong việc nhận dạng hành vi là quan tâm hay không quan tâm đến gian hàng mà đối tượng đi qua trong siêu thị, chúng tôi cần xây dựng được tập dữ liệu chứa thông tin đặc trưng với số thuộc tính cố định.

Mục tiêu là xây dựng được tập dữ liệu các thuộc tính thể hiện đặc trưng của quá trình theo dõi, chúng ta cần rút trích các thông tin quan trọng từ kết quả theo dõi đối tượng. Các kết quả đó là quỹ đạo đường đi của đối tượng và thời gian (hoặc số khung ảnh) mà đối tượng lưu lại trong vùng quan sát. Vì thế, chúng tôi đề xuất một phương pháp trích và biểu diễn đặc trưng dưới dạng véc-tơ có cùng số chiều.

Quỹ đạo chuyển động của đối tượng được mô tả từ một tập hợp điểm $P = (p_1, p_2 \dots p_n)$. Ứng với mỗi quỹ đạo chuyển động khác nhau của đối tượng thì số n cũng khác nhau. Vì vậy, để đặc trưng quỹ đạo không phụ thuộc vào n, thì quỹ đạo được chia làm k đoạn, mỗi đoạn lấy tọa độ điểm làm đặc trưng. Trong trường hợp $n < k$, cần phải bổ sung vào các điểm có tọa độ (0, 0) sao cho $n = k$.

Ta có tỷ lệ rút thông tin là $\frac{n}{k}$, tọa độ các điểm $P = \{(x_1, y_1), (x_2, y_2) \dots (x_k, y_k)\}$ được rút trích như sau:

$$P_x = \begin{cases} x_{i*\frac{n}{k}} \text{ nếu } i * \frac{n}{k} \leq n \\ x_{n-1} \text{ nếu } i * \frac{n}{k} > n \end{cases}$$

$$P_y = \begin{cases} y_{i*\frac{n}{k}} \text{ nếu } i * \frac{n}{k} \leq n \\ y_{n-1} \text{ nếu } i * \frac{n}{k} > n \end{cases}$$

Trong đó, i đi từ 0 đến k-1.

Giả sử chọn k=16 và thông tin về số khung hình mà đối tượng lưu lại trong vùng quan sát thì ta sẽ có: 16*2+1=33 chiều.

3.4 Cải thiện tốc độ xử lý của hệ thống

3.4.1 Thay đổi bộ mô tả đặc trưng

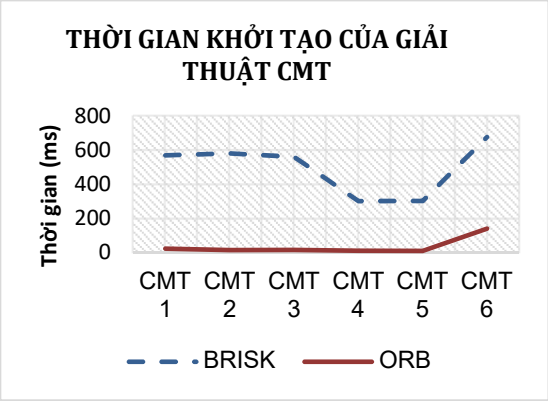
Như đã mô tả ở mục 3.2, hệ thống được cài đặt sử dụng bộ mô tả ORB trong khi thư viện libCMT sử dụng BRISK. Sau khi cài đặt thực nghiệm, tốc độ của hệ thống khi sử dụng bộ mô tả ORB tăng lên đáng kể, xem Biểu đồ 1 và Biểu đồ 2.

Nếu không sử dụng bộ mô tả ORB mà sử dụng bộ mô tả BRISK, như tác giả giải thuật sử dụng, ngưỡng so khớp thr_dist với giá trị mặc định là 0,25 có nghĩa là bộ mô tả của 2 keypoint cần so khớp phải giống nhau trên 75% mới được xem là giống nhau. Tuy nhiên, nếu áp dụng phân ngưỡng này ở môi trường là siêu thị có nền phức tạp và các keypoint tương tự nhau rất nhiều, sẽ dẫn đến so khớp sai, cần phải tăng tỷ lệ so khớp này lên từ 85% đến 95%, có nghĩa là thr_dist là 0,15 hoặc 0,05.

3.4.2 Giới hạn phạm vi đối tượng trước khi rút trích các đặc trưng

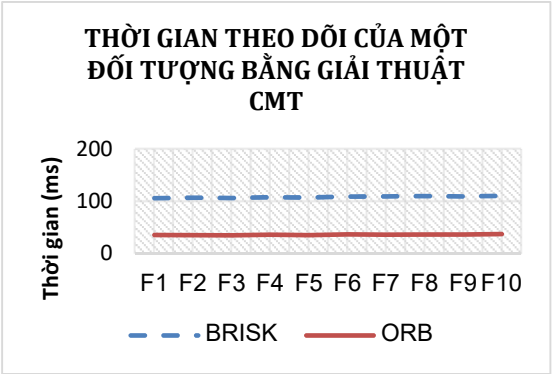
Thông thường sau khi phát hiện đối tượng sẽ dùng hình chữ nhật để bao quanh và tiến hành rút trích các đặc trưng trong hình chữ nhật đó. Tuy nhiên, đối với trường hợp đối tượng là dáng người đứng thì sẽ có khoảng trống không thuộc đối tượng người mà thuộc nền dẫn đến việc rút các đặc trưng nhầm lẫn từ nền sang người. Trong trường hợp nền có cấu trúc phức tạp, số keypoint nền bị nhầm lẫn sang người nhiều hơn số lượng keypoint thực sự thuộc người thì theo dõi bằng giải thuật CMT không còn chính xác nữa.

Để hạn chế vấn đề này, chúng tôi đã đề xuất sử dụng hình đa giác bám theo dáng người đang đứng hoặc đi khác với hình chữ nhật (Hình 7).

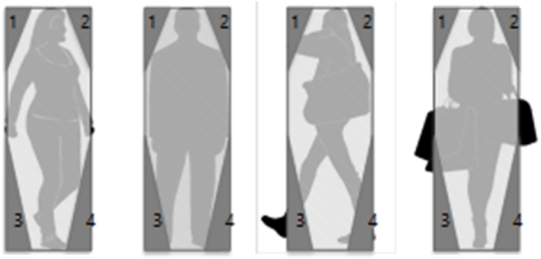


Biểu đồ 1: So sánh về thời gian khởi tạo của giải thuật CMT giữa hai bộ mô tả BRISK và ORB dựa trên số lượng các CMT được tạo ra

Ngoài ra, số lượng keypoint thuộc đối tượng giảm làm cho giải thuật theo dõi chạy nhanh hơn ở một số bước như theo dõi (tracking) ở bước 2, so khớp cục bộ (local matching) ở bước 7 (Hình 6).



Biểu đồ 2: So sánh thời gian theo dõi một đối tượng của giải thuật CMT từ khung hình thứ nhất đến khung hình thứ 10



Hình 7: Mặt nạ đa giác biểu diễn dáng người

3.4.3 Xử lý bằng kỹ thuật đa luồng

Các bước xử lý của giải thuật CMT thích nghi như mô tả trong Hình 6 đều tuần tự và đa số phụ thuộc nhau (ngoại trừ bước 2, 3 và 6, 7 có thể thực hiện song song). Chúng tôi đề xuất một giải thuật mới cải tiến tốc độ xử lý việc theo dõi bằng cách sử dụng kỹ thuật đa luồng (multithread) trên các hệ thống đa nhân (multicores).

Chúng tôi chia các bước xử lý trong Hình 6 thành 6 giai đoạn (stages): (1), (2), (3), (4, 5), (6) và (7, 8). Mỗi giai đoạn được thực thi trong một luồng (thread) khác nhau. Giải thuật hoạt động như sau: đầu tiên, thread 1 tính toán các đặc trưng cục bộ và lưu kết quả vào biến toàn cục (***A***). Thread 2 thực hiện truy vết các đặc trưng được tính ở Thread1 bằng phương pháp luồng quang học và lưu kết quả vào trong biến (***B***). Thread 3 chờ cho đến khi thread 1 hoàn thành là nó có thể bắt đầu thực hiện việc so khớp toàn cục. Kết quả của việc so khớp toàn cục sẽ được đặt vào biến lưu trữ tương ứng (***C***). Thread 4 chờ cho đến khi thread 2 và thread 3 thực hiện xong nó sẽ hợp các điểm đặc trưng lại (được lưu trữ trong biến ***B*** và ***C***) và tính toán ước lượng tỷ lệ, góc quay, tâm đối tượng đưa vào biến lưu trữ (***D***). Thread 5 chờ cho thread 4 thực hiện xong nó sẽ tìm cluster lớn nhất theo phương pháp consensus từ thông tin lưu trữ trong ***D***. Thread 6 có thể bắt đầu cùng lúc với Thread 5 sau khi thread 4 hoàn thành, thread này sẽ thực hiện so khớp cục bộ và hợp nhất các đặc trưng để cho ra kết quả sau cùng (Hình 8).

4 KẾT QUẢ THỰC NGHIỆM

Chúng tôi đã thực hiện cài đặt hệ thống bằng ngôn ngữ C++ sử dụng thư viện OpenCV và thư viện libCMT trên nền Qt. Chúng tôi đánh giá việc cải tiến tốc độ xử lý theo phương pháp đa luồng trên tập dữ liệu thu thập từ camera an ninh ghi lại diễn biến thực tế tại siêu thị.

4.1 Kết quả xây dựng mô hình nhận dạng người và hành vi

4.1.1 Xây dựng mô hình nhận dạng người sử dụng phương pháp cascade classifier

Tạo tập dữ liệu để xây mô hình cho phương pháp cascade classifier có giai đoạn quan trọng là xây dựng tập ảnh chứa đối tượng (tập positive) và tập chứa ảnh nền (tập negative). Ở bước xây dựng tập ảnh chứa đối tượng, tôi sử dụng công cụ ImageClipper của tác giả Naotoshi Seo (, 2014). Một tập dữ liệu với 18.357 file là ảnh người và tập dữ liệu với 44.648 ảnh nền đã được xây dựng. Sau khi xây dựng tập dữ liệu tiến hành xây dựng mô hình. Kết quả đánh giá mô hình trên tập dữ liệu (rút trích từ video thực tế) được thể hiện ở Bảng 1.

Với tập ảnh 1 và tập ảnh 2 được rút ra ngẫu nhiên từ đoạn 2 đoạn video dài 10 phút ghi lại hình ảnh khách hàng tại siêu thị, và dữ liệu này không nằm trong tập dữ liệu huấn luyện.

Bảng 1: Kết quả đánh giá mô hình

Tập ảnh	Tổng số ảnh	Tổng số đối tượng	% phát hiện đúng	% phát hiện sai	% không phát hiện được
Tập 1	100	657	83,08%	18,46%	16,92%
Tập 2	200	1295	84,55%	6,50%	15,45%

4.1.2 Xây dựng mô hình nhận dạng hành vi dùng máy học SVM

Dữ liệu huấn luyện sử dụng giải thuật SVM thì cần phải có định dạng là véc-tơ đặc trưng và nhãn ở dạng số. Vì vậy, cần phải xây dựng tập dữ liệu mẫu có chứa thông tin phù hợp. Tập tin đã xây dựng chứa 369 mẫu và chứa 33 thuộc tính. Trong đó, 16 cột đầu tiên là tọa độ điểm x, 16 cột sau là tọa độ điểm y, cột cuối cùng là số khung hình mà đối tượng xuất hiện trong vùng quan sát. Trong đó, 16 cột đầu tiên là tọa độ điểm x, 16 cột sau là tọa độ điểm y, cột cuối cùng là số khung hình mà đối tượng xuất hiện trong vùng quan sát.

Dữ liệu được xây dựng phân thành hai lớp, lớp thứ nhất là những khách hàng có quan tâm đến gian hàng đang trưng bày, lớp thứ hai gồm những khách hàng không quan tâm. Sử dụng thư viện libSVM xây dựng mô hình với các thông số như trong Bảng 2, sẽ đạt độ chính xác 92,1196%.

Bảng 2: Thông số xây dựng mô hình nhận dạng

Hàm nhân	c	gama	degree	coef
POLY	0,00002	0,7	1	0,25

4.2 Kết quả cải thiện tốc độ hệ thống

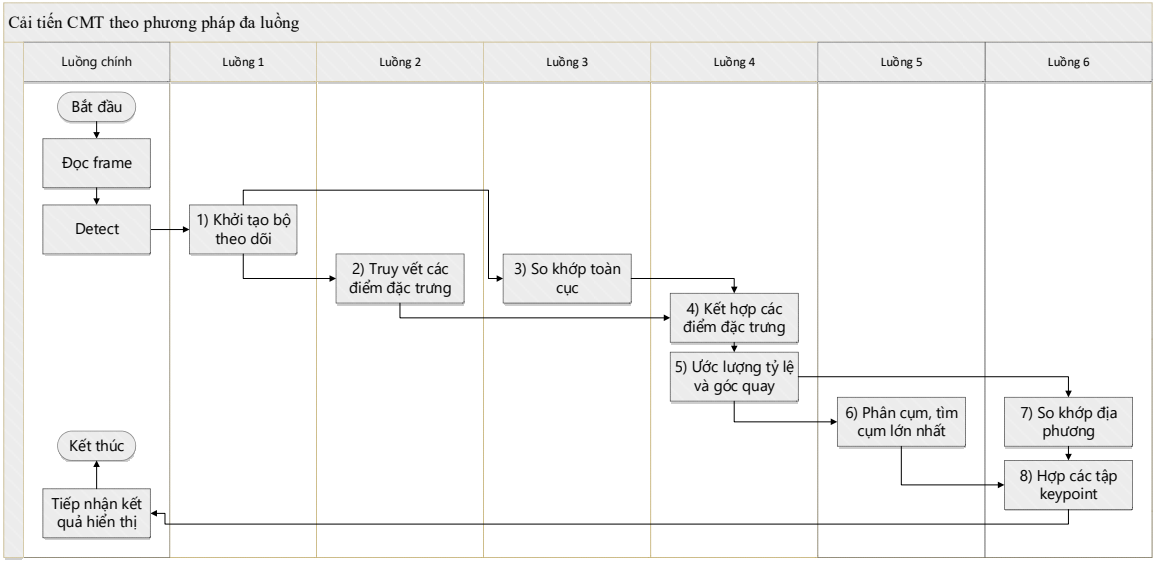
Kết quả cho thấy giải thuật cải tiến cho hiệu quả xử lý tăng lên trung bình gấp 7 lần ở giai đoạn khởi tạo bộ theo dõi (khi có đối tượng mới xuất hiện) và tăng lên trung bình 2,8 lần ở giai đoạn theo dõi khi chạy trên máy tính 2 cores (4 threads/core).

Biểu đồ 3 cho thấy rằng khi kết hợp việc thay đổi bộ mô tả từ BRISK sang ORB và phương pháp xử lý từ tuần tự sang đa luồng có thể cải thiện đáng kể thời gian thực thi. Tốc độ hệ thống tăng rõ rệt khi kết hợp cùng lúc 2 phương pháp cải tiến là sử dụng bộ mô tả ORB và xử lý đa luồng trong khi độ chính xác trước và sau khi cải tiến giải thuật là như nhau.

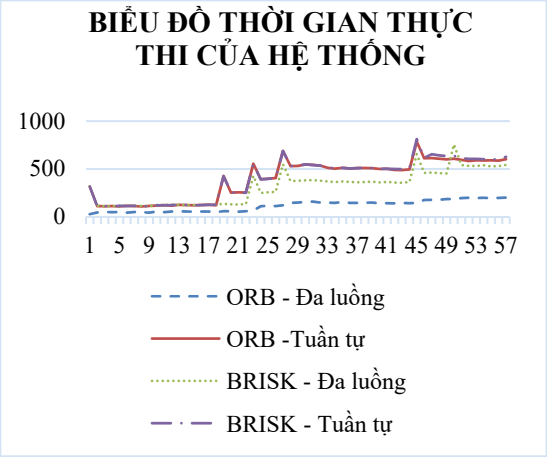
5 KẾT LUẬN

Qua quá trình nghiên cứu, chúng tôi đã xây dựng thành công mô hình nhận dạng người, theo dõi và cuối cùng là nhận dạng hành vi của họ trong siêu thị. Ngoài ra, chúng tôi còn đề xuất một phương pháp cải tiến giải thuật CMT theo hướng xử lý đa luồng nhằm cải thiện tốc độ. Kết quả là tốc độ tăng lên 2,8 lần so với giải thuật gốc. Từ đó đáp ứng tốt vào bài toán thực tế theo dõi nhiều đối tượng xuất hiện cùng lúc, Tuy nhiên, do môi trường đặt camera quan sát là siêu thị đông người, cấu trúc ảnh nền phức tạp, có quá nhiều các chi tiết nhỏ có đặc trưng giống nhau. Chính vì những đặc trưng giống nhau này gây ra việc nhầm lẫn khi nhận dạng cũng như theo dõi đối tượng, trực tiếp ảnh hưởng đến độ chính xác của cả hệ thống.

Từ mô hình này, chúng ta có thể mở rộng áp dụng cho các hệ thống thông minh hỗ trợ công tác an ninh như dựa vào nhận dạng quỹ đạo hoặc kết hợp thêm các phương pháp nhận dạng cử chỉ, hành động người phát hiện các hành vi bất thường...



Hình 8: Cải tiến giải thuật CMT theo phương pháp đa luồng



Biểu đồ 3: so sánh thời gian thực thi khi kết hợp các phương pháp cải tiến giải thuật

TÀI LIỆU THAM KHẢO

Lavee, G., Rivlin, E., & Rudzsky, M. (2009). Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 39(5), 489–504.

Leutenegger, S., Chli, M., & Siegwart, R. Y. (2011). BRISK: Binary robust invariant scalable keypoints. In 2011 International conference on computer vision (pp. 2548–2555). IEEE.

Lucas, B. D., Kanade, T., & others. (1981). An iterative image registration technique with an application to stereo vision. In IJCAI (Vol. 81, pp. 674–679).

Naotoshi Seo. (2014). ImageClipper. C++.

Nebehay, G., & Pflugfelder, R. (2014). Consensus-based matching and tracking of keypoints for object tracking. In IEEE Winter Conference on Applications of Computer Vision (pp. 862–869). IEEE.

Nebehay, G., & Pflugfelder, R. (2015). Clustering of static-adaptive correspondences for deformable object tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2784–2791).

Ragland, K., & Tharcis, P. (2014). A survey on object detection, classification and tracking methods. Int. J. Eng. Res. Technol, 3(11).

Rosten, E., & Drummond, T. (2006). Machine learning for high-speed corner detection. Computer Vision–ECCV 2006, 430–443.

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. In Computer Vision (ICCV), 2011 IEEE international conference on (pp. 2564–2571). IEEE.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on (Vol. 1, pp. I–I). IEEE.