

# Ứng dụng thị giác máy tính trong khâu thanh toán của cửa hàng không người bán

Minh Hoang Vu 

Hanoi University of Science, VNU

Phong Nguyen Nam 

Hanoi University of Science, VNU

Van Anh Nguyen Thi 

Hanoi University of Science, VNU

## Abstract

Trong bài báo này, chúng tôi đề xuất một giải pháp thuần sử dụng thị giác máy tính nhằm giải quyết khâu thanh toán trong bài toán thực tế: "Cửa hàng không người bán". Chúng tôi tập trung giải quyết hai khía cạnh chính: thanh toán bằng khuôn mặt và tự động tính giá thành sản phẩm. Với thanh toán bằng khuôn mặt, chúng tôi kết hợp việc định danh khuôn mặt (Face Recognition) đồng thời chống giả mạo bằng hình ảnh hoặc video (Liveness Detection). Phần tự động tính giá thành được triển khai bằng phát hiện và truy vết đối tượng (Object Detection & Object Tracking).

Giải pháp của chúng tôi không chỉ giảm chi phí nhân lực mà còn giúp nâng cao hiệu suất, đồng thời mở ra những triển vọng mới về ứng dụng của thị giác máy tính và trí tuệ nhân tạo trong việc giải quyết các bài toán tự động hóa. Chúng tôi tin rằng giải pháp này không chỉ là đề xuất về mặt lý thuyết mà còn mang lại giá trị ứng dụng cao, đưa ra một hướng đi triển vọng trong ngành công nghiệp bán hàng tự động. Tính khả thi và cơ sở kỹ thuật sẽ được chúng tôi đề cập trong các chương sau của bài báo.

## Keywords

computer vision, cashier-less store, face recognition, face anti-spoofing, object detection, object tracking

## 1. Introduction

Trong thời đại hiện nay, đi cùng với sự phát triển của khoa học, công nghệ, đời sống và mức sống của con người ngày càng được nâng cao, quy trình mua sắm và thanh toán cũng đang ngày càng trở nên tự động và không còn phụ thuộc vào sự can thiệp của con người. Ở các nước phát triển như Anh, Mỹ, Nhật Bản, Trung Quốc, không khó để bắt gặp robot giúp việc, robot phục vụ, đi cùng với đó là vô vàn các thử nghiệm liên quan đến chuỗi cửa hàng hoàn toàn tự động trên toàn cầu. Những ông lớn trong ngành bán lẻ như Amazon hay Alibaba cũng không đứng ngoài cuộc, họ cũng đã triển khai những cửa hàng không người bán của riêng mình (Amazon Go), nhìn chung những

thử nghiệm này đều mang lại hiệu quả tích cực. Một loạt các start-up về cửa hàng không người bán trên thế giới cũng nổi lên và dần khẳng định được vị thế, nổi bật trong số đó là BingoBox - một trong những đối thủ lớn nhất của Alibaba trong mảng này tại Trung Quốc. Tuy nhiên, một trong những thách thức lớn nhất đối diện với các doanh nghiệp là làm thế nào để thực hiện thanh toán một cách hiệu quả và an toàn trong môi trường không có sự giám sát từ người bán hàng, bên cạnh đó, vấn đề chi phí cũng là một rào cản lớn trong việc phát triển cửa hàng không người bán.

Mặc dù bài toán cửa hàng không người bán không còn mới, tuy nhiên cho đến nay, trên thế giới vẫn chưa có cửa hàng nào triển khai việc tự động hóa hoàn toàn bằng công nghệ thị giác máy tính. Ở hầu khắp các cửa hàng không người bán đều có tích hợp cảm biến nhiệt hạch hoặc thanh toán bằng mã vạch cho khâu thanh toán, vô hình chung làm phát sinh thêm nhiều chi phí cho máy móc và thiết bị cảm biến. Chính vì vậy, trong bài báo này, chúng tôi đề xuất một giải pháp nhằm giải quyết khâu thanh toán này một cách nhanh chóng và tiện lợi hoàn toàn dựa trên công nghệ trí tuệ nhân tạo và thị giác máy tính, điều này giúp giảm thiểu đáng kể chi phí cho việc vận hành các máy móc, thiết bị, đồng thời mang đến một cái nhìn độc đáo, mới lạ trong việc ứng dụng công nghệ này nhằm giải quyết các bài toán trên thực tế. Chúng tôi tập trung vào việc giải quyết hai khía cạnh: thanh toán bằng khuôn mặt và tính giá thành sản phẩm tự động. Trước đây, việc thanh toán bằng khuôn mặt thường gặp khó khăn trong việc thu thập dữ liệu khuôn mặt của khách hàng, tuy nhiên với sự phát

triển của các mô hình đào tạo trước, hiện nay, việc này hoàn toàn có thể thực hiện bằng phương pháp học tương tự (Learning Similarity) mà không cần thông qua việc đào tạo trên dữ liệu khuôn mặt của khách hàng. Sự phát triển của các mô hình phát hiện đối tượng trong thời gian thực như YOLO cũng giúp cải thiện đáng kể tốc độ trong việc phát hiện và truy xuất đối tượng, qua đó giúp đẩy nhanh tốc độ và độ chính xác trong việc tính giá thành tự động một cách đáng kể. Chi tiết về mặt kỹ thuật sẽ được chúng tôi trình bày trong mục 3. Thiết kế kỹ thuật. Các phần còn lại trong bài báo được tổ chức như sau: 2. Các nghiên cứu liên quan, 4. Thảo luận, kết quả thực nghiệm, 5. Kết luận.

## 2. Approach

---

**QR Code and RFID.** Đây là phương pháp mà hiện tại được BingoBox sử dụng cho khâu thanh toán. Mỗi vật phẩm sẽ được gắn thẻ RFID, như vậy khi đi qua cổng, máy quét sẽ xác định được số vật phẩm khách hàng đã mua, sau đó tiến hành thanh toán bằng hình thức quét QR Code. Cách làm này tồn tại nhiều hạn chế, chi phí đắt đỏ cho các thẻ RFID.

**Sử dụng thị giác và cân nặng.** Amazon Go kết hợp cả camera và cảm biến cân nặng trên kệ hàng, nơi cảm biến cân nặng có thể được sử dụng để xác định khi nào một mặt hàng được lấy ra khỏi kệ, ngay cả khi nó bị che khuất khỏi camera. Một thách thức mà hệ thống như vậy phải đối mặt là khả năng phân biệt giữa các mặt hàng có cùng trọng lượng. Điều này dẫn tới việc khách hàng có thể gian lận bằng việc đặt những viên đá.

Hơn nữa, thiết kế của họ đòi hỏi một sự triển khai lại cấu trúc của cửa hàng: cổng kiểm tra người dùng, hệ thống camera dày đặc trên trần.

### 3. Proposal Design

Dựa trên những cách tiếp cận đề tài đã có, đồng thời sửa đổi để phù hợp với khả năng và điều kiện, chúng tôi đề xuất một phương pháp thanh toán trong cửa hàng không người bán chỉ sử dụng camera. Phương pháp này bao gồm 2 thành phần:

*Object Detection/Tracking* nhằm tạo hóa đơn tự động. Đối với số lượng hàng mua ít, khách hàng có thể chụp 1 bức hình các món đồ muốn mua, và kết quả trả về sẽ là danh sách những món đồ đó, cùng với tổng tiền cần trả. Trong trường hợp khách hàng muốn mua hàng số lượng lớn, có thể đưa các món hàng lên băng chuyền. Trong đề tài này, do chưa có điều kiện, nên chúng tôi sẽ mô phỏng việc sử dụng băng chuyền bằng cách quay video các món đồ theo chiều từ trái qua phải.

*Face Recognition* giúp khách hàng có thể thanh toán bằng khuôn mặt. Khách hàng sẽ đưa mặt vào vùng camera chỉ định trước, và nếu là khách hàng cũ/đã đăng kí thì kết quả trả về sẽ là tên khách hàng. Ngược lại, nếu là một khách hàng mới, hệ thống sẽ yêu cầu khách hàng đăng kí. Ngoài ra, kết quả bán hàng và tên khách sẽ được lưu tự động vào file csv nhằm tối ưu việc quản lý và bán hàng.

*Live ness Detection* nhằm phát hiện trong việc mạo danh khuôn mặt bằng ảnh hoặc video tại quầy thanh toán bằng khuôn mặt.

## 4. Methodology

### 4.1. Object Detection/Tracking

Object Detection/Tracking có nhiệm vụ tạo hóa đơn tự động thông qua ảnh hoặc video. Đối với ảnh, ta sẽ cần xác định chính xác các món đồ có trong bức ảnh đó. Đối với video, ta sẽ cần thêm bước gán ID cho từng món đồ đó. Với mục đích đề tài nhằm áp dụng cho các cửa hàng, yếu tố tốc độ chạy sẽ được đặt lên hàng đầu. Do đó, chúng tôi lựa chọn sử dụng mô hình YOLOv7 [1] trong việc phát hiện và nhận diện đồ vật, cùng với đó là thuật toán SORT [2] cho việc theo dõi và gán ID cho đồ vật.

#### 4.1.1. Dataset

Trong đề tài này, để mô phỏng, chúng tôi sử dụng 9 loại mặt hàng: bim bim Poca, mì Hảo Hảo, mì Omachi, sữa Milo, sữa Fami, hộp bánh Custas (2 cái), cà phê lon Boss, bình nước và bút bi FlexOffice. Các mặt hàng được lựa chọn để kiểm định khả năng mô hình, bao gồm các món hàng có màu sắc giống nhau (bim bim - mì - sữa Milo), có hình dạng và màu sắc giống nhau (mì Hảo Hảo - mì Omachi; sữa Fami - hộp bánh Custas) và có kích thước nhỏ (bút bi).



Figure 4.1: Các món đồ được sử dụng

Chúng tôi đã thực hiện việc thu thập dữ liệu từ ảnh chụp đối tượng ở nhiều góc độ khác nhau. Quá trình này bao gồm việc cắt các khung hình từ nhiều cảnh quay của đối tượng.

Tiếp theo, chúng tôi đánh nhãn dữ liệu, nhằm tạo ra một tập tin chứa thông tin về tọa độ tâm, chiều dài và chiều rộng của hộp bao quanh đối tượng trong ảnh, mục đích là để phục vụ cho bài toán phát hiện đối tượng. Phương pháp gán nhãn dữ liệu của chúng tôi bao gồm cả hai cách tiếp cận: tự động và thủ công. Trong phương pháp tự động, chúng tôi sử dụng RetinaNet50 [3] để phát hiện các đối tượng và xác định tọa độ của hộp bao quanh chúng. Còn đối với phương pháp thủ công, chúng tôi thực hiện quá trình gán nhãn bằng cách vẽ các hộp xung quanh đối tượng và sau đó xuất ra định dạng YOLO. Để thực hiện phương pháp thủ công này, chúng tôi sử dụng công cụ từ trang web <https://www.makesense.ai/>.

Chúng tôi cũng giới thiệu một loạt các kỹ thuật tăng cường dữ liệu được tích hợp vào quá trình huấn luyện, nhằm nâng cao hiệu suất của mô hình hình ảnh. Các kỹ thuật này bao gồm điều chỉnh thuộc tính màu sắc theo không gian màu HSV, bao gồm tăng cường Hue, Saturation, và Value. Ngoài ra, để đảm bảo sự đa dạng trong dữ liệu đào tạo, chúng tôi áp dụng kỹ thuật dịch chuyển hình ảnh theo chiều ngang và chiều dọc, cũng như điều chỉnh tỷ lệ kích thước của hình ảnh. Đồng thời, để tăng cường khả năng nhận diện, chúng tôi sử dụng kỹ thuật lật hình ảnh theo chiều ngang với xác suất xác định. Một phương pháp mới được giới thiệu trong

nghiên cứu này là kỹ thuật Mosaic, trong đó nhiều hình ảnh được kết hợp lại với nhau với xác suất nhất định, đóng góp vào sự đa dạng của dữ liệu. Chúng tôi cũng đề cập đến kỹ thuật Mixup, trong đó các hình ảnh khác nhau được kết hợp lại để tạo ra các ví dụ mới. Kỹ thuật Copy Paste là một phần quan trọng khác của phương pháp tăng cường dữ liệu của chúng tôi, nơi một phần của hình ảnh được sao chép và dán vào hình ảnh gốc. Mỗi kỹ thuật được kiểm soát thông qua giá trị tham số để điều chỉnh mức độ tăng cường dữ liệu. Kết quả của nghiên cứu này chứng minh rằng sự kết hợp linh hoạt của các kỹ thuật tăng cường dữ liệu này có thể cải thiện đáng kể hiệu suất của mô hình hình ảnh trong các ứng dụng nhận dạng và phân loại.

Kết quả của quá trình chuẩn bị dữ liệu, là bộ 2.301 ảnh bao gồm 9 nhãn.

#### 4.1.2. Xử lý kết quả trên video

Khi xử lý video, ngoài việc nhận diện từng đồ vật bằng YOLOv7 ở mọi khung hình, ta còn cần theo dõi và gán ID bằng SORT cho đồ vật để tránh việc trùng lặp đồ ở các khung hình. Mỗi ID sẽ đi kèm theo duy nhất 1 nhãn, và nhờ đó ta sẽ có một danh sách các món đồ chính xác.

Tuy nhiên, khi thực nghiệm, có một số vấn đề xảy ra. Đầu tiên là việc mô hình nhận diện sai đồ vật do chỉ có một phần đồ vật trong khung hình, dù vẫn nhận diện đúng khi đồ vật xuất hiện đầy đủ. Các nhãn bị nhận sai cũng sẽ bị thêm vào danh sách mua hàng. Do vậy, để hạn chế tối đa trường hợp nhận sai nhãn, ta sẽ chỉ thêm vào danh sách những món đồ có tâm của bounding box nằm trong

một khoảng cụ thể của khung hình.



**Figure 4.2:** Đường phân cách

Các món đồ có tâm hộp bao nằm giữa 2 đường màu xanh lá sẽ được thêm vào danh sách. Một vấn đề khác xảy ra về ID, khi có hiện tượng ID bị nhảy nhãn khi tâm bounding box nằm trong khoảng. Điều này khiến một ID có thể có nhiều nhãn. Cách xử lý hiện tại là với mỗi ID, ta sẽ nhận nhãn tương ứng cuối cùng.

## 4.2. Face Recognition

Face Recognition sẽ giúp khách hàng có thể thanh toán bằng khuôn mặt. Quá trình này sẽ bao gồm 3 bước: Face Detection, Liveness Detection và Face Embedding.

*Face Detection* nhằm phát hiện khuôn mặt khách hàng. Sau khi thử nghiệm một số mô hình, chúng tôi lựa chọn OpenCV cho phần này nhằm đảm bảo sự cân bằng về tốc độ chạy và sự chính xác. Một bước quan trọng đó là *Liveness Detection* để phát hiện các trường hợp giả mạo, ví dụ như sử dụng ảnh giấy hoặc ảnh trên điện thoại có chứa khuôn mặt người khác để gian lận. Chúng tôi sẽ sử dụng MiniFASNet của công ty Minivision [4] (Trung Quốc). Mạng này sử dụng quang

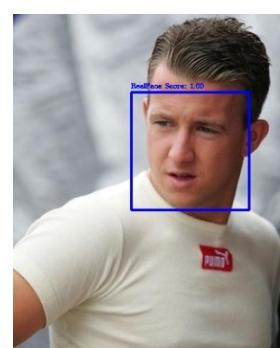
phổ FFT để phát hiện khuôn mặt giả mạo, với việc được thiết kế chủ yếu để hoạt động trong các ứng dụng mobile, mạng sẽ đạt tốc độ tối đa nếu tỉ lệ chiều rộng / chiều dài của khung ảnh là 3/4. Vì vậy, chúng tôi thực hiện chuyển hình ảnh về kích thước đó trước khi đưa vào mạng, theo ước tính tốc độ có thể nhanh lên tới 60%. Nếu khuôn mặt trong camera là giả mạo, chương trình sẽ thoát ra. Ngược lại thì ta sẽ đến với bước tiếp theo.

*Face Embedding*, sử dụng VGGFace [5] để vector hóa khuôn mặt khách hàng và đổi chiều với các khuôn mặt có sẵn trong cơ sở dữ liệu. Nếu khuôn mặt hiện tại không giống các khuôn mặt đã được lưu, ta có thể xác định đây là khách hàng mới và yêu cầu đăng ký. Kết quả sẽ được ghi lại vào một file csv nhằm theo dõi quản lý và tối ưu hóa việc bán hàng.

## 5. Result

### 5.1. Face Recognition

Các trường hợp thật được minh họa bằng đường bao màu xanh và ngược lại, các trường hợp giả mạo được minh họa bằng đường bao màu đỏ.



**Figure 5.3:**  
Trường hợp người thật



**Figure 5.4:**  
Trường hợp giả mạo

Mô hình có thể phân biệt và phát hiện các trường hợp người thật và giả mạo với độ chính xác cao trong thời gian rất ngắn.



**Figure 5.5:** Kết quả nhận diện

Đồng thời, kết quả nhận diện cũng là chính xác, với FPS vẫn giữ ổn định 30 FPS khi sử dụng webcam.

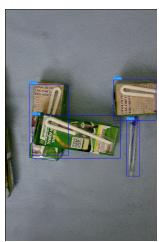
## 5.2. Object Detection/Tracking



**Figure 5.6:**  
Kết quả trên  
ảnh



**Figure 5.7:** Video lộn  
**Figure 5.8:** Video  
xộn



**Figure 5.8:** Video  
chồng đồ

Mô hình xử lí ảnh khá tốt, khi dù có nhiều đồ xếp lộn xộn và có cả đồ vật nhỏ như bút vẫn có thể nhận diện chính xác và đầy đủ.

Tương tự đối với video, khi các món đồ xếp lộn xộn và có màu sắc, hình dáng gần giống nhau vẫn được nhận diện đầy đủ.

Cuối cùng, khi có một số món đồ chồng lên nhau một phần, mô hình vẫn có thể phát hiện được. Tuy nhiên, trong các trường hợp

các món đồ chồng trùng tâm thì kết quả sẽ không tốt.

## 6. Related works

Trên thực tế, các cửa hàng không người bán trên toàn cầu không công khai toàn bộ những gì họ thực hiện, vậy nên chúng tôi viết dựa trên hiểu biết về những thông tin đã được công bố.

**Hệ thống mua sắm không thu ngân trong thương mại.** Amazon Go là chuỗi cửa hàng đầu tiên thực hiện triển khai hệ thống bán lẻ không thu ngân, sự tiên phong đó đã kéo theo rất nhiều các cửa hàng không thu ngân khác trên thế giới, nổi bật hơn cả là chuỗi cửa hàng thử nghiệm của Alibaba và BingoBox - đối thủ chính của Alibaba trong lĩnh vực này tại Trung Quốc. Alibaba và BingoBox sử dụng đầu đọc RFID [6] để quét các mặt hàng trong siêu thị nhằm phục vụ cho khâu thanh toán, trong khi Amazon Go thực hiện bằng sự kết hợp giữa hệ thống camera khổng lồ và các thiết bị cảm biến. Imager sử dụng một camera đặt trong giỏ hàng nhằm phát hiện sản phẩm mà khách hàng đã mua, điều này có phần tương đồng với giải pháp được chúng tôi trình bày trong bài báo.

**Weight scales.** Một nhóm các nhà nghiên cứu từ Trung Quốc cũng đề ra giải pháp sử dụng cân để nhận biết một vật biến mất khỏi giá để hàng, điều này cũng giúp khâu thanh toán diễn ra một cách nhanh hơn. Tuy nhiên phương án này cũng tiềm ẩn nhiều rủi ro, khi ta có thể gian lận về cân nặng. Họ cũng trình bày một vài giải pháp nhằm tránh các trường hợp gian lận, chi tiết hãy tham khảo bài báo

"Grab: Fast and Accurate Sensor Processing for Cashier-Free Shopping", Xiaochen Liu, Yurong Jiang, Kyu-Han Kim, Ramesh Govindan [7].

**Items detection and tracking.** Trước đây bài toán này đã được thực hiện bằng Google Glass [8] nhưng các thiết bị đó hiện vẫn chưa được triển khai rộng rãi. Trong bài báo hiện tại, chúng tôi đề xuất đến việc phát hiện và theo dõi các mặt hàng dựa trên thị giác máy tính.

**Face recognition.** Nhận diện khuôn mặt trong hệ thống an ninh, nhận diện khuôn mặt trên các thiết bị di động. Đặc biệt, nhận diện khuôn mặt dựa trên Deep Learning đã đem lại những đột phá đáng kể trong lĩnh vực nhận diện khuôn mặt. Các mô hình như FaceNet [9] và VGG-Face sử dụng kiến trúc mạng neural sâu để học các đặc trưng phức tạp từ dữ liệu khuôn mặt. Các phương pháp này không chỉ cải thiện độ chính xác mà còn giảm thiểu ảnh hưởng của biến đổi hình thái và đa dạng trong dữ liệu. Chúng tôi tận dụng sức mạnh từ những mô hình học sâu đào tạo trước để thực hiện việc nhận diện khuôn mặt bằng phương pháp học tương tự.

**Live ness Detection.** Bài toán này đã được nhiều bài báo, nghiên cứu thực hiện bằng cách sử dụng các mô hình phát hiện và phân loại trên thời gian thực như YOLOv8, YOLOv7... Chúng tôi sử dụng giải pháp của Minivision AI .

## Conclusion

Triển khai cửa hàng không người bán bằng giải pháp của chúng tôi không chỉ làm tăng

trải nghiệm mua sắm của khách hàng mà còn giúp giảm đáng kể chi phí về thiết bị cảm biến. Chúng tôi đã đạt được độ chính xác cao và tốc độ xử lý ổn định trong cả hai bài toán thanh toán bằng khuôn mặt và tự động tính giá thành, các trường hợp giả mạo khuôn mặt cũng đã được nhận biết.

Tuy nhiên, giải pháp hiện tại vẫn còn những hạn chế. Chúng tôi chưa thể xử lý một số tình huống như các đồ vật bị chồng trùng tâm lên nhau trong quá trình thanh toán, và số nhãn hiện tại vẫn còn khá nhỏ so với thực tế. Trong tương lai, chúng tôi cam kết tiếp tục nghiên cứu và phát triển để đưa ra các giải pháp mới nhằm giải quyết những thách thức này. Chúng tôi đặt mục tiêu mở rộng ứng dụng của thị giác máy tính và trí tuệ nhân tạo ra tất cả các khâu của cửa hàng không người bán, tạo nên một mô hình cửa hàng hoàn toàn dựa trên công nghệ thị giác. Điều này không chỉ là một bước tiến vững chắc trong việc tự động hóa mua sắm mà còn mở ra những triển vọng mới và sáng tạo trong ngành công nghiệp này.

## References

- 
- [1] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022.
  - [2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft, SIMPLE ONLINE AND REALTIME TRACKING, 2017.
  - [3] Tsung-Yi Lin, Priya Goyal, Ross Gir-

shick, Kaiming He, and Piotr Dollar, Focal Loss for Dense Object Detection, 2018.

[4] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, "VGGFace2: A dataset for recognising faces across pose and age"

[5] Silent Face Anti Spoofing - Minivison AI, [github.com/minivision-ai/Silent-Face-Anti-Spoofing](https://github.com/minivision-ai/Silent-Face-Anti-Spoofing).

[6] L. Shangguan, Z. Yang, A. X. Liu, Z. Zhou, and Y. Liu. Relative Localization of Rfid Tags using Spatial-temporal Phase profiling. In NSDI, 2015.

[7] Xiaochen Liu, Yurong Jiang, Kyu-Han Kim, Ramesh Govindan, In "Grab: Fast and Accurate Sensor Processing for Cashier-Free Shopping", 2020.

[8] K. Ha, Z. Chen, W. Hu, W. Richter, P. Pillai, and M. Satyanarayanan. Towards Wearable Cognitive assistance. In Proceedings of the 12th annual international conference on Mobile systems, applications, and services, 2014.

[9] Florian Schroff, Dmitry Kalenichenko, James Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering", 2015.