Ứng dụng Xử lý ảnh số và video số
Seminar

Nguyễn Bá Công
22127046
nbcong22@clc.fitus.edu.vn

Nguyễn Huỳnh Hải Đăng
22127052
nhhdang22@clc.fitus.edu.vn

Đặng Trần Anh Khoa
22127024
dtakhoa22@clc.fitus.edu.vn

Lời Giới Thiệu

Bài báo cáo này là kết quả của các phần làm việc trong đồ án môn học và đã được trình bày trong buổi seminar của lớp. Báo cáo tổng hợp các nghiên cứu, phân tích và kết quả đạt được từ các hoạt động nhóm và cá nhân liên quan đến nội dung môn học. Qua đó, báo cáo không chỉ thể hiện sự hiểu biết về các vấn đề lý thuyết mà còn ứng dụng vào thực tiễn, giúp làm rõ các khái niệm và kỹ thuật đã học. Bài trình bày seminar đã tạo cơ hội cho việc trao đổi ý tưởng và nhận phản hồi từ giảng viên và bạn bè, từ đó nâng cao khả năng thảo luận và áp dụng kiến thức vào các tình huống thực tế.

Mục lục

1 NHẬN DIỆN HÀNG HOÁ BÁN LẺ

1.1 Giới thiệu

Nhận diện hàng hoá bán lẻ là quá trình ứng dụng Thị giác máy tính để tự động xác định, phân loại và theo dõi sản phẩm trong các siêu thị, cửa hàng.

Ví dụ là Amazon Go, mô hình cửa hàng không thu ngân, ứng dụng nhiều công nghệ hiện đại để nhận diện người dùng, tính tiền tự động theo đơn hàng.

1.2 Phát biểu bài toán

Đầu vào:
- Ảnh hoặc Video của sản phẩm
- Ảnh hoặc Video của kệ hàng

Đầu ra: Thông tin sản phẩm, bao gồm vị trí trên kệ, tên sản phẩm, giá cả, hạn sử dụng, v.v.

1.3 Phương pháp

1.3.1 Truyền thống

1. Template Matching: So khớp đặc trưng là một phương pháp để truy tìm vùng ảnh có chứa đặc điểm hoặc vật thể cụ thể bằng cách so sánh những đặc điểm của ảnh đầu vào với ảnh mục tiêu.

Nhược điểm: Không chống chọi được với phép biến đổi, tức là nếu hình ảnh bị xoay ngang, dọc, chéo, phóng to, thu nhỏ thì cũng không thể nhận dạng được sản phẩm. Nếu vật bị che khuất thì phương pháp này cũng hoạt động kém.

Phương pháp này không phù hợp với bán lẻ.

2. Đặc trưng SIFT (Scale-Invariant Feature Transform): Phương pháp này dựa vào trích xuất đặc trưng, chuyên xác định những điểm đặc trưng (keypoint) không bị ảnh hưởng khi phóng to, thu nhỏ, xoay, cũng như biến đổi affine.

Phương pháp này có 4 bước chính:
- Phát hiện các điểm đặc trưng trong không gian
- Định vị điểm đặc trưng (Keypoint Localization)
- Gán hướng (Orientation Assignment)
- Tạo mô tả về đặc trưng (Keypoint Descriptor)

Ưu điểm:
- Bất biến 1 phần trước phép quay, độ sáng, góc nhìn
- Có thể hoạt động cả khi bị che khuất một phần

Nhược điểm:
- Độ phức tạp tính toán cao
- Hoạt động kém với sản phẩm có ít đặc trưng

1.3.2 Deep Learning

1.3.2.1 Trích xuất đặc trưng

Darknet19 là backbone (mạng nền) dùng trong YOLOv2, gồm 19 lớp tích chập (Convolutional Layers) và 5 lớp pooling.

Cách làm:
- Bỏ lớp cuối cùng (lớp fully-connected và softmax dùng cho phân loại).
- Dùng output của lớp convolution cuối cùng (hoặc trung gian) làm đặc trưng đầu ra.
- Nếu cần, có thể áp dụng thêm Global Average Pooling (GAP) để biến tensor thành vector cố định.

ResNet50 là mạng sâu 50 lớp, sử dụng các khối residual (các shortcut connections) để giải quyết vấn đề gradient biến mất trong mạng rất sâu.

Cách làm:
- Cắt bỏ lớp fully-connected cuối cùng (layer `fc` dùng cho phân loại ImageNet).
- Lấy đầu ra tại phần Global Average Pooling (`avg_pool`) để làm vector đặc trưng.

- Đầu ra là một vector 2048 chiều (vì ResNet50 sau pooling có 2048 channels).

1.3.2.2 Đề xuất vùng

- Thay vì dự đoán trực tiếp toạ độ bounding box như các phương pháp cũ, YOLOv2 sử dụng anchor boxes - các hộp có hình dạng/kích thước cố định làm mẫu dự đoán.
- Các anchor boxes được tối ưu bằng thuật toán K-means Clustering trên tập dữ liệu huấn luyện, nhằm tìm ra các kích thước phổ biến của vật thể (hàng hóa) cần nhận diện.
- Trong bài toán này, sử dụng 5 anchor boxes, đại diện cho 5 nhóm kích thước điển hình của các sản phẩm bán lẻ.

Các bước: Mỗi ảnh đầu vào sẽ được chia thành một lưới 13×13 ô. Tại mỗi ô lưới, mô hình dự đoán 5 bounding boxes, tương ứng với 5 anchor boxes đã xác định trước.

Với mỗi bounding box, mô hình dự đoán:

- Offsets: (t_x, t_y, t_w, t_h) - giá trị điều chỉnh (dịch chuyển và thay đổi kích thước) so với anchor box gốc.
- Confidence score: Xác suất có vật thể nằm trong bounding box, đồng thời phản ánh độ chính xác (IoU) của dự đoán.

Region Proposal Network - RPN là một mạng tích chập nhỏ được gắn trực tiếp lên trên backbone (mạng trích xuất đặc trưng), có nhiệm vụ tự động đề xuất các vùng có khả năng chứa vật thể.

Cấu trúc

RPN là một mạng chia làm 2 nhánh song song:

Object Classifier (Nhánh phân loại nhị phân):

- Phân loại mỗi anchor box thành 2 lớp: Object (có vật thể) hoặc Background (không có vật thể).
- Output: Xác suất (score) thể hiện mức độ tin cậy vật thể có tồn tại ở vị trí anchor đó.

Object Regressor (Nhánh hồi quy toạ độ):

- Dự đoán chính xác hơn vị trí và kích thước của bounding box so với anchor ban đầu.
- Dự đoán các giá trị dịch chuyển và tỉ lệ thay đổi: offsets (t_x, t_y, t_w, t_h) cho các anchor có điểm objectness cao.

1.3.2.3 Object Recognition

1.4 Nhận xét

- Nhóm vẫn chưa phát biểu được về cách thức phân loại trong những tình huống cụ thể, như phân biệt các sản phẩm cùng loại, khác nhãn hiệu (Coca-Cola với Pepsi, Sữa Vinamilk và sữa TH, ...)
- Trong phần 2, nhóm không nêu rõ được mình sẽ mục tiêu thực hiện của công trình là gì.

2 HỆ THỐNG TRUY VẾT ĐỐI TƯỢNG DỰA VÀO CÂU MÔ TẢ

2.1 Bối Cảnh Chung

Hệ thống truy vết đối tượng dựa vào câu mô tả là một nhiệm vụ quan trọng trong lĩnh vực xử lý ảnh và video, đặc biệt trong các ứng dụng như xe tự hành, giao thông và an ninh. Việc theo dõi đối tượng trong môi trường phức tạp đối diện với các thách thức như:

- Hạn chế ngôn ngữ mô tả
- Phân biệt đối tượng mục tiêu
- Truy vết trong điều kiện thay đổi môi trường và góc nhìn.

2.2 Phát Biểu Bài Toán

Vấn đề chính của hệ thống là phát hiện và theo dõi đối tượng trong video dựa trên các mô tả ngữ nghĩa. Mô hình phải có khả năng xử lý cả thông tin hình ảnh và ngữ nghĩa, đảm bảo việc truy vết chính xác dù đối tượng có thay đổi hoặc biến mất trong một số frame.

2.3 Các Công Trình Liên Quan

2.3.1 TP-GMOT: Tracking Generic Multiple Object by Textual Prompt with Motion Appearance Cost SORT

Phương pháp TP-GMOT sử dụng mô tả văn bản và tính toán chi phí chuyển động để theo dõi đối tượng đa dạng trong video.

2.3.2 DTLLM-VLT:

Tại một frame bất kỳ trong video, có thể gặp hai trường hợp quan trọng: một là đối tượng bị biến mất, hai là đối tượng xuất hiện trở lại. Trong cả hai trường hợp, câu mô tả phải đóng vai trò quan trọng trong việc giúp mô hình nhận diện và theo dõi đối tượng.

2.4 Phương Pháp Tiếp Cận

Để giải quyết bài toán này, chúng ta áp dụng các phương pháp học sâu kết hợp hai loại dữ liệu: hình ảnh và ngôn ngữ. Cách tiếp cận chính của bài toán là sử dụng một mô hình Transformer với các phương pháp căn chỉnh đa phương thức giữa ngôn ngữ và hình ảnh.

Các bước chính trong phương pháp này bao gồm:

- **Căn Chỉnh Đặc Trưng Đa Phương Thức (Multi-Modal Alignment)**: Các đặc trưng hình ảnh và ngôn ngữ được căn chỉnh với nhau để tạo ra một không gian biểu diễn chung. Quá trình này sử dụng phương pháp Cross-Modal Alignment (CMA) và Intra-Modal Alignment (IMA) để đảm bảo rằng thông tin hình ảnh và ngữ nghĩa có thể tương tác một cách hiệu quả trong quá trình truy vết.
- **Chuyển Đổi và Kết Hợp Đặc Trưng (Transformer Backbone)**: Sau khi các đặc trưng hình ảnh và ngôn ngữ đã được căn chỉnh, chúng được đưa vào một mô hình Transformer để tính toán mối quan hệ giữa các đối tượng trong video và mối liên kết ngữ nghĩa với câu mô tả. Mô hình Transformer sẽ sử dụng các lớp attention để học được mối quan hệ giữa các đặc trưng.
- **Đầu Ra Dự Đoán (Tracking Head)**: Mô hình dự đoán vị trí và kích thước của đối tượng trong video. Đầu ra được chia thành hai nhánh: phân loại (classifying whether an object is present in a region) và hồi quy (regressing the bounding box of the object).

2.5 Cách Làm Chi Tiết

Quá trình xử lý trong hệ thống được thực hiện theo các bước sau:

1. **Tạo Mô Tả Ngôn Ngữ**: Lời nhắc ngôn ngữ được tạo ra để mô tả đối tượng cần theo dõi. Câu mô tả này có thể bao gồm các tính năng như màu sắc, hình dáng, kích thước, hoặc các đặc điểm khác của đối tượng.

2. **Căn Chỉnh Các Đặc Trưng Hình Ảnh và Ngôn Ngữ**: Phương pháp Cross-Modal Alignment Loss (CMA) được sử dụng để tối đa hóa thông tin chung giữa các đặc trưng ngữ nghĩa (câu mô tả) và đặc trưng hình ảnh (vùng tìm kiếm và vùng mẫu). Trong quá trình này, các đặc trưng tương đồng sẽ được gom nhóm lại, còn các đặc trưng khác biệt sẽ được phân biệt.
3. **Sử Dụng Transformer Để Xử Lý Thông Tin**: Các đặc trưng hình ảnh và ngôn ngữ được đưa vào mô hình Transformer, nơi các lớp attention sẽ học được mối quan hệ giữa các đối tượng trong các frame khác nhau. Mô hình sẽ dự đoán vị trí của đối tượng trong các frame tiếp theo.
4. **Dự Đoán Vị Trí và Kích Thước Đối Tượng**: Cuối cùng, hệ thống sẽ sử dụng một nhánh phân loại và một nhánh hồi quy trong tracking head để xác định đối tượng và dự đoán chính xác vị trí của nó. Nhánh phân loại sẽ xác định xem đối tượng có mặt trong một vùng nhất định hay không, trong khi nhánh hồi quy sẽ dự đoán kích thước hộp bao quanh đối tượng.
5. **Cải Tiến và Huấn Luyện Mô Hình**: Các mô hình được huấn luyện và kiểm thử trên các bộ dữ liệu như LaSOT và WebUAV-3M. Quá trình huấn luyện sẽ tối ưu hóa các hàm loss như CMA và IMA để cải thiện độ chính xác của hệ thống trong việc theo dõi đối tượng.

Mô hình được huấn luyện với dữ liệu từ các bộ dữ liệu video thực tế để kiểm tra khả năng của nó trong các tình huống phức tạp. Qua đó, mô hình có thể cải thiện khả năng nhận diện và theo dõi các đối tượng ngay cả khi chúng thay đổi vị trí hoặc xuất hiện lại sau khi bị biến mất trong một số frame.

2.6 Kiến Trúc Mô Hình

Mô hình tổng thể của hệ thống bao gồm:

- **Multi-Modal Alignment Module**: Sử dụng phương pháp Cross-Modal Alignment Loss (CMA) để căn chỉnh đặc trưng giữa ngôn ngữ và hình ảnh.
- **All-in-One Transformer Backbone**: Kết hợp đặc trưng hình ảnh và ngôn ngữ qua Modal Mixup và sử dụng các lớp Transformer để học mối quan hệ giữa chúng.
- **Tracking Head**: Bao gồm hai nhánh phân loại và hồi quy, giúp dự đoán vị trí và kích thước của đối tượng.

2.7 Kết Quả và Đánh Giá

Hệ thống đã được huấn luyện và kiểm thử trên các bộ dữ liệu như LaSOT và WebUAV-3M. Các kết quả cho thấy mô hình đạt hiệu suất cao, đặc biệt trong việc xử lý các tác động từ lời nhắc ngôn ngữ mơ hồ. Các thí nghiệm cho thấy rằng phương pháp All-in-One giúp cải thiện độ chính xác trong việc theo dõi đối tượng trong các tình huống phức tạp.

2.8 Nhận xét

Mô hình All-in-One Transformer kết hợp ngôn ngữ và hình ảnh một cách hiệu quả, đạt được kết quả vượt trội trong việc theo dõi đối tượng. Tuy nhiên, cần cải thiện khả năng xử lý các lời nhắc ngôn ngữ không chính xác hoặc mơ hồ để tăng cường tính chính xác và linh hoạt của hệ thống.

3 PHÁT HIỆN BẤT THƯỜNG TRONG GIAO Thông

3.1 Phát Biểu Bài Toán

3.1.1 Mục Tiêu

Bài toán đặt ra là phát hiện các bất thường trong giao thông bằng cách phân tích video từ camera hành trình hoặc camera an ninh. Hệ thống mục tiêu là tính toán xác suất xảy ra tai nạn tại mỗi frame của video và cảnh báo mức độ nguy hiểm nếu vượt qua ngưỡng định trước. Việc xác định tai nạn và cảnh báo kịp thời là quan trọng để nâng cao hiệu quả giám sát và giảm thiểu tai nạn giao thông.

3.1.2 Đầu Vào và Đầu Ra

Đầu vào: Một đoạn video thu từ camera hành trình hoặc camera an ninh.

- Video ghi lại cảnh giao thông từ một hoặc nhiều camera.
- Dữ liệu từ các khung hình video, bao gồm hình ảnh và chuyển động của các phương tiện.

Đầu ra:

- Xác suất tai nạn trong mỗi frame của video.
- Cảnh báo nếu xác suất vượt qua ngưỡng định trước (Threshold) để cảnh báo mức độ nguy hiểm.

3.2 Phương Pháp Tiếp Cận

Hệ thống này sử dụng phương pháp MEDAVET (Traffic Vehicle Anomaly Detection Mechanism based on spatial and temporal structures in vehicle traffic), kết hợp với YOLOv7 để phát hiện và theo dõi các phương tiện trong video. Các phương pháp chính trong hệ thống bao gồm:

1. **Phát hiện đối tượng với YOLOv7**: YOLOv7 (You Only Look Once version 7) được sử dụng để phát hiện các phương tiện trong các khung hình video. Đây là một mô hình học sâu có khả năng phát hiện đối tượng trong ảnh với tốc độ và độ chính xác cao.
2. **Theo dõi hành trình phương tiện**: Sau khi phát hiện đối tượng, hệ thống sử dụng đồ thị để theo dõi hành trình của các phương tiện qua các khung hình liên tiếp. Mỗi phương tiện sẽ được theo dõi dựa trên các đặc trưng vị trí và chuyển động trong không gian và thời gian.
3. **Cấu trúc dữ liệu QuadTree**: QuadTree là một cấu trúc dữ liệu giúp tổ chức không gian trong video và phân tích hành vi của các phương tiện. Cấu trúc này giúp giảm thiểu độ phức tạp tính toán trong việc phân tích và theo dõi các phương tiện qua thời gian.
4. **Xác suất tai nạn**: Dựa trên hành vi của các phương tiện, hệ thống tính toán xác suất xảy ra tai nạn trong mỗi frame. Các yếu tố như tốc độ, khoảng cách giữa các phương tiện, và các bất thường trong hành vi di chuyển sẽ ảnh hưởng đến xác suất này.

3.3 Nhận Xét

- **Chưa giải thích cơ chế tìm chiều di chuyển và vận tốc của phương tiện**: Một yếu tố quan trọng trong việc tính toán xác suất tai nạn là việc xác định hướng di chuyển và vận tốc của các phương tiện. Việc này chưa được giải thích rõ ràng, mặc dù đó là yếu tố quan trọng trong việc dự đoán và cảnh báo tai nạn.
- **Thiếu giải thích chung về hệ thống trước khi đi vào thuật toán và biểu đồ**: Các biểu đồ và thuật toán mô tả trong hệ thống cần phải được giải thích đầy đủ hơn về ý nghĩa và cách thức hoạt động của chúng trước khi đi sâu vào chi tiết. Cần phải có phần giải thích tổng quan về hệ thống trước khi thảo luận về các thuật toán cụ thể.

- **Cần giải thích rõ về dữ liệu không gian và thời gian (spatial-temporal)**: Việc xử lý dữ liệu không gian và thời gian là một thách thức lớn trong bài toán này. Cần nêu rõ các phương pháp được sử dụng để phân tích dữ liệu này, chẳng hạn như cách mà hệ thống theo dõi phương tiện qua không gian và thời gian, và cách tính toán sự bất thường trong hành vi giao thông.
- **Cần hiểu rõ "Thế nào là tai nạn?"**: Để hệ thống có thể xác định chính xác thời điểm xảy ra tai nạn, cần có một định nghĩa rõ ràng về tai nạn. Điều này sẽ giúp xác định được chính xác khi nào hệ thống cần cảnh báo.
- **Cần làm rõ đối tượng cần quan tâm trong mỗi frame**: Hệ thống cần chỉ rõ đối tượng nào cần được theo dõi và tính toán xác suất tai nạn trong mỗi frame. Liệu tất cả các phương tiện hay chỉ các phương tiện di chuyển không bình thường mới cần được tính toán xác suất tai nạn?
- **Xử lý dữ liệu từ xe bị tai nạn**: Dữ liệu thu thập từ các xe đã bị tai nạn là rất quan trọng, nhưng cách thức gán nhãn cho dữ liệu này chưa được làm rõ. Cần phải giải thích quy trình gán nhãn cho các video hoặc dữ liệu giao thông để huấn luyện hệ thống.

3.4 Tổng Quan

Hệ thống sử dụng YOLOv7 và cấu trúc dữ liệu QuadTree để phát hiện và theo dõi các phương tiện trong giao thông, đồng thời tính toán xác suất xảy ra tai nạn. Tuy nhiên, một số yếu tố như việc xác định vận tốc, chiều di chuyển của phương tiện, và cách gán nhãn cho dữ liệu cần được làm rõ để cải thiện độ chính xác và khả năng hoạt động của hệ thống trong thực tế. Việc hiểu rõ hơn về các bất thường trong hành vi giao thông và cách xác định tai nạn sẽ giúp hệ thống hoạt động hiệu quả hơn và có thể cảnh báo kịp thời.

4 BONE DISEASE VQA BASED ON MULTIMODAL TRANSFORMER

4.1 Phương pháp

Decoder:
- Decoder giải mã và tìm cách liên kết với encoder

Encoder:
- Ảnh chụp y khoa được đưa vào Vision Encoder là SWIN
- Câu hỏi của bác sĩ được đưa vào Text Encoder là ViHealthBERT
- Kết quả của 2 encoder được đưa vào Fusion, gọi là CMAN.
- Chuyển tiếp qua Decoder có Learnable Answer
- MLP: có Sigmoid, Cross Entropy, AdamW (có weight decay để tránh làm ảnh hưởng đến Gradient khi loss thay đổi nhiều)
- Output là Class ID.

Vấn đề là cơ chế Generation tốn quá nhiều tài nguyên, nên chọn cơ chế Classification.

Dataset kết hợp hình ảnh xét nghiệm và

Training:
- Giai đoạn 1: Train 6 epoch, trọng số không đổi
- Giai đoạn 2: Train 3 epoch, có cho thay đổi trọng số

Thời gian train là cho cả 2 giai đoạn là hơn 10 tiếng.

⇒

4.2 Nhận xét

- Cần chắt lọc dữ liệu lại, nếu dữ liệu y khoa quá lớn
- Cần tự bổ sung thêm dữ liệu bằng cách đặt câu hỏi tương ứng.
- Nên sắp xếp câu hỏi theo category: What?, Where?
- Chưa giải thích rõ được cách đưa dữ liệu học vào mô hình: tức là 1 bảng, các cột là hình ảnh - câu hỏi - câu trả lời. Việc đưa raw data vào mô hình là vô lý.
- Cần trình bày câu hỏi, kết quả theo từng nhóm bệnh

5 ĐIỂM DANH LỚP HỌC VÀ ĐÁNH GIÁ ĐƯỜNG CONG THÁI ĐỘ HỌC TẬP

5.1 Phát Biểu Bài Toán

5.1.1 Mục Tiêu

Mục tiêu của hệ thống là tự động hóa quá trình điểm danh và đánh giá thái độ học tập của sinh viên. Qua đó, giúp giảng viên có thể theo dõi tình trạng học tập của từng sinh viên và cả lớp, điều chỉnh nội dung giảng dạy kịp thời để cải thiện chất lượng học tập. Hệ thống sử dụng công nghệ nhận diện khuôn mặt và phân tích cảm xúc qua các video lớp học thu được từ camera.

5.1.2 Đầu Vào và Đầu Ra

Đầu vào: Dãy ảnh hoặc video từ camera giám sát lớp học.

- Video lớp học ghi lại trong suốt quá trình giảng dạy.
- Các hình ảnh có chứa khuôn mặt của sinh viên trong lớp.

Đầu ra:

- Bảng điểm danh tự động.
- Đường cong thái độ học tập của từng sinh viên và của toàn bộ lớp.
- Lưu trữ và phân tích xu hướng thái độ học tập qua thời gian.
- Phân tích nguyên nhân tiêu cực nếu có, giúp giảng viên nhận diện sinh viên có thái độ học tập không tích cực.

5.2 Phương Pháp Tiếp Cận

Hệ thống này áp dụng các phương pháp xử lý ảnh số và AI, bao gồm các bước chính sau:

1. **Face Detection (Phát hiện khuôn mặt)**: Sử dụng các thuật toán phát hiện khuôn mặt để nhận diện khuôn mặt của các sinh viên trong lớp học qua từng khung hình video.
2. **Face Recognition (Nhận diện khuôn mặt)**: Xác định và phân loại khuôn mặt của sinh viên dựa trên cơ sở dữ liệu có sẵn, để liên kết khuôn mặt với danh tính của từng sinh viên.
3. **Face Emotion Recognition (Nhận diện cảm xúc khuôn mặt)**: Phân tích các biểu cảm khuôn mặt của sinh viên, từ đó tính toán các giá trị Valence và Arousal, phản ánh mức độ tích cực, tiêu cực và mức độ hứng thú của sinh viên trong suốt buổi học.

4. **Action Reception (Nhận diện hành động)**: Phát hiện các hành động như giơ tay, đọc sách, sử dụng điện thoại, cúi đầu hoặc ngủ, giúp đánh giá mức độ tập trung của sinh viên trong lớp học.

5.3 Đường Cong Thái Độ Học Tập

Sử dụng mô hình Valence-Arousal để xây dựng đường cong thái độ học tập của sinh viên:

- **Valence**: Mức độ tích cực hoặc tiêu cực của cảm xúc.
- **Arousal**: Mức độ kích thích hoặc mức độ tỉnh táo của cảm xúc.

Đường cong thái độ học tập được tổng hợp từ các giá trị Valence và Arousal theo thời gian, giúp giảng viên đánh giá sự thay đổi trong thái độ học tập của sinh viên và có biện pháp điều chỉnh phù hợp.

5.4 Thách Thức

- **Nhận diện khuôn mặt trong điều kiện phức tạp**: Các vấn đề như ánh sáng thay đổi, góc nhìn khác nhau hoặc khuôn mặt bị che khuất có thể gây khó khăn trong việc phát hiện chính xác khuôn mặt sinh viên. - **Phân tích cảm xúc chính xác**: Biểu cảm cảm xúc có thể đa dạng và tinh tế, và có thể có nhiều sinh viên cùng lúc, gây khó khăn trong việc phân tích chính xác. - **Xây dựng đường cong có ý nghĩa sư phạm**: Việc chuyển đổi dữ liệu cảm xúc thô thành thông tin hữu ích cho giảng viên để cải thiện phương pháp giảng dạy là một thách thức lớn.

5.5 Ý Nghĩa Ứng Dụng

- **Tự động hóa điểm danh** giúp tiết kiệm thời gian cho giảng viên và làm tăng tính chính xác trong việc ghi nhận sự có mặt của sinh viên. - **Đánh giá thái độ học tập** qua đường cong Valence-Arousal cung cấp dữ liệu khách quan giúp giảng viên điều chỉnh phương pháp giảng dạy theo tình trạng cảm xúc của sinh viên. - **Phát hiện sinh viên có thái độ tiêu cực** giúp giảng viên nhận diện sớm các vấn đề tâm lý của sinh viên, từ đó can thiệp kịp thời, hỗ trợ tư vấn tâm lý nếu cần.

Hệ thống không chỉ có ứng dụng trong việc cải thiện chất lượng giảng dạy mà còn hỗ trợ các hoạt động tư vấn tâm lý, giúp phát hiện các dấu hiệu học sinh bị cô lập hoặc gặp vấn đề về cảm xúc.

Tên ứng dụng: Giám sát thái độ học tập của sinh viên

5.6 Nhận xét:

- Phương pháp sẽ chạy chậm vì áp dụng quá nhiều tác vụ khác nhau, gây tốn kém không cần thiết.
- Cách làm tương đối tốt, nhưng trình bày khó hiểu

6 ĐỊNH VỊ VÀ TÁI TẠO MÔI TRƯỜNG XUNG QUANH

6.1 Introduction

Đối với một robot di động khám phá một môi trường tĩnh chưa biết, việc đ ịnh vị chính xác vị trí của nó đồng thời xây dựng bản đồ là một vấn đề "gà và trứng", được biết đến với tên gọi Định vị và xây dựng bản đồ đồng thời (Simultaneous Localization And Mapping).

6.2 Động lực nghiên cứu

Khoa học: Để cải thiện khả năng nhận diện đặc trưng, theo dõi, và tái lập bản đồ trong điều kiện phức tạp (ánh sáng kém, chuyển động nhanh, cảnh lặp…).

Trong nghiên cứu này, chúng ta sẽ xem xét cách xây dựng bản đồ 3D dựa trên mô hình đồ thị bằng cách:

1. Theo dõi các đặc trưng t hị giác như SIFT/SURK
2. Tính toán các phép biến đổi hình học với RANSA>
3. Áp dụng các kỹ t huật tối ưu phi tuyến để ước lượng quỹ đạo di chuyển.

6.3 Phát biểu bài toán

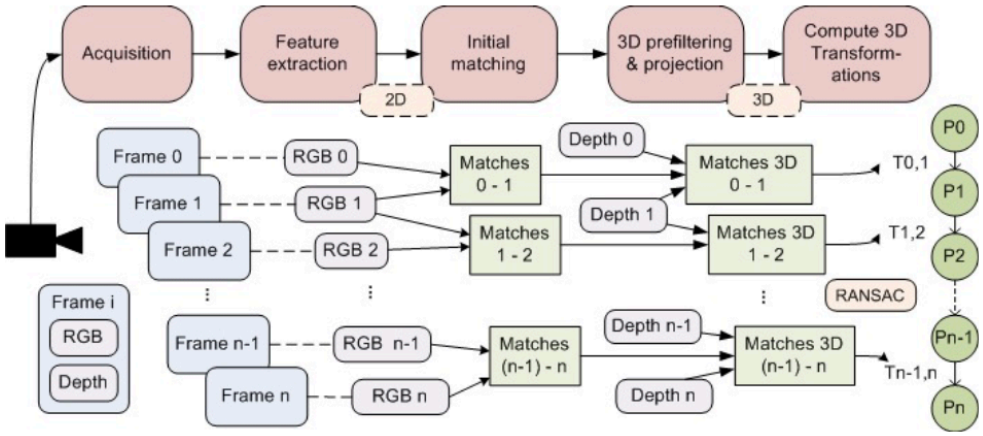
Input: $\{f(t), D(t), (t = 1, ..., n)\}$

1. f(t) với tọa độ (x,y): là hình ảnh thông thường (RGB) được thu thập từ cảm biến màu của hệ thống. Với f(t), t đại diện cho thời gian hoặc số khung hình mà cảm biến thu thập tại thời điểm đó. Dữ liệu này chứa thông tin màu sắc của các đối tượng trong cảnh quan.
2. D(t): là dữ liệu độ sâu (depth data) thu được từ cảm biến RGB-D, chẳng hạn như Kinect. Dữ liệu này chứa thông tin về khoảng cách từ camera đến các đối tượng trong không gian.

Output: { Oxyz(t), Point cloud(t) (.ply)

1. Oxyz(t): là vị trí 3D của camera hoặc robot tại thời điểm t, có thể biểu diễn dưới dạng một vector 3D, O(t) = Ox(t), Oy(t), Oz(t). Đây là tọa độ 3D của camera trong không gian.
2. Point cloud(t): liên quan đến 3D reconstruction, đại diện cho một tập hợp các điểm 3D trong không gian. Các point clouds là kết quả của việc định vị và xây dựng bản đồ. Khi camera di chuyển, mỗi khung hình tạo ra một point cloud mới, giúp xây dựng một bản đồ không gian 3D của môi trường.

6.4 Sợ đồ hệ thống:



6.5 Phát Biểu Bài Toán

Bài toán trong nghiên cứu này là về việc xây dựng hệ thống định vị và tạo bản đồ đồng thời (SLAM) cho robot hoặc xe tự hành sử dụng cảm biến RGB-D. Hệ thống này sử dụng

dữ liệu ảnh màu (RGB) và dữ liệu độ sâu (depth) để xác định vị trí của robot và xây dựng bản đồ 3D của môi trường xung quanh. Cụ thể, bài toán đặt ra là làm sao để có thể xử lý dữ liệu hình ảnh từ các cảm biến RGB-D, từ đó ước lượng quỹ đạo di chuyển của robot, xác định vị trí của các đặc trưng trong môi trường, và xây dựng một bản đồ 3D chính xác trong thời gian thực.

6.6 Phương Pháp Tiếp Cận

Để giải quyết bài toán này, hệ thống SLAM sử dụng các phương pháp nhận diện đặc trưng và khớp các điểm đặc trưng giữa các khung hình liên tiếp. Các đặc trưng này bao gồm những điểm đặc trưng mạnh mẽ được trích xuất từ hình ảnh sử dụng các thuật toán như SIFT (Scale-Invariant Feature Transform) và SURF (Speeded-Up Robust Features). Sau khi trích xuất đặc trưng, hệ thống sử dụng các phương pháp tối ưu hóa để tính toán phép biến đổi hình học giữa các khung hình và tạo ra bản đồ không gian 3D.

6.7 Cách Làm Chi Tiết

Bài toán được giải quyết qua các bước sau:

1. **Dữ Liệu Đầu Vào**: Hệ thống sử dụng hai loại dữ liệu chính:
    - $f(t)$ là hình ảnh màu RGB thu được từ cảm biến tại thời điểm t. Đây là các khung hình mà cảm biến ghi nhận, chứa thông tin về màu sắc của các đối tượng trong cảnh.
    - $D(t)$ là dữ liệu độ sâu thu được từ cảm biến RGB-D (ví dụ như Kinect), giúp xác định khoảng cách từ camera đến các đối tượng trong không gian.
2. **Trích Xuất Đặc Trưng**: Các đặc trưng hình ảnh được trích xuất từ các khung hình sử dụng thuật toán SIFT hoặc SURF. Các đặc trưng này cho phép hệ thống nhận diện các điểm quan trọng trong môi trường và theo dõi chúng qua các khung hình khác nhau.
3. **Khớp Các Đặc Trưng**: Sau khi trích xuất các đặc trưng, hệ thống sử dụng kỹ thuật khớp đặc trưng giữa các khung hình. Các điểm đặc trưng này sẽ được ghép nối thông qua phương pháp KD-tree, giúp tối ưu hóa quá trình tìm kiếm các cặp điểm tương ứng trong không gian đặc trưng.
4. **Ước Lượng Biến Hình 3D**: Dựa trên các cặp điểm đã được khớp, hệ thống sẽ tính toán một phép biến hình 3D (bao gồm các phép quay và dịch chuyển) để chuyển đổi các điểm từ khung hình nguồn sang khung hình đích. Phép biến hình này được tính toán bằng phương pháp bình phương tối thiểu, giúp đảm bảo rằng các điểm khớp với độ chính xác cao.
5. **Tạo Bản Đồ 3D**: Khi đã ước lượng được quỹ đạo di chuyển của camera, hệ thống tiếp tục tạo ra một bản đồ 3D của môi trường. Các point cloud được tạo ra từ dữ liệu RGB và độ sâu sẽ được kết hợp và đăng ký (registration) lại để tạo ra bản đồ không gian 3D hoàn chỉnh.
6. **Tối Ưu Hóa Đồ Thị**: Các phép biến đổi giữa các khung hình được lưu trữ trong một đồ thị, trong đó mỗi nút đại diện cho một vị trí của camera (keypose). Các cạnh trong đồ thị biểu thị sự chuyển động giữa các vị trí này. Đồ thị sẽ được tối ưu hóa bằng phương pháp Levenberg-Marquardt để cải thiện độ chính xác của quỹ đạo và bản đồ.

6.8 Kết quả của bài toán là:

- $Oxyz(t)$: Vị trí 3D của camera hoặc robot tại thời điểm t, được biểu diễn dưới dạng vector 3D $O(t) = (Ox(t), Oy(t), Oz(t))$.

- Point cloud (t): Là bộ dữ liệu 3D thể hiện các điểm trong không gian, giúp tạo thành bản đồ 3D của môi trường xung quanh.

Bài toán này không chỉ là vấn đề về định vị mà còn liên quan đến việc xây dựng bản đồ trong môi trường thực tế, giúp các robot có thể tự động di chuyển và nhận diện các đối tượng xung quanh một cách chính xác và hiệu quả.

6.9 Nhận xét

Hệ thống SLAM sử dụng cảm biến RGB-D để giải quyết bài toán định vị và xây dựng bản đồ đồng thời đã thể hiện được tính hiệu quả trong môi trường thực tế. Phương pháp trích xuất và khớp đặc trưng sử dụng các thuật toán mạnh mẽ như SIFT và SURF đã cho phép hệ thống nhận diện và theo dõi các đối tượng trong môi trường với độ chính xác cao, ngay cả khi có sự thay đổi về góc nhìn và ánh sáng.

Một điểm mạnh của hệ thống là khả năng sử dụng dữ liệu độ sâu từ cảm biến RGB-D để xác định chính xác vị trí 3D của camera hoặc robot trong không gian, đồng thời xây dựng bản đồ 3D của môi trường. Việc áp dụng phương pháp bình phương tối thiểu để ước lượng các phép biến hình cũng giúp cải thiện độ chính xác của các phép chuyển đổi giữa các khung hình.

Tuy nhiên, hệ thống cũng còn một số hạn chế. Việc xử lý các point cloud trong môi trường có nhiều vật thể di động có thể gặp khó khăn do sự thay đổi nhanh chóng của các đặc trưng. Mặc dù phương pháp tối ưu hóa đồ thị giúp cải thiện quỹ đạo di chuyển và bản đồ, nhưng trong các tình huống phức tạp như môi trường đông đúc hoặc ánh sáng yếu, độ chính xác của hệ thống có thể giảm.

Ngoài ra, việc xây dựng bản đồ 3D từ các point cloud yêu cầu một lượng tính toán lớn, điều này có thể gây khó khăn khi triển khai trên các hệ thống có tài nguyên tính toán hạn chế. Việc cải tiến hiệu suất và giảm độ trễ trong quá trình xử lý vẫn là một thách thức cần được giải quyết trong các nghiên cứu tiếp theo.

Mặc dù vậy, phương pháp SLAM sử dụng cảm biến RGB-D này vẫn là một bước tiến quan trọng trong việc phát triển các hệ thống tự động hóa, đặc biệt trong các ứng dụng như robot di động, xe tự hành, và các hệ thống giám sát an ninh.

7 XÂY DỰNG HỆ THỐNG TÍNH CHỈ SỐ HẤP THỤ CÁC BON TỪ CÂY TRỒNG

Carbon index (Chỉ số các-bon) là một thước đo dùng để đánh giá khả năng hâp thụ và lưu trữ CO2 của cây trồng hoặc hệ sinh thái.

7.1 Ý nghĩa khoa học

- Định lượng CO2 hấp thụ từ cây trồng tạo điều kiện nghiên cứu về môi trường và biến đổi khí hậu
- Hỗ trợ cho nghiên cứu của các lĩnh vực sinh học và tài nguyên - môi TRƯỜNG

7.2 Ý nghĩa ứng dụng

- Phục vụ quản lý tài nguyên và môi trường rừng
- Cơ sở xây dựng chính sách về bảo vệ môi trường và chống biến đổi khí hậu
- Công cụ chính yếu cho thị trường tín chỉ các bon

⇒ Một phần không thể thiếu trong bảo vệ môi trường

7.3 Phát biểu bài toán

Đầu vào:
- Dữ liệu quang học: Ảnh RGB, ảnh đa phổ, ảnh siêu phổ
- Dữ liệu cao độ: LiDAR, GEDI, Photogrammetry
- Dữ liệu kiểm chứng: Các kết quả nghiên cứu có sẵn, Sinh khối thực tế, DBH thực tế

Đầu ra:
- Bản đồ Carbon Density: Dữ liệu raster hiển thị mật độ carbon trên toàn bộ khu vực rừng
- Bảng số liệu thống kê tổng hợp mật độ carbon theo khu vực
- Dữ liệu tích hợp GIS

Các công đoạn chính:
1. Tiền xử lý dữ liệu: Tải và xử lý, chuẩn hoá dữ liệu ảnh Sentinel-2.
2. Xử lý, trích xuất đặc trưng từ ảnh vệ tinh: Tính NDVI để xác định vùng rừng
3. Dự đoán mật độ các bon với mô hình AI
4. Xuất kết quả & phân tích: Tạo bản đồ Carbon Density, tạo báo cáo về mật độ các bon, so sánh kết quả với dữ liệu kiểm chứng

Đóng góp: Phát triển hệ thống tính toán tự động, cung cấp dữ liệu chính xác cho nghiên cứu và ứng dụng thực tế.
- Tăng độ chính xác trong việc ước tính lượng các-bon lưu trữ.
- Giảm chi phí và thời gian so với các phương pháp truyền thống.
- Phát hiện nhanh chóng các thay đổi trong diện tích rừng, đặc biệt là do phá rừng để nhanh chóng thông báo và xử lý

7.4 Công trình nghiên cứu liên quan

7.4.1 Giai đoạn truyền thống

Đây là giai đoạn diễn ra vào những năm 1980 - 2000, tập trung vào kiểm kê rừng và công thức sinh khối

Phương pháp sinh khối toàn cây:
- Sử dụng phương trình toán học để tính toán biomass từ đường kính thân cây
- Thực hiện hồi quy tuyến tính từ khảo sát thực địa
- Ưu điểm: Dễ thực hiện
- Nhược điểm: Tốn nhiều công sức, và hiệu quả phụ thuộc vào dữ liệu thực địa

Phương pháp đánh giá bằng ô tiêu chuẩn
- Đo đạc thực tế qua các ô tiêu chuẩn, sau đó nội suy cho toàn khu vực
- Thống kê từ khảo sát thực địa
- Ưu điểm: Phù hợp với diện tích nhỏ, độ chính xác cao
- Nhược điểm: Chỉ phù hợp nếu thu thập dữ liệu rộng và đầy đủ

7.4.2 Giai đoạn viễn thám truyền thống (2000s - 2015)

Tóm tắt:
- Dữ liệu ảnh vệ tinh (MODIS, Landsat) giúp mở rộng quy mô tính toán carbon.
- Radar SAR khắc phục được nhược điểm mây che, nhưng xử lý phức tạp.
- LiDAR cho kết quả chính xác nhất, nhưng chi phí cao.
- Ưu điểm: Khả năng tính toán trên diện rộng, không cần khảo sát thực địa nhiều.

- Nhược điểm: Các mô hình còn đơn giản (hồi quy tuyên tính), chưa tận dụng AI

Phương pháp Chỉ số thực vật NDVI:

- Sử dụng chỉ số thực vật NDVI từ ảnh vệ tinh để ước tính ảnh sinh khối
- Phương pháp: Hồi quy tuyến tính NDVI
- Sử dụng Dataset Landsat 5/7
- Ưu điểm: Dễ áp dụng, không cần khảo sát thực địa
- Nhược điểm: Độ chính xác thấp ở rừng nhiệt đới, bị ảnh hưởng bởi đất trống.

Tham khảo