BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

**ĐẠI HỌC QUỐC GIA HÀ NỘI**



**BÁO CÁO MÔN TRÍ TUỆ NHÂN TẠO**

**ĐỀ TÀI**

**GIAO TIẾP, NHẬN THỨC, HÀNH ĐỘNG**

Nhóm: 9

|  |  |  |
| --- | --- | --- |
| Giảng viên | : | TS. Trần Hồng Việt |
| Mã lớp học phần | : | INT3401\_20 |
| Sinh viên | : | Nguyễn Cảnh Quân - 21021357  Phùng Gia Quang – 21021355  Lê Đăng Quân – 21020157  Nguyễn T. Minh Quân – 21021358  Trần Anh Quân - 21020593 |

|  |  |  |
| --- | --- | --- |
|  |  |  |

**Mục lục**

[**I.** **Xử lý ngôn ngữ NLP** 1](#_Toc167436020)

[**1.** **Các loại và giai đoạn của NLP** 1](#_Toc167436021)

[**2.** **Ngôn ngữ tự nhiên cho giao tiếp và nhận thức** 2](#_Toc167436022)

[**II.** **Tri giác (perception)** 4](#_Toc167436023)

[**1.** **Giới thiệu** 4](#_Toc167436024)

[**2.** **Tạo ảnh** 4](#_Toc167436025)

[**3.** **Hoạt động xử lý hình ảnh** 9](#_Toc167436026)

[**4.** **Thu nhập thông tin 3 chiều sử dụng hình ảnh hoặc dữ liệu** 13](#_Toc167436027)

[**5.** **Sử dụng tầm nhìn để điều khiển và điều hướng** 14](#_Toc167436028)

[**6.** **Hiển thị và nhận diện đối tượng** 16](#_Toc167436029)

[**7.** **Nhận diện giọng nói** 20](#_Toc167436030)

[**III.** **Robotics** 28](#_Toc167436031)

[**1.** **Khái niệm robot là gì?** 28](#_Toc167436032)

[**2.** **Robot làm được những gì ?** 29](#_Toc167436033)

[**3.** **Robot được làm từ những gì?** 30](#_Toc167436034)

[**4.** **Thu nhập thông tin 3 chiều sử dụng hình ảnh hoặc dữ liệu** 35](#_Toc167436035)

# **Xử lý ngôn ngữ NLP**

Xử lý ngôn ngữ tự nhiên (NLP) là một nhánh của trí tuệ nhân tạo (AI) tập trung vào sự tương tác giữa máy tính và con người thông qua ngôn ngữ tự nhiên. Nó cho phép máy tính hiểu, diễn giải và tạo ra cả ngôn ngữ của con người theo cách có ý nghĩa và hữu ích. Ngôn ngữ học tính toán NLP kết hợp, khoa học máy tính và Kỹ thuật Alto thu hẹp khoảng cách giữa giao tiếp của con người và sự hiểu biết của máy tính. Một số nhiệm vụ chính của NLP bao gồm:

· Phân tích văn bản: Chia văn bản thành các phần cấu thành của nó, chẳng hạn như từ, câu và câu.

· Nhận dạng thực thể được đặt tên (NER): Xác định và phân loại các thực thể được đặt tên được đề cập trong văn bản, chẳng hạn như con người, tổ chức, địa điểm, ngày tháng, v.v.

· Gắn thẻ phần lời nói (POS): Gán các phần ngữ pháp của lời nói cho các từ trong câu (ví dụ: danh từ, động từ, tính từ).

· Phân tích cảm xúc: Xác định cảm xúc hoặc quan điểm được thể hiện trong một đoạn văn bản, dù đó là tích cực, tiêu cực hay trung tính.

· Dịch máy: Dịch văn bản từ ngôn ngữ này sang ngôn ngữ khác một cách tự động.

· Trả lời câu hỏi: Hiểu các câu hỏi được đặt ra bằng ngôn ngữ tự nhiên và đưa ra câu trả lời chính xác. Do đó, việc trả lời câu hỏi tập trung vào việc chúng tôi xây dựng các hệ thống tự động trả lời các câu hỏi của con người bằng ngôn ngữ tự nhiên.

· Tóm tắt văn bản: thông tin chính. Tạo bản tóm tắt ngắn gọn của văn bản dài hơn trong khi vẫn bảo tồn

· Phát hiện thư rác: Được sử dụng để phát hiện các email không mong muốn gửi đến hộp thư đến của người dùng.

NLP có nhiều ứng dụng trên nhiều lĩnh vực khác nhau, bao gồm chatbot dịch vụ khách hàng, trợ lý ảo, hệ thống truy xuất thông tin và phân tích tình cảm trên mạng xã hội, dịch vụ dịch ngôn ngữ, v.v. Khi các kỹ thuật NLP tiếp tục phát triển, khả năng ứng dụng của nó không ngừng mở rộng.

## **Các loại và giai đoạn của NLP**

Xử lý ngôn ngữ tự nhiên (NLP) bao gồm nhiều loại và giai đoạn khác nhau, mỗi loại giải quyết các khía cạnh khác nhau của việc hiểu và xử lý ngôn ngữ. Dưới đây là tổng quan về các loại và giai đoạn của NLP:

Các loại NLP

· NLP dựa trên quy tắc: Trong cách tiếp cận này, các quy tắc ngôn ngữ được các chuyên gia xác định một cách thủ công để phân tích và xử lý văn bản. Những quy tắc này chi phối cách hệ thống diễn giải và tạo ra ngôn ngữ. Mặc dù hiệu quả đối với một số nhiệm vụ nhất định, NLP dựa trên quy tắc có thể bị hạn chế trong việc xử lý các trường hợp mơ hồ và phức tạp.

· NLP thống kê: Các phương pháp thống kê liên quan đến việc sử dụng lượng lớn dữ liệu để đào tạo các mô hình có thể tự động tìm hiểu các mẫu và mối liên kết trong ngôn ngữ. Các kỹ thuật như thuật toán học máy, mô hình xác suất và phân tích thống kê được sử dụng để phân tích và tạo văn bản.

· NLP lai: Cách tiếp cận này kết hợp các kỹ thuật thống kê và dựa trên quy tắc để tận dụng điểm mạnh của cả hai phương pháp. Nó có thể cung cấp khả năng xử lý ngôn ngữ mạnh mẽ và chính xác hơn bằng cách kết hợp các quy tắc ngôn ngữ với các mô hình thống kê dựa trên dữ liệu.

· NLP học sâu: Các kỹ thuật deep learning, đặc biệt là mạng lưới thần kinh, đã cách mạng hóa NLP trong những năm gần đây. Các mô hình như mạng thần kinh tái phát (RNN), mạng thần kinh tích chập (CNN) và kiến trúc biến đổi (ví dụ: BERT, GPT) được sử dụng cho các tác vụ như phân loại văn bản, tạo ngôn ngữ và dịch máy.

## **Ngôn ngữ tự nhiên cho giao tiếp và nhận thức**

"Ngôn ngữ tự nhiên để giao tiếp và nhận thức" đề cập đến khả năng của con người và ngày càng nhiều máy móc hiểu, xử lý và tạo ra ngôn ngữ để giao tiếp hiệu quả và nhận thức thế giới xung quanh. Khái niệm này bao gồm cả việc sử dụng ngôn ngữ như một phương tiện giao tiếp giữa các cá nhân và như một công cụ để nhận thức và giải thích thông tin từ môi trường.

**Communication**

· Giao tiếp giữa các cá nhân: Con người sử dụng ngôn ngữ tự nhiên để truyền đạt suy nghĩ, ý tưởng, cảm xúc và ý định cho người khác. Điều này bao gồm ngôn ngữ nói trong các cuộc trò chuyện trực tiếp, ngôn ngữ viết trong thư và email cũng như giao tiếp kỹ thuật số thông qua tin nhắn trò chuyện và mạng xã hội.

· Tương tác giữa người và máy tính (HCI): Giao diện ngôn ngữ tự nhiên cho phép người dùng tương tác với máy tính và thiết bị bằng ngôn ngữ nói hoặc viết, thay vì các phương thức nhập liệu truyền thống như bàn phím hoặc màn hình cảm ứng. Trợ lý ảo, chatbot và hệ thống điều khiển bằng giọng nói là những ví dụ về ứng dụng HCI tận dụng công nghệ xử lý ngôn ngữ tự nhiên (NLP).

· Giao tiếp đa ngôn ngữ: Ngôn ngữ tự nhiên đóng vai trò là cầu nối giữa những người nói các ngôn ngữ khác nhau, tạo điều kiện giao tiếp và hiểu biết vượt qua các rào cản văn hóa và ngôn ngữ. Các dịch vụ biên dịch và phiên dịch cũng như các công cụ học ngôn ngữ đóng một vai trò quan trọng trong việc hỗ trợ giao tiếp đa ngôn ngữ.

**Perception**

· Hiểu văn bản: Xử lý ngôn ngữ tự nhiên cho phép máy phân tích và diễn giải dữ liệu văn bản, trích xuất ý nghĩa, xác định các thực thể và mối quan hệ cũng như đưa ra các suy luận. Công nghệ hiểu văn bản được sử dụng trong nhiều ứng dụng khác nhau như truy xuất thông tin, phân tích cảm xúc và phân loại nội dung.

· Nhận dạng giọng nói: Máy có thể nhận biết và chuyển ngôn ngữ nói thành văn bản bằng hệ thống nhận dạng giọng nói tự động (ASR). Công nghệ này cho phép giao diện điều khiển bằng giọng nói, phần mềm đọc chính tả và dịch vụ phiên âm theo thời gian thực.

· Tạo ngôn ngữ: Kỹ thuật tạo ngôn ngữ tự nhiên (NLG) cho phép máy tạo ra văn bản giống con người dựa trên dữ liệu đầu vào hoặc các quy tắc được xác định trước. Các ứng dụng NLG bao gồm tạo báo cáo, viết tóm tắt và soạn tin nhắn được cá nhân hóa.

· Hiểu biết theo ngữ cảnh: Giao tiếp và nhận thức hiệu quả đòi hỏi phải hiểu ngôn ngữ trong ngữ cảnh, xem xét các yếu tố như giọng điệu, ý định và sắc thái văn hóa. Hiểu biết theo ngữ cảnh là điều cần thiết cho các nhiệm vụ như phân tích tình cảm, tác nhân đàm thoại và đề xuất được cá nhân hóa.

· Nhận thức trực quan và đa phương thức: Ngôn ngữ tự nhiên có thể được tích hợp với các phương thức khác, chẳng hạn như hình ảnh, video và dữ liệu cảm biến, để nâng cao nhận thức và hiểu biết. Công nghệ đa phương thức cho phép các ứng dụng như chú thích hình ảnh, trả lời câu hỏi trực quan và giao diện thực tế tăng cường.

Nói chung, ngôn ngữ tự nhiên đóng vai trò là một công cụ mạnh mẽ để giao tiếp và nhận thức, cho phép con người và máy móc tương tác với nhau và với thế giới theo những cách phong phú hơn, trực quan hơn. Những tiến bộ trong xử lý và hiểu ngôn ngữ tự nhiên tiếp tục thúc đẩy sự đổi mới trên nhiều lĩnh vực, từ công nghệ truyền thông và trí tuệ nhân tạo đến tương tác giữa người và máy tính và khoa học nhận thức.

Pros:

· Giúp người dùng đặt câu hỏi về bất kỳ chủ đề nào và nhận được câu trả lời trực tiếp trong vòng vài giây.

· Giúp máy tính giao tiếp với con người bằng ngôn ngữ của họ.

· Nó rất hiệu quả về mặt thời gian.

· Hầu hết các công ty sử dụng NLP để nâng cao hiệu quả của quy trình tài liệu, độ chính xác của tài liệu và xác định thông tin trà từ cơ sở dữ liệu lớn.

Cons:

· NLP có thể không hiển thị được bối cảnh xung quanh.

· NLP không thể đoán trước được

· NLP có thể yêu cầu gõ phím nhiều hơn.

· NLP không thể thích ứng với miền mới và nó có chức năng hạn chế, đó là lý do tại sao NLP chỉ được xây dựng cho một nhiệm vụ cụ thể và duy nhất.

# **Tri giác (perception)**

## **Giới thiệu**

Tri giác cung cấp cho các tác tử thông tin về thế giới đang sống. Tri giác được bắt đầu bởi các cảm biến. Cảm biến là bất cứ thứ gì có thể thay đổi trạng thái tính toán của tác nhân để đáp ứng với sự thay đổi trạng thái của thế giới. Nó có thể đơn giản như một cảm biến một bit phát hiện xem công tắc đang bật hay tắt, hoặc phức tạp như võng mạc của mắt người, nơi chứa hơn một trăm triệu phần tử cảm quang.

.Có nhiều phương thức cảm giác khác nhau dành cho các tác tử nhân tạo. Những thứ họ chia sẻ với con người bao gồm thị giác, thính giác và xúc giác. Trong phần này, nhóm em sẽ tập trung vào tầm nhìn vì đây là giác quan hữu ích nhất để tiếp xúc với thế giới vật chất. Nhóm em đã rút ra được 2 điều sau khi nghiên cứu về phần tri giác này, đó là:

* Thứ nhất, chúng ta không thể nhìn xung quanh ở mọi góc nhìn, vì vậy chúng ta không thể phục hồi tất cả các khía cạnh của trạng thái thế giới hiện tại. Hơn nữa, ngay cả phần chúng ta có thể nhìn thấy cũng vô cùng mơ hồ. Một khía cạnh quan trọng của nghiên cứu về tri giác là hiểu những thông tin bổ sung nào có thể được sử dụng để giải quyết sự mơ hồ.
* Hạn chế thứ hai và có lẽ quan trọng hơn của cách tiếp cận trực tiếp là nó đang cố gắng giải quyết một vấn đề quá khó. Trong nhiều trường hợp, tác tử không cần biết mọi thứ về thế giới.

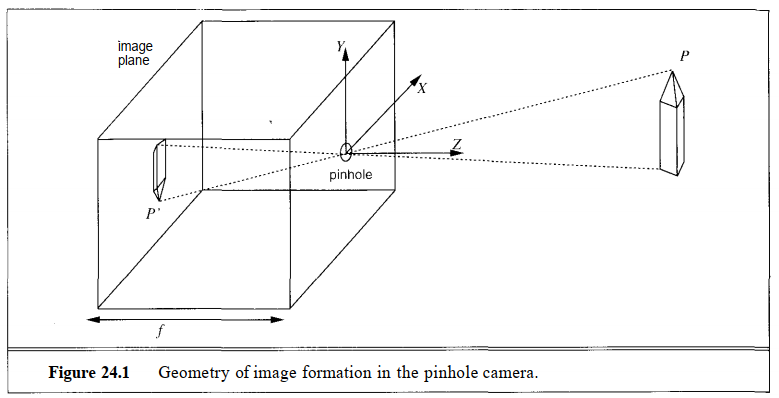
Vì vậy trong phần này, nhóm nghiên cứu quá trình hình thành hình ảnh ở mục 2. Nhóm em cũng đề cập đến cả hình học của quy trình, quy định vị trí một điểm nhất định sẽ được chụp ảnh, và phép đo trắc quang của quy trình, quy định mức độ sáng của điểm sẽ xuất hiện. Mục 3 xử lý các thao tác xử lý ảnh cơ bản thường được sử dụng trong thị giác sớm. Chúng tạo tiền đề cho quá trình phân tích sau này nhằm trích xuất thông tin cần thiết cho các nhiệm vụ như thao tác, điều hướng và nhận dạng. Trong mục 4, nhóm em sẽ nghiên cứu các tín hiệu khác nhau trong hình ảnh có thể được khai thác cho mục đích này, bao gồm chuyển động, lập thể, kết cấu, bóng đổ và đường viền. Mục 5 là thảo luận về thông tin cần thiết cho thao tác và điều hướng được hướng dẫn trực quan, còn mục 6 đề cập đến các cách tiếp cận khác nhau để nhận dạng đối tượng. Cuối cùng, mục 7 đề cập đến vấn đề tri giác trong bối cảnh nhận dạng giọng nói, từ đó giúp xác định chính xác các vấn đề nảy sinh trong tri giác nói chung.

## **Tạo ảnh**

Tầm nhìn hoạt động bằng cách thu thập ánh sáng rải rác từ các vật thể trong khung cảnh và tạo ra hình ảnh 2 chiều. Để sử dụng hình ảnh này để thu được thông tin về khung cảnh, chúng ta phải hiểu hình dạng của quá trình.

**Máy ảnh lỗ kim (Pinhole Camera)**

Cách đơn giản nhất để tạo thành một hình ảnh là sử dụng máy ảnh lỗ kim (Hình 24.1). Gọi P là một điểm trong cảnh, có tọa độ (X, Y, Z) và P' là ảnh của nó trên mặt phẳng ảnh, có tọa độ (*x, y, z*). Nếu f là khoảng cách từ lỗ kim O đến mặt phẳng ảnh thì bằng các tam giác đồng dạng, có thể rút ra các phương trình sau:



Lưu ý rằng hình ảnh bị đảo ngược, cả từ trái sang phải và từ trên xuống, so với cảnh được biểu thị trong các phương trình bằng dấu âm. Các phương trình này xác định một quá trình hình thành hình ảnh được gọi là phép chiếu phối cảnh.

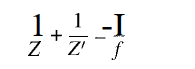
Tương tự, chúng ta có thể mô hình hóa quá trình chiếu phối cảnh với mặt phẳng chiếu ở khoảng cách f phía trước lỗ kim. Thiết bị tưởng tượng một bề mặt chiếu phía trước này lần đầu tiên được Alberti khuyên dùng cho các họa sĩ thời Phục hưng Ý vào năm 1435 như một kỹ thuật để xây dựng các mô tả chính xác về mặt hình học của khung cảnh ba chiều. Đối với mục đích của chúng tôi, ưu điểm chính của mô hình này là nó tránh được sự nghịch đảo ngang và do đó loại bỏ dấu âm trong các phương trình chiếu phối cảnh.

Dưới phép chiếu phối cảnh, các đường song song dường như hội tụ về một điểm trên đường chân trời - hãy nghĩ đến đường ray xe lửa. Chúng ta biết từ phép tính vectơ rằng một điểm P tùy ý trên đường thẳng đi qua (X0, Y0, Z0) theo hướng (U, V, W) được cho bởi, với A thay đổi giữa + ∞ và -∞ . Hình chiếu của Px trên mặt phẳng ảnh được cho bởi công thức:

Khi A = + ∞ hoặc - ∞, công thức này sẽ là Nhóm em gọi điểm điểm ảo liên kết với họ các đường thẳng có hướng (U, V, W). Nó không phụ thuộc vào điểm (X0, Y0, Z0) mà đường thẳng trong cảnh đi qua, chỉ về hướng. Nếu đối tượng tương đối nông so với khoảng cách của nó với camera, chúng ta có thể ước chừng phép chiếu phối cảnh bằng phép chiếu trực giao tỷ lệ. Ý tưởng như sau. Nếu độ sâu Z của các điểm trên đối tượng thay đổi trong phạm vi nào đó Zo ± AZ, với AZ << Zo, thì hệ số tỷ lệ phối cảnhf/Z có thể được tính gần đúng bằng hằng số . Phương trình chiếu từ tọa độ cảnh (X, Y, Z) lên mặt phẳng ảnh trở thành *x = sX* và *y = sY*. Lưu ý rằng phép chiếu trực giao theo tỷ lệ chỉ là phép tính gần đúng đối với các phần của cảnh có độ sâu bên trong không thay đổi nhiều. Nó không nên được sử dụng để nghiên cứu các thuộc tính "trong phạm vi rộng".

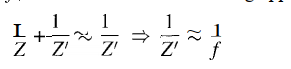
**Hệ thống ống kính**

Mắt của động vật có xương sống và máy ảnh thật đều sử dụng ống kính. Thấu kính rộng hơn nhiều so với lỗ kim, cho phép nó thu được nhiều ánh sáng hơn. Điều này được bù đắp bởi thực tế là không phải tất cả các cảnh đều có thể được lấy nét sắc nét cùng một lúc. Ảnh của một vật ở khoảng cách Z trong cảnh được tạo ra ở một khoảng cách cố định so với thấu kính Z', trong đó mối quan hệ giữa Z và Z' được cho bởi phương trình thấu kính sau:



trong đó f là tiêu cự của thấu kính. Với một lựa chọn nhất định về khoảng cách ảnh Zo giữa điểm nút của thấu kính và mặt phẳng ảnh, các điểm cảnh có độ sâu trong phạm vi xung quanh Zo, trong đó Zo là khoảng cách đối tượng tương ứng, sẽ được chụp ảnh với tiêu cự sắc nét hợp lý. Độ sâu trong ngữ cảnh này được gọi là độ sâu trường ảnh.

Vì khoảng cách đối tượng Z thường lớn hơn nhiều so với khoảng cách ảnh Z' hoặc f nên chúng ta thường thực hiện phép tính gần đúng sau:



Do đó, khoảng cách ảnh Z' ≈f. Do đó, chúng ta có thể tiếp tục sử dụng các phương trình chiếu phối cảnh camera lỗ kim để mô tả dạng hình học của sự hình thành ảnh trong hệ thấu kính.

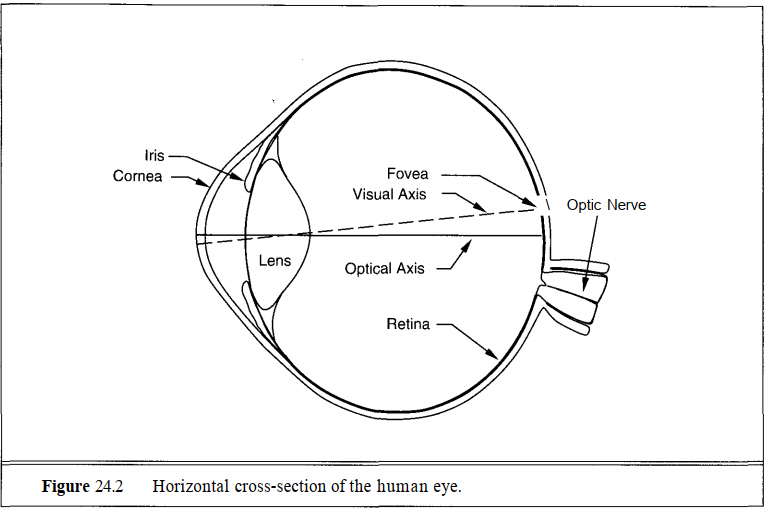
Để lấy nét các vật thể ở các khoảng cách Z khác nhau, thấu kính trong mắt (xem Hình 24.2) thay đổi hình dạng, trong khi thấu kính trong máy ảnh di chuyển theo hướng Z.

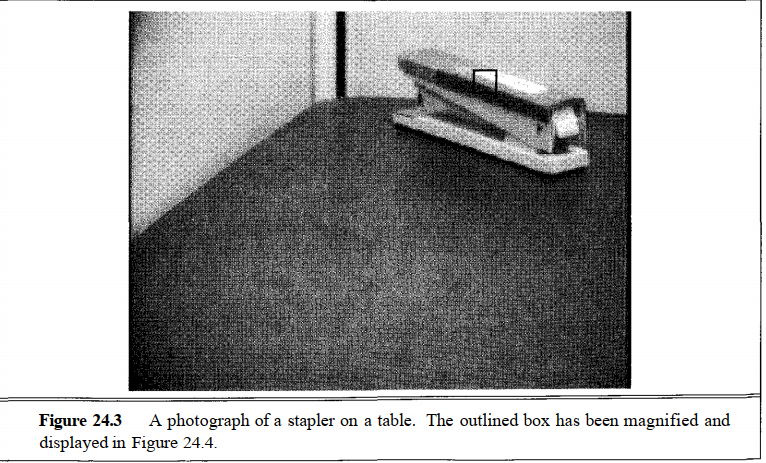
Mặt phẳng hình ảnh được phủ bằng vật liệu cảm quang:

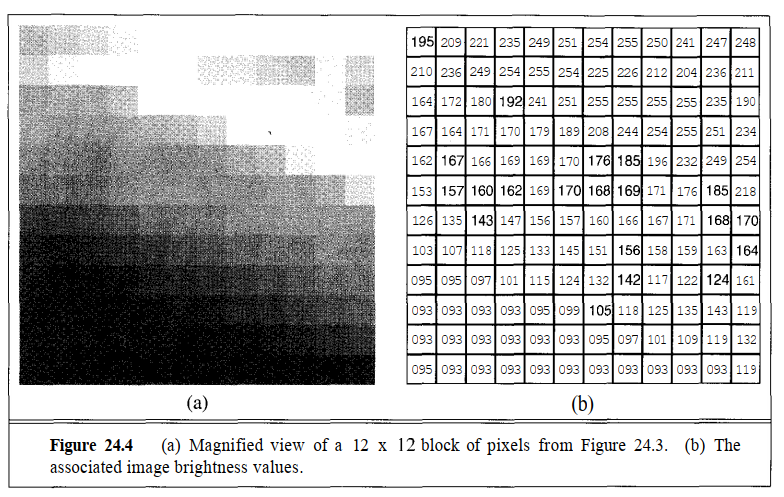
* Bạc halogenua trên phim ảnh.
* Rhodopsin và các biến thể ở võng mạc.
* Mạch silicon trong camera CCD (thiết bị ghép điện tích). Mỗi vị trí trong CCD tích hợp các electron được giải phóng do sự hấp thụ photon trong một khoảng thời gian cố định.

Trong mắt và camera CCD, mặt phẳng hình ảnh được chia thành các **pixel**: thường là 512×512 (0,25 x 106) trong camera CCD, được sắp xếp theo lưới hình chữ nhật; 120 x 106 que và 6 x 106 hình nón trong mắt, sắp xếp theo kiểu khảm lục giác.

Trong cả hai trường hợp, chúng ta có thể mô hình hóa tín hiệu được phát hiện trong mặt phẳng ảnh bằng sự thay đổi độ sáng của ảnh theo thời gian: *I(x,y,t)*. Hình 24.3 hiển thị hình ảnh số hóa của chiếc máy dập ghim trên bàn và Hình 4 hiển thị một mảng các giá trị độ sáng của hình ảnh được liên kết với khối pixel 12 x 12 được trích xuất từ hình ảnh chiếc máy dập ghim. Một chương trình máy tính cố gắng diễn giải hình ảnh sẽ phải bắt đầu từ cách biểu diễn như vậy.







**Phép đo quang học**

Độ sáng của pixelp trong ảnh tỷ lệ thuận với lượng ánh sáng hướng về phía máy ảnh bởi miếng vá bề mặt S, chiếu tới pixel p. Điều này lại phụ thuộc vào đặc tính phản xạ của Sp, vị trí và sự phân bố của nguồn sáng. Ngoài ra còn có sự phụ thuộc vào đặc tính phản xạ của phần còn lại của khung cảnh vì các bề mặt khung cảnh khác có thể đóng vai trò là nguồn sáng gián tiếp bằng cách phản chiếu ánh sáng mà chúng nhận được lên Sp.

Ánh sáng phản xạ từ một vật thể được đặc trưng là phản xạ khuếch tán hoặc phản xạ gương. Ánh sáng phản xạ khuếch tán là ánh sáng xuyên qua bề mặt của vật thể, bị hấp thụ và sau đó phát ra lại. Bề mặt có vẻ sáng như nhau đối với người quan sát ở bất kỳ hướng nào. Định luật cosin Lambert được sử dụng để mô tả sự phản xạ ánh sáng từ bề mặt khuếch tán hoàn hảo hoặc bề mặt **Lambertian**. Cường độ E của ánh sáng phản xạ từ một bộ khuếch tán hoàn hảo được cho bởi



trong đó Eo là cường độ của nguồn sáng; p là suất phản chiếu, thay đổi từ 0 (đối với bề mặt hoàn toàn đen) đến 1 (đối với bề mặt trắng tinh); và 0 là góc giữa hướng ánh sáng và bề mặt bình thường.

Ánh sáng phản xạ đặc biệt được phản xạ từ bề mặt bên ngoài của vật thể. Ở đây, năng lượng của ánh sáng phản xạ tập trung chủ yếu theo một hướng cụ thể - hướng mà tia phản xạ nằm trong cùng mặt phẳng với tia tới và thỏa mãn điều kiện góc phản xạ bằng góc tới. Đây là hành vi của một tấm gương hoàn hảo.

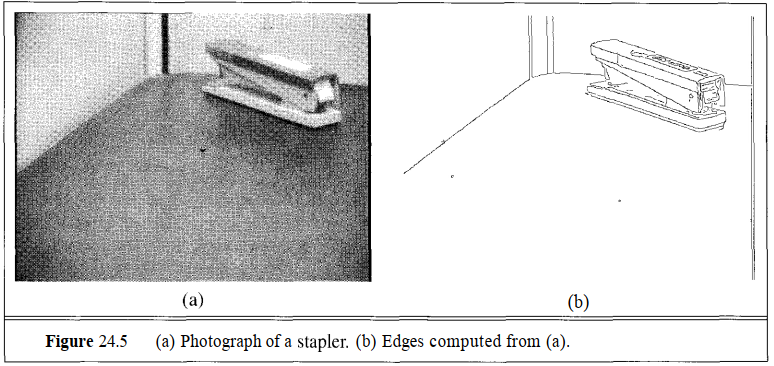
Trong cuộc sống thực, các bề mặt thể hiện sự kết hợp của các đặc tính khuếch tán và đặc tính phản chiếu. Mô hình hóa điều này trên máy tính là công việc chính của đồ họa máy tính. Việc hiển thị hình ảnh thực tế thường được thực hiện bằng phương pháp dò tia, nhằm mục đích mô phỏng quá trình vật lý của ánh sáng bắt nguồn từ các nguồn sáng và bị phản xạ đi phản xạ lại nhiều lần. Bài toán tạo bóng từ hình dạng trong thị giác máy tính nhằm mục đích đảo ngược quá trình, tức là bắt đầu từ hình ảnh "được kết xuất" và tìm ra bố cục của ảnh ba chiều đã tạo ra nó. Chúng em sẽ nói về điều này sâu hơn ở mục 4.

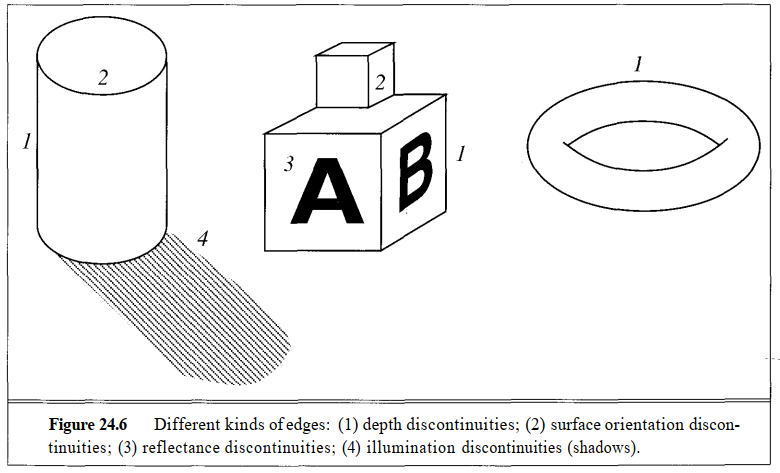
**Quang phổ của sự hình thành hình ảnh**

Về cường độ ảnh I(x,y,t), vui vẻ bỏ qua thực tế là ánh sáng khả kiến có nhiều bước sóng khác nhau - từ 400 nm ở đầu tím của quang phổ đến 700 nm ở đầu đỏ. Cho rằng có sự liên tục của các bước sóng, việc chúng ta có ba màu cơ bản có nghĩa là gì? Lời giải thích là màu sắc hoàn toàn theo đúng nghĩa đen trong mắt người nhìn. Có ba loại hình nón khác nhau trong mắt với ba đường cong độ nhạy quang phổ khác nhau Rk(). Đầu ra của hình nón thứ k tại vị trí (x, y) tại thời điểm t khi đó là Ik(x, y, t) =. Không gian bước sóng vô hạn được chiếu thành không gian màu ba chiều. Điều này có nghĩa là chúng ta nên coi như một vectơ ba chiều tại (x,y,t). Bởi vì mắt ánh xạ nhiều phổ tần số khác nhau vào cùng một cảm giác màu sắc, nên chúng ta có thể hy vọng rằng có tồn tại các quang phổ ánh sáng khác nhau của các chất chuyển hóa có vẻ giống nhau đối với con người.

## **Hoạt động xử lý hình ảnh**

Hình 24.5 hiển thị hình ảnh của một cảnh có chứa một cái dập ghim nằm trên bàn cũng như các cạnh được phát hiện trên hình ảnh này. Các cạnh là các đường cong trong mặt phẳng ảnh mà qua đó có sự thay đổi "đáng kể" về độ sáng của ảnh. Mục tiêu cuối cùng của việc phát hiện cạnh là xây dựng một đường vẽ lý tưởng như Hình 24.6. Động lực là các đường viền cạnh trong ảnh tương ứng với các đường viền cảnh quan trọng. Trong ví dụ này, nhóm em nhận thấy có các điểm gián đoạn về độ sâu, được gắn nhãn 1; sự gián đoạn định hướng bề mặt, được dán nhãn 2; sự gián đoạn phản xạ, được dán nhãn 3; và sự gián đoạn chiếu sáng (bóng), được dán nhãn 4.

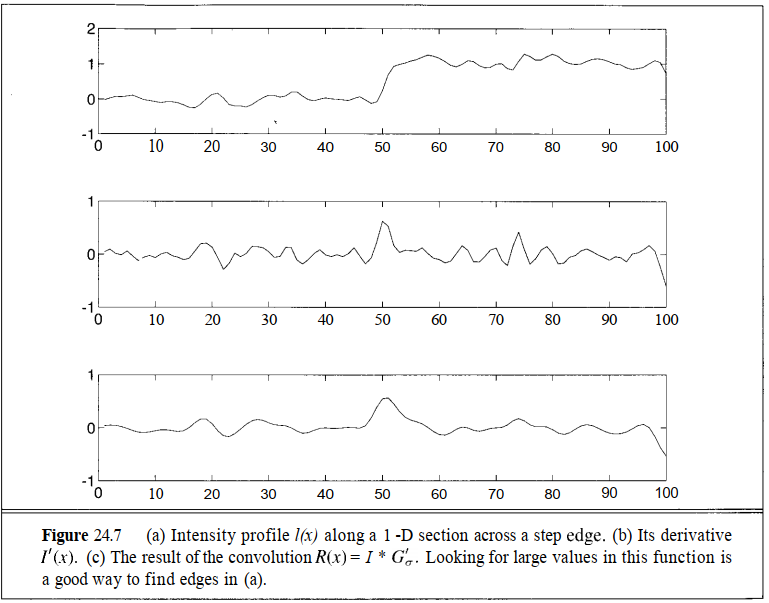




Có sự khác biệt lớn giữa đầu ra của bộ dò cạnh như trong Hình 24.5(b) và một bản vẽ đường lý tưởng. Thông thường, có những đường viền bị thiếu (chẳng hạn như cạnh trên của kim bấm), cũng như các đường viền nhiễu không tương ứng với bất kỳ điều gì có ý nghĩa trong cảnh. Các giai đoạn xử lý sau dựa trên các cạnh phải tính đến các lỗi này.

Để phát hiện các cạnh trong một hình ảnh, nhóm em đã xem xét cấu hình độ sáng của hình ảnh dọc theo mặt cắt 1-D vuông góc với một cạnh, ví dụ, mặt cắt giữa cạnh trái của bàn và tường. Nó trông giống như những gì được hiển thị trong Hình 24.7 (a). Vị trí của cạnh tương ứng với x = 50.

Bởi vì các cạnh tương ứng với các vị trí trong ảnh mà độ sáng trải qua sự thay đổi rõ rệt, nên một ý tưởng ngây thơ sẽ là vi phân ảnh và tìm những vị trí có độ lớn của đạo hàm I'(x) lớn. Vâng, điều đó gần như có tác dụng. Trong Hình 24.7(b) chúng ta thấy rằng mặc dù có một đỉnh ở x = 50, nhưng cũng có những đỉnh phụ ở các vị trí khác (ví dụ: x = 75) có thể bị nhầm lẫn là các cạnh thực. Những điều này phát sinh do sự hiện diện của nhiễu trong hình ảnh. Chúng ta nhận được kết quả tốt hơn nhiều bằng cách kết hợp phép toán vi phân với **phương pháp làm mịn**.



Kết quả là Hình 24.7(c), trong đó đỉnh trung tâm tại x = 50 vẫn còn và các đỉnh phụ bị giảm đi nhiều. Điều này cho phép nhóm em tìm thấy các cạnh mà không bị nhầm lẫn bởi tiếng ồn.

Để hiểu rõ hơn những ý tưởng này, ta cần khái niệm toán học về tích chập. Nhiều thao tác xử lý hình ảnh hữu ích như làm mịn và phân biệt có thể được thực hiện bằng cách cuộn hình ảnh với các chức năng phù hợp.

**Tích phân chập với các bộ lọc tuyến tính**

Kết quả của việc tích phân chập hai hàm f và g là hàm h mới, ký hiệu là h g, được xác định bởi 2 phương trình



Thông thường, các hàm f và g mà chúng ta làm việc chỉ nhận các giá trị khác 0 trong một khoảng hữu hạn nào đó, do đó phép tính tổng có thể dễ dàng thực hiện trên máy tính.

Việc khái quát hóa các hàm được xác định trên hai chiều (chẳng hạn như hình ảnh) là đơn giản. Chúng tôi thay thế tích phân 1 -D (hoặc tổng) bằng tích phân 2-D (hoặc tổng). Kết quả của việc tích chập hai hàm f và g là hàm h mới, ký hiệu là h = f\* g, được xác định bởi

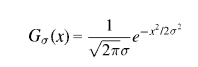


hoặc nếu chúng ta lấy miền xác định của hai hàm là rời rạc



**Phát hiện các cạnh**

Chúng ta hãy quay lại ví dụ 1-D trong Hình 24.7. Chúng tôi muốn đưa ra khái niệm làm mịn hình ảnh chính xác hơn. Một dạng làm mịn tiêu chuẩn là làm xoắn ảnh bằng hàm Gaussian



Bây giờ có thể chỉ ra rằng với mọi hàm f và g, f\* g' = (f\* g)', do đó làm mịn ảnh bằng cách tích chập với một Gaussian G, và sau đó vi phân tương đương với việc tích chập ảnh với G'(x ), đạo hàm bậc nhất của Gaussian:



Vì vậy, chúng ta có một thuật toán đơn giản để phát hiện cạnh 1 -D:

1. Xoay ảnh / với G' để thu được R.

2. Tìm giá trị tuyệt đối của R.

3. Đánh dấu các đỉnh đó trong ||R|| cao hơn một số ngưỡng T được xác định trước. Ngưỡng này được chọn để loại bỏ các đỉnh giả do nhiễu.

Trong hai chiều, chúng ta cần phải đối mặt với thực tế là cạnh có thể ở bất kỳ góc nào bằng 0. Để phát hiện các cạnh thẳng đứng, chúng ta có một chiến lược rõ ràng: tích chập với G'(x)G(y). Theo hướng y, hiệu ứng chỉ là làm mịn (do tích chập Gaussian) và theo hướng x, hiệu ứng là hiệu ứng vi phân đi kèm với làm mịn. Thuật toán phát hiện các cạnh dọc như sau:

1. Xoay ảnh I(x, y) với fv(x, y) G(x)G, (y để thu được Ry(x, y).

2. Tìm giá trị tuyệt đối của RV(X, y).

3. Đánh dấu các đỉnh trong ||Rv|\(x,y) nằm trên ngưỡng T được chỉ định trước.

Để phát hiện một cạnh theo hướng tùy ý, chúng ta cần tích hợp hình ảnh với hai bộ lọc fv = G(x)G, (y) và f₁ = G(y)G(x), chỉ được xoay fy một góc 90° . Thuật toán phát hiện các cạnh ở các hướng tùy ý là

1. Xoay ảnh I(x, y) với fv(x, y) và f(xy) để thu được Ry(x,y) và RH(x,y), tương ứng. Xác định R(x, y) = R(x, y) + R(x, y)

2. Tìm giá trị tuyệt đối của R(x,y).

3. Đánh dấu các đỉnh trong \\R\ (x, y) nằm trên ngưỡng Tn được chỉ định trước.

Khi chúng ta đã đánh dấu các pixel cạnh bằng thuật toán này, giai đoạn tiếp theo là liên kết các pixel đó thuộc cùng một đường cong cạnh. Điều này có thể được thực hiện bằng cách giả sử rằng bất kỳ hai pixel lân cận nào đều là pixel cạnh có hướng nhất quán phải thuộc cùng một đường cong cạnh. Chúng tôi vừa phác thảo quy trình cơ bản được sử dụng trong bộ dò biên Canny (Canny, 1986), đây là một thuật toán tiêu chuẩn được sử dụng rộng rãi để phát hiện các cạnh trong ảnh.

## **Thu nhập thông tin 3 chiều sử dụng hình ảnh hoặc dữ liệu**

1. Phân chia cảnh thành các đối tượng riêng biệt.

2. Xác định vị trí và hướng của mỗi đối tượng so với người quan sát.

3. Xác định hình dạng của mỗi đối tượng.

Phân đoạn hình ảnh là một bước quan trọng hướng tới việc tổ chức mảng pixel hình ảnh thành các vùng tương ứng với các thực thể có ý nghĩa về mặt ngữ nghĩa trong cảnh. Để nhận biết, chúng tôi muốn biết những đặc điểm nào thuộc về nhau để có thể so sánh chúng với các mô hình được lưu trữ; để nắm bắt một đối tượng, người ta cần biết những gì thuộc về một đối tượng

**Chuyển động**

Nếu máy ảnh di chuyển so với cảnh ba chiều thì chuyển động biểu kiến ​​tạo ra trong ảnh được gọi là luồng quang học. Điều này mô tả hướng và tốc độ chuyển động của các đối tượng trong ảnh do chuyển động tương đối giữa người xem và cảnh.

**Binocular stereopsis**: lập thể 2 mắt

Lập thể hai mắt Ý tưởng ở đây khá giống với thị sai chuyển động, ngoại trừ việc thay vì sử dụng hình ảnh theo thời gian, chúng tôi sử dụng hai (hoặc nhiều) hình ảnh tách biệt trong không gian, chẳng hạn như được cung cấp bởi mắt hướng về phía trước của con người. Bởi vì một đặc điểm nhất định trong cảnh sẽ ở một vị trí khác so với trục z của mỗi mặt phẳng hình ảnh, nên nếu chúng ta đặt chồng hai hình ảnh lên nhau thì sẽ có sự chênh lệch về vị trí của đặc điểm đó.

**Độ dốc của kết cấu**

Kết cấu trong ngôn ngữ hàng ngày là một đặc tính của các bề mặt liên quan đến chất lượng xúc giác mà chúng gợi ý (kết cấu có cùng gốc với dệt). Trong tầm nhìn tính toán, nó đề cập đến một khái niệm có liên quan chặt chẽ, đó là mô hình lặp lại trong không gian trên một bề mặt có thể cảm nhận được bằng mắt. Ví dụ Phần 24.4. Trích xuất thông tin 3-D bằng Vision 743 bao gồm mô hình cửa sổ trên tòa nhà, đường khâu trên áo len, đốm trên da báo, ngọn cỏ trên bãi cỏ, sỏi trên bãi biển hoặc đám đông người trong sân vận động. Đôi khi sự sắp xếp khá tuần hoàn, chẳng hạn như các mũi khâu trên áo len; trong những trường hợp khác, như đối với những viên sỏi trên bãi biển, tính đều đặn chỉ mang tính thống kê - mật độ của những viên sỏi gần như giống nhau ở những phần khác nhau của bãi biển.

**Bóng**

Bóng đổ—sự thay đổi về cường độ ánh sáng nhận được từ các phần khác nhau của bề mặt trong khung cảnh—được xác định bởi hình dạng khung cảnh và đặc tính phản xạ của các bề mặt. Trong đồ họa máy tính, mục tiêu là xác định độ sáng hình ảnh I(x,y) dựa trên các đặc tính hình học và phản xạ của khung cảnh. Trong thị giác máy tính, hy vọng của chúng ta có thể là đảo ngược quá trình, nghĩa là khôi phục các đặc tính hình học và phản xạ của cảnh với độ sáng hình ảnh I(x, y). Điều này tỏ ra khó thực hiện trong mọi trường hợp ngoại trừ trường hợp đơn giản nhất

**Đường viền**

Đường viền Khi chúng ta nhìn vào một bản vẽ đường, chẳng hạn như Hình 24.19, chúng ta có được nhận thức sống động về hình dạng và bố cục 3-D. Làm sao? Rốt cuộc, chúng ta đã thấy trước đó rằng có vô số cấu hình cảnh có thể dẫn đến cùng một đường vẽ. Lưu ý rằng chúng ta thậm chí còn nhận được cảm nhận về độ nghiêng và độ nghiêng bề mặt. Nó có thể là do sự kết hợp của kiến ​​thức cấp cao về các hình dạng điển hình cũng như một số hạn chế ở cấp độ thấp. Chúng ta sẽ xem xét kiến ​​thức định tính có sẵn từ một bản vẽ đường.

## **Sử dụng tầm nhìn để điều khiển và điều hướng**

Một trong những công dụng chính của thị giác là cung cấp thông tin để điều khiển các vật thể - nhặt, nắm, xoay, v.v. - cũng như điều hướng trong một cảnh trong khi tránh chướng ngại vật.

Khả năng sử dụng thị giác cho những mục đích này hiện diện trong hệ thống thị giác nguyên thủy nhất của động vật. Có lẽ nguồn gốc tiến hóa của giác quan thị giác có thể bắt nguồn từ sự hiện diện của một điểm cảm quang ở một đầu của sinh vật giúp nó tự định hướng về phía (hoặc tránh xa) ánh sáng. Ruồi sử dụng tầm nhìn dựa trên luồng quang học để kiểm soát phản ứng hạ cánh của chúng. Robot di động di chuyển trong môi trường cần biết chướng ngại vật ở đâu, có hành lang không gian trống ở đâu, v.v.

Nhóm em đã xem xét nhiệm vụ của người điều khiển ô tô ở góc dưới bên trái của hình 24.24.

1. Tiếp tục di chuyển với tốc độ vừa phải.

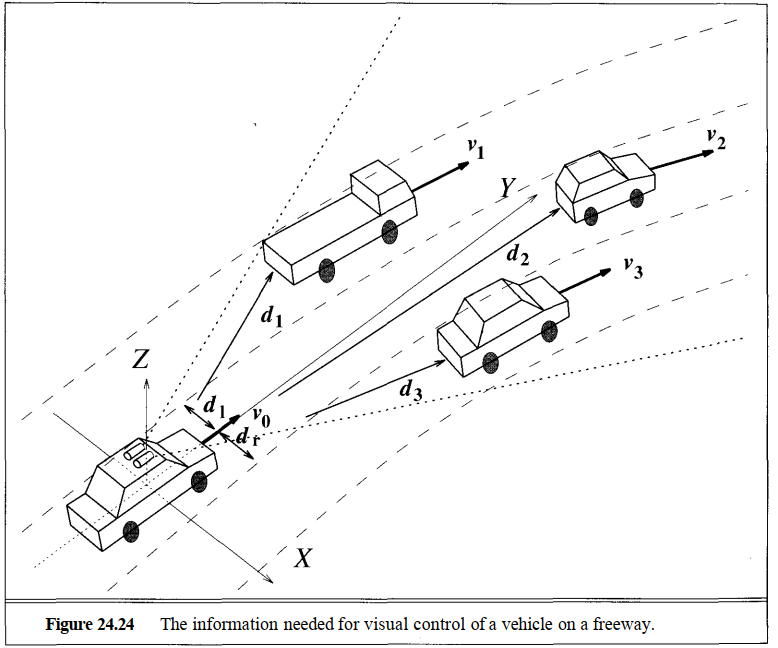
2. Kiểm soát bên - đảm bảo rằng phương tiện vẫn an toàn trong làn đường của mình, nghĩa là giữ d₁ = dr.

3. Kiểm soát dọc đảm bảo khoảng cách an toàn d2 giữa xe đó và xe phía trước.

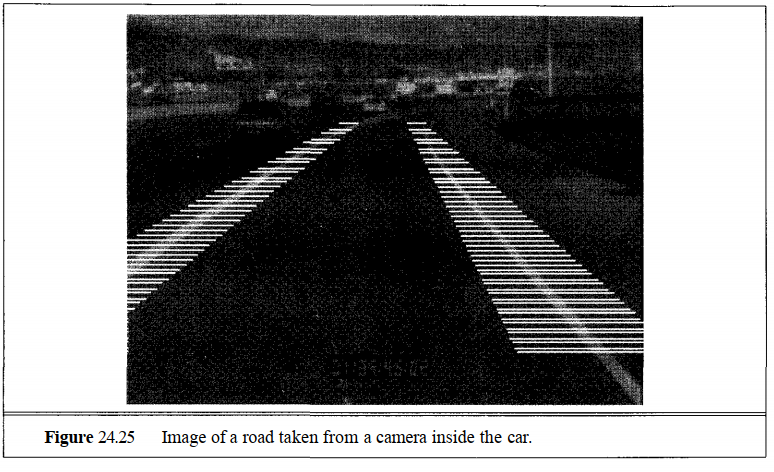
4. Giám sát các phương tiện ở các làn đường lân cận (ở khoảng cách d₁ và d3) và chuẩn bị các biện pháp di chuyển phù hợp nếu một trong số các phương tiện đó quyết định chuyển làn.

Vấn đề đối với người lái xe là tạo ra các hành động lái, dẫn động hoặc phanh thích hợp để hoàn thành tốt nhất các nhiệm vụ này. Để điều khiển 2 bên, cần duy trì sự thể hiện vị trí và hướng của ô tô so với làn đường. Hình ảnh đường đi từ camera gắn trên ô tô được thể hiện trong hình 24.25. Nhóm em phát hiện các cạnh tương ứng với các đoạn đánh dấu làn đường và sau đó điều chỉnh các đường cong mượt mà cho các đoạn này. Các thông số của những đường cong này mang thông tin về vị trí ngang của ô tô, hướng mà ô tô hướng tới so với làn đường và độ cong của làn đường. Thông tin này cùng với động lực học của ô tô là tất cả những gì hệ thống điều khiển lái cần. Cần lưu ý rằng vì từ mỗi khung hình này sang khung hình tiếp theo chỉ có một thay đổi nhỏ về vị trí hình chiếu của làn đường trong ảnh nên người ta biết nơi cần tìm các điểm đánh dấu làn đường trong hình ảnh - trong hình, điều này được chỉ ra bằng cách hiển thị các cửa sổ tìm kiếm.

Để điều khiển theo chiều dọc, người ta cần biết khoảng cách với các xe phía trước. Điều này có thể được thực hiện bằng cách sử dụng phương pháp lập thể hai mắt hoặc dòng quang học. Cả hai cách tiếp cận này đều có thể được đơn giản hóa bằng cách khai thác các ràng buộc về miền xuất phát từ thực tế là người ta đang lái xe trên một bề mặt phẳng. Sử dụng những kỹ thuật này, Dickmanns và Zapp (1987) đã chứng minh việc lái ô tô được điều khiển bằng trực quan trên đường cao tốc ở tốc độ cao. Pomerleau (1993) đã đạt được hiệu suất tương tự bằng cách sử dụng phương pháp tiếp cận mạng lưới thần kinh.



Ví dụ lái xe cho thấy rất rõ một điểm: đối với một tác vụ cụ thể, người ta không cần khôi phục tất cả thông tin mà về nguyên tắc có thể khôi phục được từ hình ảnh. Ta không cần phải khôi phục hình dạng chính xác của mọi phương tiện, giải quyết hình dạng từ kết cấu trên bề mặt cỏ cạnh xa lộ, v.v. Nhu cầu của nhiệm vụ chỉ yêu cầu một số loại thông tin nhất định và người ta có thể đạt được tốc độ tính toán đáng kể và độ bền bằng cách chỉ khôi phục thông tin đó và khai thác triệt để các ràng buộc về miền. Mục đích của nhóm em khi thảo luận về các phương pháp tiếp cận chung trong phần trước là chúng hình thành nên lý thuyết cơ bản mà có thể chuyên môn hóa cho nhu cầu của các nhiệm vụ cụ thể.



## **Hiển thị và nhận diện đối tượng**

**Phương pháp căn chỉnh**

Chúng ta sẽ xem xét một dạng cụ thể của bài toán một cách chi tiết hơn - chúng ta được yêu cầu nhận dạng một vật thể ba chiều từ hình chiếu của nó lên mặt phẳng ảnh. Để thuận tiện, quá trình chiếu được mô hình hóa dưới dạng phép chiếu trực giao tỷ lệ. Chúng ta không biết tư thế của vật thể - vị trí và hướng của nó đối với máy ảnh.

Đối tượng được biểu diễn bằng một tập hợp m đặc điểm hoặc các điểm phân biệt 41, 42, µm trong không gian ba chiều - có thể là các đỉnh đối với một đối tượng đa diện. Chúng được đo trong một số hệ tọa độ tự nhiên của đối tượng. Sau đó, các điểm phải chịu một phép quay 3-D R không xác định, sau đó được dịch theo lượng t và phép chiếu không xác định để tạo ra các điểm đặc trưng của ảnh p1, p2,..., pn trên mặt phẳng ảnh. Nói chung, n≠m vì một số điểm mô hình có thể bị che khuất. Bộ phát hiện đặc điểm trong ảnh cũng có thể bỏ sót các đặc điểm thực và đánh dấu các đặc điểm sai do nhiễu. Chúng ta có thể diễn đạt điều này theo phương trình sau:

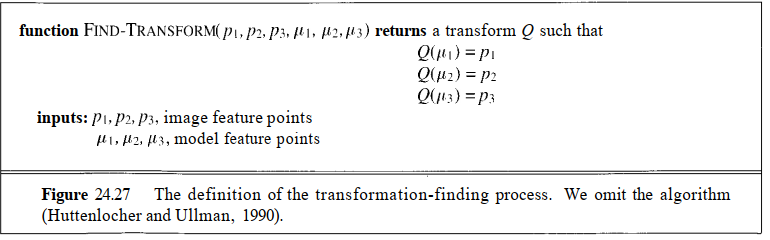


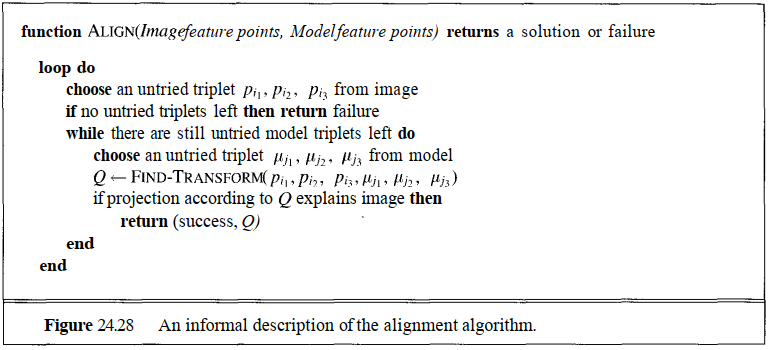
đối với điểm mô hình 3-D; và điểm ảnh tương ứng p₁. Ở đây nhóm em biểu thị phép chiếu phối cảnh hoặc một trong những phép tính gần đúng của nó, chẳng hạn như phép chiếu chính tả tỷ lệ. Ta có thể tóm tắt điều này bằng phương trình p₁ = Qu; trong đó Q là phép biến đổi (chưa xác định) đưa các điểm mô hình thẳng hàng với hình ảnh. Giả sử đối tượng là cứng nhắc, phép biến đổi Q là như nhau đối với tất cả các điểm mô hình.

Taa có thể giải Q với tọa độ 3-D của ba điểm mô hình và hình chiếu 2-D của chúng. Trực giác như sau: người ta có thể viết ra các phương trình liên hệ tọa độ của Pi với tọa độ của ui. Trong các phương trình này, các đại lượng chưa biết tương ứng với các tham số của ma trận quay R và vectơ tịnh tiến t. Nếu chúng ta có đủ nhiều phương trình, chúng ta có thể giải được Q. Chúng ta sẽ không đưa ra bất kỳ chứng minh nào ở đây mà chỉ nêu kết quả sau (Huttenlocher và Ullman, 1990):

Cho ba điểm không thẳng hàng μι, μ2 và 13 trong mô hình và các hình chiếu của chúng trên mặt phẳng ảnh, pl, p2 và p3 dưới phép chiếu trực giao tỷ lệ, tồn tại chính xác hai phép biến đổi từ hệ tọa độ mô hình ba chiều sang hệ tọa độ hai- khung tọa độ hình ảnh chiều.

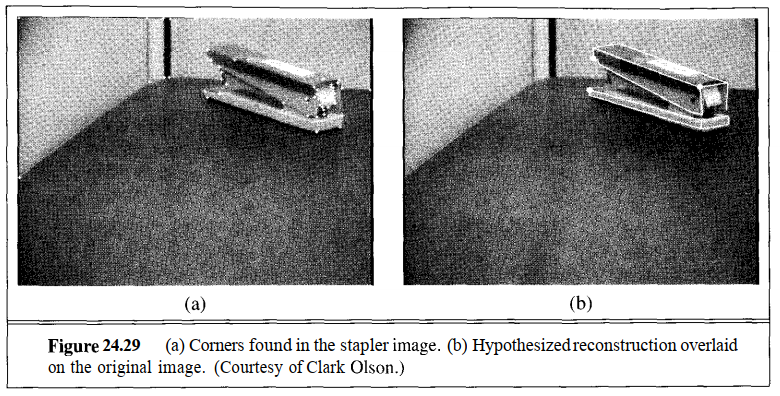
Những phép biến đổi này có liên quan bởi sự phản chiếu xung quanh mặt phẳng ảnh và có thể được tính toán bằng một giải pháp dạng đóng đơn giản. Chúng ta sẽ chỉ giả sử rằng tồn tại một hàm FIND-TRANSFORM, như trong Hình 24.27.





Đây là cơ sở của thuật toán ALIGN, thuật toán tìm cách tìm tư thế cho một mô hình nhất định và trả về lỗi nếu ngược lại (xem Hình 24.28). Độ phức tạp về thời gian trong trường hợp xấu nhất của thuật toán tỷ lệ thuận với số lượng kết hợp của bộ ba mô hình và bộ ba hình ảnh - điều này cho biết số lần Q phải được tính - nhân với chi phí xác minh. Điều này mang lại (2) (") gấp lần chi phí xác minh. Chi phí xác minh là m logn, vì chúng ta phải dự đoán vị trí hình ảnh của từng điểm trong số m điểm mô hình và tìm khoảng cách đến điểm hình ảnh gần nhất, thao tác log n nếu các điểm hình ảnh được sắp xếp theo cấu trúc dữ liệu phù hợp. Do đó, độ phức tạp trong trường hợp xấu nhất của thuật toán căn chỉnh là O(m+n³log n), trong đó m và n lần lượt là số lượng điểm mô hình và điểm hình ảnh.

Ta có thể giảm độ phức tạp về thời gian bằng một số ý tưởng. Một kỹ thuật đơn giản là chỉ đưa ra giả thuyết về sự trùng khớp giữa các cặp điểm hình ảnh và mô hình. Cho hai điểm ảnh và các cạnh tại các điểm này, điểm ảo thứ ba có thể được xây dựng bằng cách mở rộng các cạnh và tìm giao điểm. Điều này làm giảm độ phức tạp xuống O(m³n²logn). Các kỹ thuật dựa trên việc sử dụng phân cụm tư thế kết hợp với ngẫu nhiên hóa (Olson, 1994) có thể được sử dụng để giảm độ phức tạp xuống O(mn³). Kết quả từ việc áp dụng thuật toán này vào hình ảnh máy dập ghim được thể hiện trong Hình 24.29.

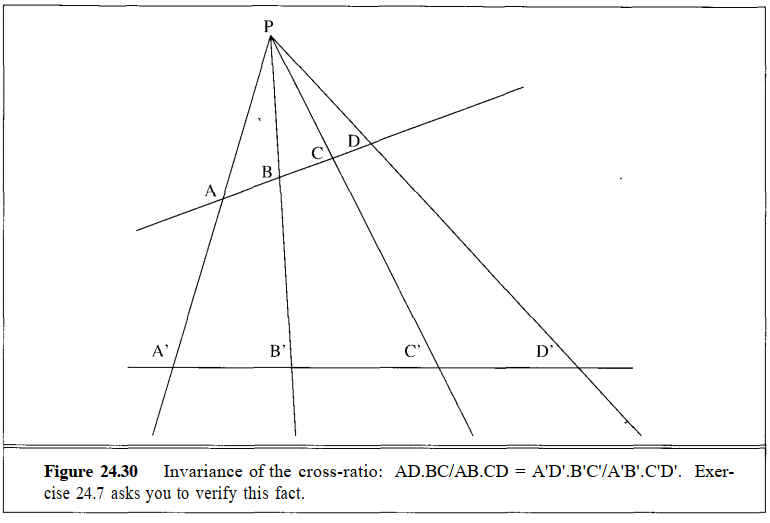


**Sử dụng bất biến xạ ảnh**

Việc căn chỉnh bằng cách sử dụng hình học đường viền và nhận dạng được coi là thành công nếu hình học đường viền trong ảnh có thể được giải thích dưới dạng hình chiếu phối cảnh của mô hình hình học của đối tượng. Một bất lợi là điều này liên quan đến việc thử từng mô hình trong thư viện mô hình, dẫn đến độ phức tạp nhận dạng tỷ lệ thuận với số lượng mô hình trong thư viện.

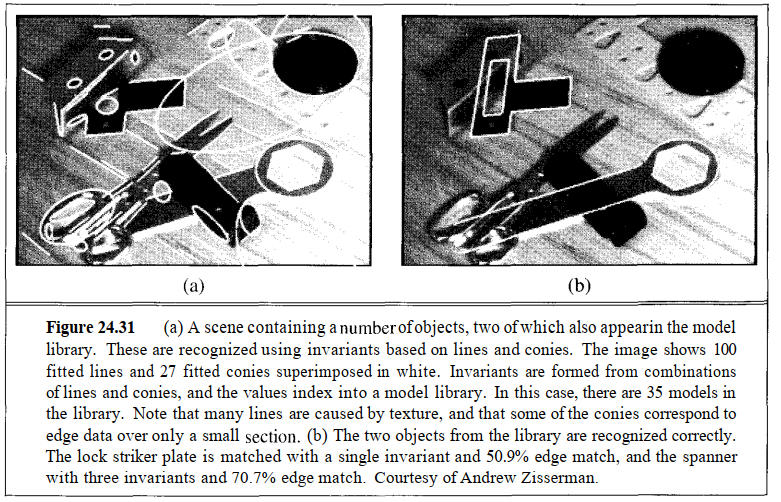
Một giải pháp được cung cấp bằng cách sử dụng các **bất biến hình học** làm biểu diễn hình dạng. Các bộ mô tả hình dạng này là bất biến quan điểm, nghĩa là chúng có cùng giá trị được đo trên đối tượng hoặc được đo từ hình ảnh phối cảnh của đối tượng và không bị ảnh hưởng bởi tư thế đối tượng. Ví dụ đơn giản nhất của bất biến xạ ảnh là “tỷ lệ chéo” của bốn điểm trên một đường thẳng, được minh họa trong Hình 24.30. Trong phép chiếu phối cảnh, tỷ lệ khoảng cách không được bảo toàn—hãy nghĩ đến khoảng cách của tà vẹt trên hình ảnh đường ray đang lùi dần. Khoảng cách không đổi trên thế giới nhưng giảm dần theo khoảng cách từ máy ảnh trong ảnh. Tuy nhiên, tỷ lệ tỷ lệ khoảng cách trên một đường thẳng được bảo toàn, nghĩa là nó giống nhau được đo trên vật thể hoặc trong ảnh.

Các bất biến có ý nghĩa quan trọng trong tầm nhìn vì chúng có thể được sử dụng làm hàm chỉ mục, do đó giá trị được đo trong hình ảnh sẽ lập chỉ mục trực tiếp cho mô hình trong thư viện. Lấy một ví dụ đơn giản, giả sử có ba mô hình (A, B, C} trong thư viện, mỗi mô hình có một giá trị bất biến tương ứng và riêng biệt {I(A), I(B), I(C)}. sau: Sau khi phát hiện và nhóm cạnh, các bất biến được đo từ các đường cong của ảnh. Nếu giá trị / = I(B) được đo thì có bằng chứng cho thấy đối tượng B có mặt. Không cần thiết phải xem xét thêm các đối tượng A và Cany. có thể là đối với một cơ sở mô hình lớn, tất cả các bất biến đều không khác biệt (nghĩa là một số mô hình có thể chia sẻ các giá trị bất biến). Do đó, khi một bất biến được đo trong ảnh tương ứng với một giá trị trong thư viện, thì một giả thuyết nhận dạng sẽ được tạo ra. tương ứng với cùng một đối tượng sẽ được hợp nhất nếu tương thích.



Một ưu điểm khác của việc biểu diễn hình dạng bất biến là các mô hình có thể thu được trực tiếp từ hình ảnh. Không cần thiết phải thực hiện các phép đo trên đối tượng thực tế vì các bộ mô tả hình dạng có cùng giá trị khi được đo trong bất kỳ hình ảnh nào. Điều này giúp đơn giản hóa và tạo điều kiện tự động hóa việc thu thập mô hình. Nó đặc biệt hữu ích trong các ứng dụng như nhận dạng từ hình ảnh vệ tinh.

Mặc dù hai cách tiếp cận nhận dạng đối tượng mà chúng tôi đã mô tả đều hữu ích trong thực tế, nhưng cần lưu ý rằng chúng ta còn rất xa so với năng lực của con người. Việc tạo ra các biểu diễn mang tính mô tả và phong phú từ hình ảnh, phân đoạn và nhóm để xác định các đặc điểm thuộc về nhau và việc kết hợp chúng với các mô hình đối tượng là những vấn đề nghiên cứu khó khăn khi được điều tra tích cực.

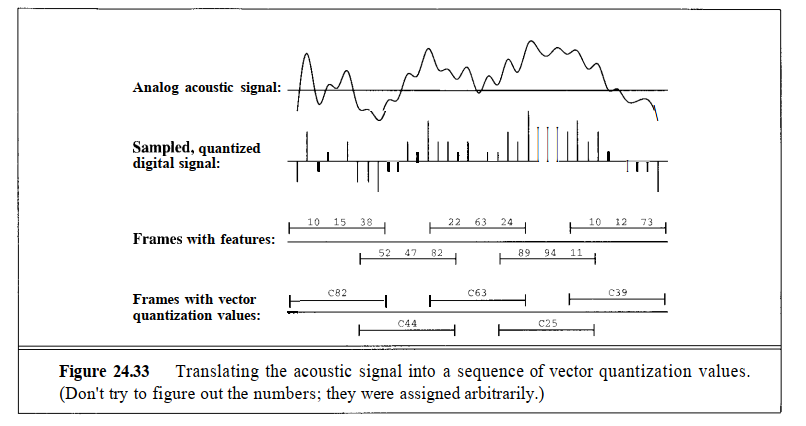


## **Nhận diện giọng nói**

**Xử lý tín hiệu**

Âm thanh là một nguồn năng lượng tương tự. Khi sóng âm thanh chạm vào micrô, nó sẽ được chuyển đổi thành dòng điện, dòng điện này có thể được truyền tới bộ chuyển đổi tương tự sang kỹ thuật số để tạo ra dòng bit đại diện cho âm thanh. Ta có hai lựa chọn trong việc quyết định nên giữ lại bao nhiêu bit. Đầu tiên, tốc độ lấy mẫu là tần số mà chúng ta xem xét tín hiệu. Đối với giọng nói, tốc độ lấy mẫu từ 8 đến 16 KHz (tức là 8 đến 16.000 lần mỗi giây) là điển hình. Điện thoại chỉ cung cấp khoảng 3 KHz. Thứ hai, hệ số lượng tử xác định độ chính xác mà năng lượng tại mỗi điểm lấy mẫu được ghi lại. Bộ nhận dạng giọng nói thường giữ 8 đến 12 bit. Điều đó có nghĩa là một hệ thống cấp thấp, lấy mẫu ở tần số 8 KHz với lượng tử hóa 8 bit, sẽ yêu cầu gần nửa megabyte mỗi phút giọng nói. Đây là rất nhiều thông tin cần thao tác, và tệ hơn, nó khiến chúng ta tiến rất xa mục tiêu khám phá những chiếc điện thoại tạo ra tín hiệu.

Bước đầu tiên để đưa ra cách biểu diễn tín hiệu tốt hơn là nhóm các mẫu lại với nhau thành các khối lớn hơn gọi là khung. Điều này giúp có thể phân tích toàn bộ khung hình để tìm sự xuất hiện của các hiện tượng lời nói như tăng hoặc giảm tần số, hoặc sự khởi đầu hoặc ngừng năng lượng đột ngột. Độ dài khung hình khoảng 10 ms (tức là 80 mẫu ở 8 KHz) dường như đủ dài để có thể phát hiện được hầu hết các hiện tượng như vậy và một số hiện tượng trong thời gian ngắn sẽ bị bỏ qua. Trong mỗi khung hình, chúng tôi thể hiện những gì đang xảy ra bằng một vectơ đặc điểm. Ví dụ: chúng ta có thể muốn mô tả lượng năng lượng ở mỗi dải tần số.



Các tính năng quan trọng khác bao gồm năng lượng tổng thể trong một khung hình và sự khác biệt so với khung hình trước đó. Chọn ra các đặc điểm từ tín hiệu giọng nói cũng giống như nghe một dàn nhạc và nói "ở đây tiếng kèn Pháp đang chơi to và tiếng vĩ cầm chơi nhẹ nhàng". Việc chia âm thanh thành các thành phần như thế này sẽ hữu ích hơn nhiều so với việc để nó dưới dạng một nguồn âm thanh không phân biệt duy nhất. Hình 33 hiển thị các khung có vectơ gồm ba đặc điểm. Lưu ý rằng các khung chồng lên nhau; điều này giúp chúng tôi không bị mất thông tin nếu một sự kiện âm thanh quan trọng tình cờ rơi vào ranh giới khung.

Bước cuối cùng trong nhiều hệ thống xử lý tín hiệu tiếng nói là lượng tử hóa vector. Nếu có n đặc điểm trong một khung, chúng ta có thể coi đây là một không gian n chiều chứa nhiều điểm. Lượng tử hóa vectơ chia không gian n chiều này thành 256 vùng được dán nhãn C1 đến C256. Khi đó, mỗi khung có thể được biểu diễn bằng một nhãn thay vì một vectơ gồm n số. Vì vậy, chúng tôi kết thúc chỉ với một byte cho mỗi khung hình, cải thiện khoảng 100 lần so với nửa megabyte mỗi phút ban đầu. Tất nhiên, một số thông tin bị mất khi chuyển từ vectơ đặc trưng sang nhãn tóm tắt toàn bộ vùng lân cận xung quanh vectơ, nhưng có các phương pháp tự động để chọn lượng tử hóa tối ưu của không gian vectơ đặc trưng sao cho có ít hoặc không có độ chính xác được đưa ra (Jelinek , 1990).

Có hai điểm cho toàn bộ bài tập này. Đầu tiên, nhóm em thu được một biểu diễn nhỏ gọn của tín hiệu giọng nói. Nhưng quan trọng hơn, nhóm em có một biểu diễn có khả năng mã hóa các đặc điểm của tín hiệu sẽ hữu ích cho việc nhận dạng từ. Một âm thanh lời nói nhất định có thể được phát âm theo nhiều cách: to hay nhẹ, nhanh hay chậm, the thé hay thấp, trên nền im lặng hoặc tiếng ồn, và bởi bất kỳ người nói nào trong số hàng triệu người nói khác nhau, mỗi người có giọng và giọng hát khác nhau. Quá trình xử lý tín hiệu hy vọng sẽ nắm bắt đủ các tính năng quan trọng để có thể chọn ra những điểm chung xác định âm thanh từ bối cảnh biến đổi này. (Vấn đề kép, nhận dạng người nói, đòi hỏi người ta phải tập trung vào sự khác biệt thay vì những điểm chung để quyết định ai đang nói.

**Xác định mô hình nhận dạng giọng nói tổng thể**

Nhận dạng giọng nói là nhiệm vụ chẩn đoán nhằm khôi phục các từ tạo ra tín hiệu âm thanh nhất định. Đó là một ví dụ kinh điển về lý luận với sự không chắc chắn. Chúng tôi không chắc chắn về việc micrô (và phần cứng số hóa) đã thu được âm thanh thực tế tốt đến mức nào, chúng tôi không chắc chắn về việc điện thoại nào sẽ phát ra tín hiệu và chúng tôi không chắc chắn về những từ nào sẽ phát ra tín hiệu cho điện thoại. Như thường lệ, nhiệm vụ chẩn đoán có thể được tiếp cận tốt nhất bằng mô hình nhân quả - các từ gây ra tín hiệu. Chúng ta có thể chia phần này thành các thành phần theo quy tắc Bayes:



Cho một tín hiệu, nhiệm vụ của chúng ta là tìm chuỗi từ làm cho P(tín hiệu từ) cực đại hóa. Trong ba thành phần ở vế phải, P(tín hiệu) là hằng số chuẩn hóa mà chúng ta có thể bỏ qua. P (từ) được gọi là mô hình ngôn ngữ. Đó là điều cho chúng ta biết, khi chúng ta không chắc mình đã nghe thấy "bad boy" hay "pad boy" thì khả năng xảy ra là "bad boy" hay "pad boy" cao hơn. Cuối cùng, P(từ tín hiệu) là mô hình âm thanh. Đó là điều cho chúng ta biết rằng "mèo" rất có thể được phát âm là [kæt].

**Mô hình ngôn ngữ: P(words)**

Thoạt nhìn, nhiệm vụ mô hình ngôn ngữ có vẻ khó khăn. Chúng ta phải gán xác suất cho mỗi số chuỗi (có thể là vô hạn). Các ngữ pháp phi ngữ cảnh không giúp ích được gì cho nhiệm vụ này, nhưng các ngữ pháp phi ngữ cảnh xác suất (PCFG) lại đầy hứa hẹn. Thật không may, PCFG không thể hiện tốt các hiệu ứng theo ngữ cảnh. Trong phần này, chúng ta tiếp cận vấn đề bằng cách sử dụng chiến lược tiêu chuẩn xác định xác suất của một sự kiện phức tạp là tích của xác suất của các sự kiện đơn giản hơn. Sử dụng ký hiệu W₁... W để biểu thị một chuỗi gồm n từ và w, để biểu thị từ thứ i của chuỗi, chúng ta có thể viết biểu thức xác suất của một chuỗi như sau:



Ưu điểm lớn của mô hình bigram là dễ dàng huấn luyện mô hình bằng cách đếm số lần mỗi cặp từ xuất hiện trong một tập chuỗi đại diện và sử dụng số đếm để ước tính xác suất. Ví dụ: nếu "a" xuất hiện 10.000 lần trong kho dữ liệu huấn luyện và theo sau là "súng" 37 lần thì P() = 37/10.000, trong đó P chúng ta muốn nói đến xác suất ước tính. Sau quá trình huấn luyện như vậy, người ta sẽ mong đợi "Tôi có" và "một khẩu súng" có xác suất ước tính tương đối cao, trong khi "Tôi có" và "một khẩu súng" sẽ có xác suất ước tính thấp. Một vấn đề là kho ngữ liệu huấn luyện có thể sẽ không chứa "gub" và quan trọng hơn, nó cũng sẽ thiếu nhiều từ tiếng Anh hợp lệ, vì vậy những từ này sẽ được gán xác suất ước tính bằng 0. Do đó, người ta thường dành một phần nhỏ phân bố xác suất cho các từ không xuất hiện trong kho ngữ liệu huấn luyện. Hình 34 cho thấy một số số lượng bigram bắt nguồn từ các từ trong chương này.



Có thể đi đến mô hình bát quái cung cấp các giá trị cho P(wi Wi-1Wi-2). Đây là một mô hình ngôn ngữ mạnh mẽ hơn, có khả năng xác định rằng "ăn một quả chuối" có nhiều khả năng hơn là "ăn một chiếc khăn rằn". Vấn đề là có quá nhiều tham số trong mô hình bát quái nên khó có đủ dữ liệu huấn luyện để đưa ra ước tính xác suất chính xác. Một sự thỏa hiệp tốt là sử dụng một mô hình bao gồm tổng có trọng số của các mô hình bát quái, bigram và unigram (tức là tần số từ). Mô hình được xác định theo công thức sau (với cl + C2+ C3 = 1):



**Mô hình âm thanh: P (word|signal)**

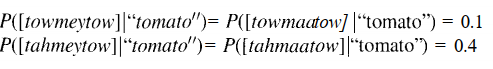
Mô hình âm thanh chịu trách nhiệm cho biết âm thanh nào sẽ được tạo ra khi một chuỗi từ nhất định được thốt ra. Chúng tôi chia mô hình thành hai phần. Đầu tiên, chúng tôi trình bày cách mỗi từ được mô tả dưới dạng một chuỗi các điện thoại và sau đó chúng tôi trình bày cách mỗi điện thoại liên quan đến các giá trị lượng tử hóa vectơ được trích xuất từ tín hiệu âm thanh.

Một số từ có mô hình phát âm rất đơn giản. Ví dụ: từ "cat" luôn được phát âm bằng ba âm fk a t]. Tuy nhiên, có hai nguồn biến đổi ngữ âm. Đầu tiên, các phương ngữ khác nhau có cách phát âm khác nhau. Phần trên của Hình 24.35 đưa ra một ví dụ về điều này: với "tomato", bạn nói [tow mey tow] và tôi nói [tow maa tow]. Các cách phát âm thay thế được chỉ định theo mô hình Markov. Nói chung, mô hình Markov là cách mô tả một quá trình trải qua một loạt trạng thái. Mô hình mô tả tất cả các đường đi có thể có trong không gian trạng thái và gán xác suất cho mỗi đường đi. Xác suất chuyển từ trạng thái hiện tại sang trạng thái khác chỉ phụ thuộc vào trạng thái hiện tại chứ không phụ thuộc vào bất kỳ phần nào trước đó của đường dẫn.

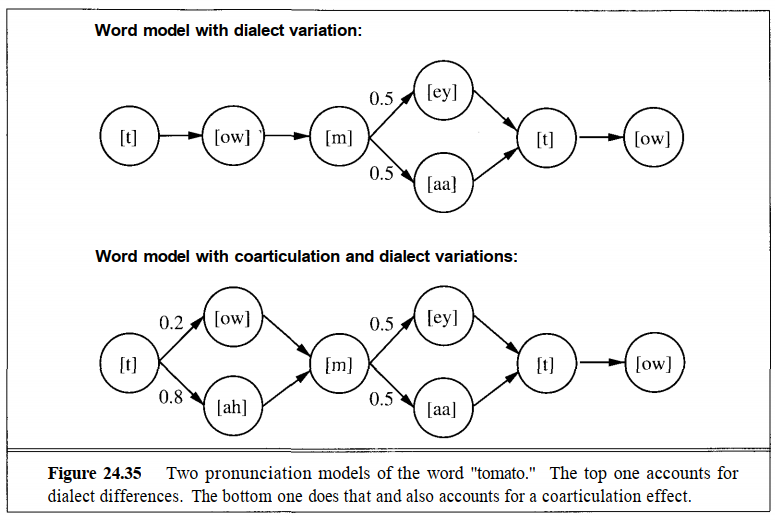
Phần trên của Hình 35 là mô hình Markov với bảy trạng thái (vòng tròn), mỗi trạng thái tương ứng với việc sản xuất một chiếc điện thoại. Các mũi tên biểu thị sự chuyển đổi được phép giữa các trạng thái và mỗi chuyển đổi có một xác suất liên quan đến nó. Chỉ có hai đường dẫn có thể có thông qua mô hình, một đường dẫn tương ứng với chuỗi điện thoại [tow mey tow] và đường còn lại tương ứng với [tow maa tow]. Xác suất của một đường đi là tích của các xác suất trên các cung tạo nên đường đi đó. Trong trường hợp này, hầu hết các xác suất cung là 1 và chúng ta có



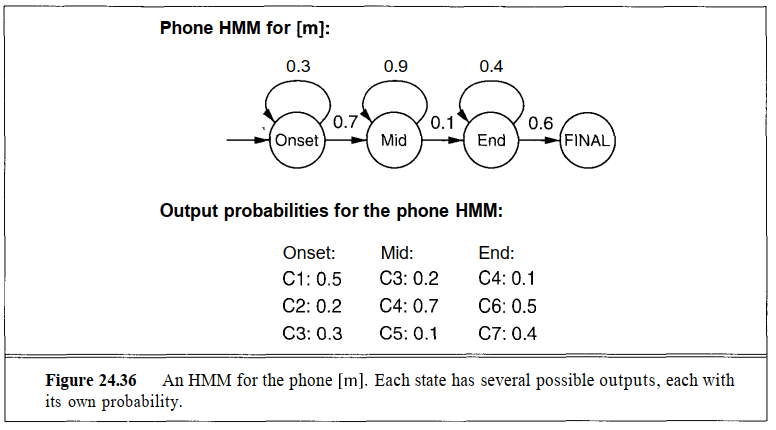
Nguồn biến đổi ngữ âm thứ hai là sự kết âm. Hãy nhớ rằng âm thanh lời nói được tạo ra bằng cách di chuyển lưỡi và hàm và đẩy không khí qua đường thanh quản. Khi người nói nói chậm rãi và thận trọng, sẽ có thời gian để đặt lưỡi vào đúng vị trí trước khi bắt đầu nói chuyện. Nhưng khi người nói nói nhanh (hoặc thậm chí đôi khi ở tốc độ bình thường), các chuyển động sẽ kết hợp với nhau. Ví dụ: âm [t] được phát âm với lưỡi ở phía trên miệng, trong khi âm [ow] có lưỡi ở gần phía dưới. Khi nói nhanh, lưỡi thường chuyển sang vị trí trung gian và chúng ta nhận được [tah] thay vì [kéo]. Nửa dưới của Hình 24.35 đưa ra một mô hình phát âm phức tạp hơn cho từ "tomato" có tính đến hiệu ứng đồng âm này. Trong mô hình này có bốn đường đi riêng biệt và chúng ta có



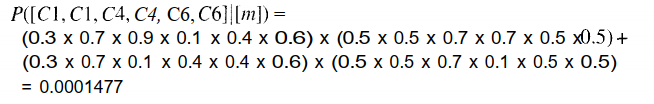
Các mô hình tương tự sẽ được xây dựng cho mọi từ mà chúng ta muốn có thể nhận ra. Bây giờ nếu tín hiệu giọng nói là một danh sách các điện thoại thì chúng ta sẽ hoàn thành mô hình âm thanh. Chúng ta có thể lấy một tín hiệu đầu vào nhất định (ví dụ: [towmeytow]) và tính P (word|signal) cho các chuỗi từ khác nhau (ví dụ: "tomato", "toe may tow", v.v.). Sau đó, chúng ta có thể kết hợp những giá trị này với các giá trị P(từ) lấy từ mô hình ngôn ngữ để đạt được những từ tối đa hóa P(word|signal).



Thật không may, việc xử lý tín hiệu không mang lại cho chúng ta một chuỗi điện thoại. Vì vậy, tất cả những gì chúng ta có thể làm cho đến nay là tối đa hóa P(words phone). Hình 24.36 cho thấy cách chúng ta có thể tính toán P (điện thoại tín hiệu) bằng cách sử dụng mô hình được gọi là mô hình Markov ẩn hay HMM. Kiểu này dành cho một điện thoại cụ thể [m], nhưng tất cả các điện thoại sẽ có kiểu có cấu trúc liên kết tương tự. Mô hình Markov ẩn giống mô hình Markov thông thường ở chỗ nó mô tả một quá trình trải qua một chuỗi các trạng thái. Sự khác biệt là trong mô hình Markov thông thường, đầu ra là một chuỗi các tên trạng thái và vì mỗi trạng thái có một tên duy nhất nên đầu ra xác định duy nhất đường dẫn xuyên qua mô hình. Trong mô hình Markov ẩn, mỗi trạng thái có phân bố xác suất của các đầu ra có thể có và cùng một đầu ra có thể xuất hiện ở nhiều trạng thái.4 HMM được gọi là mô hình ẩn vì trạng thái thực của mô hình bị ẩn khỏi người quan sát. Nói chung, khi bạn thấy HMM xuất ra một số ký hiệu, bạn không thể chắc chắn ký hiệu đó đến từ trạng thái nào.



Giả sử tín hiệu giọng nói của chúng ta được xử lý để tạo ra chuỗi các giá trị lượng tử hóa vectơ [C1,C4,C6]. Từ HMM trong Hình 36, chúng ta có thể tính xác suất chuỗi này được tạo bởi điện thoại [m] như sau. Đầu tiên, chúng tôi lưu ý rằng chỉ có một đường dẫn xuyên suốt mô hình có thể tạo ra chuỗi này: đường dẫn từ Bắt đầu đến Giữa đến Cuối, trong đó nhãn đầu ra từ ba trạng thái lần lượt là C1, C4 và C6. Nhìn vào xác suất trên các cung chuyển tiếp, chúng ta thấy xác suất của đường đi này là 0,7 x 0,1 x 0,6 (đây là các giá trị trên ba mũi tên ngang ở giữa Hình 24.36). Tiếp theo, chúng ta xem xét xác suất đầu ra của các trạng thái này để thấy rằng xác suất của [C1,C4,C6] cho đường dẫn này là 0,5 x 0,7 x 0,5 (đây là các giá trị cho P(C1|Onset), P(C4|Mid ) và P(C6|End), tương ứng). Vậy xác suất của [C1,C4,C6] cho mô hình [m] là:



Chúng ta có thể lặp lại phép tính cho tất cả các kiểu điện thoại khác để xem kiểu nào có khả năng là nguồn tín hiệu giọng nói nhất.

Trên thực tế, hầu hết các điện thoại đều có thời lượng 50-100 mili giây hoặc 5-10 khung hình ở tốc độ 10 mili giây/khung hình. Vì vậy, chuỗi [C1,C4,C6] nhanh một cách bất thường. Giả sử chúng ta có một diễn giả điển hình hơn, người tạo ra chuỗi [C1,C1,C4,C4,C6,C6] trong khi sản xuất điện thoại. Hóa ra có hai đường dẫn thông qua mô hình tạo ra chuỗi này. Ở một trong số chúng, cả hai C4 đều đến từ trạng thái Giữa (lưu ý các cung lặp lại) và ở cái còn lại, C4 thứ hai đến từ trạng thái Kết thúc. Chúng tôi tính xác suất để chuỗi này đến từ mô hình [m] theo cách tương tự: lấy tổng trên tất cả các đường dẫn có thể có của xác suất của đường dẫn nhân với xác suất mà đường dẫn tạo ra chuỗi.

Chúng tôi thấy rằng các vòng lặp trong kiểu điện thoại cho phép mô hình thể hiện cả giọng nói nhanh và giọng nói chậm, một nguồn biến thể rất quan trọng. Nhiều giá trị lượng tử hóa vectơ trên mỗi trạng thái biểu thị các nguồn biến thể khác. Nhìn chung, điều này tạo nên một mô hình khá mạnh mẽ. Phần khó là nhận được giá trị xác suất tốt cho tất cả các tham số. May mắn thay, có nhiều cách để thu được những con số này từ dữ liệu, như chúng ta sẽ thấy.

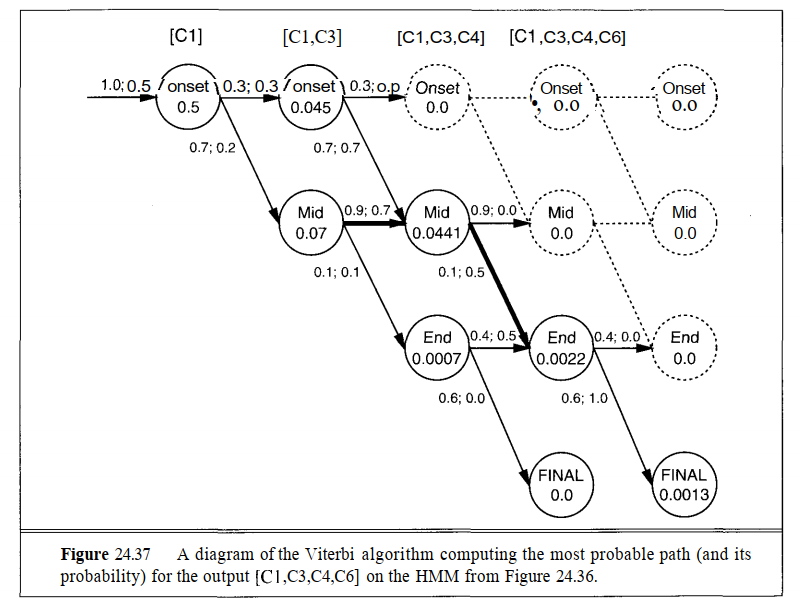
**Tập hợp các mô hình lại với nhau**

Nhóm em đã mô tả ba mô hình. Mô hình bigram ngôn ngữ cho chúng ta P(word; word;-1). Cách phát âm từ HMM cho chúng ta P(từ điện thoại). Điện thoại HMM cho chúng ta P (điện thoại tín hiệu). Nếu chúng ta muốn tính P(tín hiệu từ), chúng ta sẽ cần kết hợp các mô hình này theo một cách nào đó. Một cách tiếp cận là kết hợp tất cả chúng thành một HMM lớn. Mô hình bigram có thể được coi như một HMM trong đó mọi trạng thái tương ứng với một từ và mỗi từ có một cung chuyển tiếp sang mỗi từ khác. Bây giờ hãy thay thế từng trạng thái từ bằng mô hình từ thích hợp, mang lại một mô hình lớn hơn trong đó mỗi trạng thái tương ứng với một chiếc điện thoại. Cuối cùng, thay thế từng trạng thái điện thoại bằng kiểu điện thoại thích hợp, tạo ra một mô hình thậm chí còn lớn hơn trong đó mỗi trạng thái tương ứng với một phân bố các giá trị lượng tử hóa vectơ.

Một số hệ thống nhận dạng giọng nói làm phức tạp hình ảnh bằng cách xử lý các hiệu ứng khớp âm ở cấp độ từ/từ hoặc điện thoại/điện thoại. Ví dụ: chúng ta có thể sử dụng một kiểu điện thoại cho [ow] khi nó theo sau [t] và một kiểu điện thoại khác cho [ow] khi nó theo sau [g]. Có rất nhiều sự đánh đổi để tạo ra một mô hình phức tạp hơn có thể xử lý các hiệu ứng tinh vi, nhưng sẽ khó đào tạo hơn. Bất kể chi tiết thế nào, chúng ta đều có được một HMM lớn có thể được sử dụng để tính P (tín hiệu từ).

**Thuật toán tìm kiếm**

Từ quan điểm lý thuyết, chúng ta có đúng thứ mà chúng tôi yêu cầu: một mô hình tính toán P (tín hiệu từ). Tất cả những gì chúng ta phải làm là liệt kê tất cả các chuỗi từ có thể có và gán xác suất cho mỗi chuỗi. Tất nhiên, trên thực tế, điều này là không khả thi vì có quá nhiều chuỗi từ ứng cử viên. May mắn thay, có một cách tốt hơn.



Thuật toán Viterbi lấy mô hình HMM và một chuỗi đầu ra, [C1, C2,., Cn], rồi trả về đường dẫn có khả năng xảy ra cao nhất thông qua HMM tạo ra chuỗi đó. Nó cũng trả về xác suất của đường đi. Hãy coi nó như một thuật toán lặp trước tiên tìm tất cả các đường dẫn tạo ra ký hiệu đầu tiên, Cl. Sau đó, đối với mỗi đường dẫn đó, nó sẽ tìm đường dẫn có khả năng xảy ra nhất tạo ra phần còn lại của chuỗi, với điều kiện là chúng ta đã chọn một đường dẫn cụ thể cho C₁. Cho đến nay điều này nghe có vẻ không mấy hứa hẹn. Nếu độ dài của chuỗi là n và có M trạng thái khác nhau trong mô hình thì thuật toán này có vẻ ít nhất là O(M").

Điểm mấu chốt của thuật toán Viterbi là sử dụng thuộc tính Markov để làm cho nó hiệu quả hơn. Thuộc tính Markov nói rằng đường đi có thể xảy ra nhất cho phần còn lại của bất kỳ chuỗi nào chỉ có thể phụ thuộc vào trạng thái mà nó bắt đầu chứ không phụ thuộc vào bất kỳ điều gì khác về đường đi đến đó. Điều đó có nghĩa là chúng ta không cần phải xem xét tất cả các con đường có thể dẫn đến một trạng thái nhất định; đối với mỗi trạng thái, chúng ta chỉ cần theo dõi con đường có khả năng xảy ra cao nhất sẽ kết thúc ở trạng thái đó. Vì vậy, thuật toán Viterbi là một ví dụ của quy hoạch động.

Hình 24.37 cho thấy thuật toán làm việc trên HMM từ Hình 24.36 và kết quả đầu ra

trình tự [C1,C3,C4,C61. Mỗi cột đại diện cho một lần lặp của thuật toán. Ở ngoài cùng bên trái

cột, chúng ta thấy rằng chỉ có một cách duy nhất để tạo ra chuỗi [Cl], với đường dẫn [Onset]. Hình bầu dục có nhãn "Khởi động 0,5" có nghĩa là đường dẫn kết thúc ở trạng thái Khởi động và có xác suất 0,5. Cung dẫn vào hình bầu dục có nhãn "1,0; 0,5", có nghĩa là xác suất thực hiện quá trình chuyển đổi này là 1,0 và xác suất tạo ra Cl, với điều kiện quá trình chuyển đổi được thực hiện, là 0,5. Trong cột thứ hai, chúng tôi xem xét tất cả các khả năng tiếp tục của các đường dẫn trong cột đầu tiên có thể dẫn đến đầu ra [C1,C3]. Có hai đường dẫn như vậy, một đường kết thúc ở trạng thái Khởi động và một ở trạng thái Giữa. Ở cột thứ ba, nó trở nên thú vị hơn. Có hai con đường dẫn đến trạng thái Mid, một từ Onset và một từ Mid. Mũi tên đậm chỉ ra rằng đường đi từ Mid có nhiều khả năng xảy ra hơn (nó có xác suất 0,0441), vì vậy đó là đường đi duy nhất

**Đào tạo mẫu**

Phương pháp HMM được sử dụng trong nhận dạng giọng nói vì hai lý do. Đầu tiên, nó là một phần tử có hiệu suất khá tốt - chúng tôi thấy rằng thuật toán Viterbi tuyến tính theo độ dài của đầu vào. Quan trọng hơn, HMM có thể được học trực tiếp từ tập huấn luyện gồm các cặp [tín hiệu, từ]. Điều này rất quan trọng vì rất khó để xác định tất cả các thông số bằng tay. Có những cách tiếp cận khác tạo ra các yếu tố hiệu suất tốt hơn HMM, nhưng chúng yêu cầu dữ liệu đào tạo phải được gắn nhãn trên cơ sở từng điện thoại thay vì trên cơ sở từng câu và đó cũng là một nhiệm vụ khó khăn. Thuật toán tiêu chuẩn để huấn luyện HMM được gọi là thuật toán Baum-Welch hoặc thuật toán tiến-lùi. Rabiner (1990) đưa ra hướng dẫn về thuật toán này và các thuật toán HMM khác.

Các hệ thống nhận dạng giọng nói tốt nhất hiện nay nhận dạng chính xác từ khoảng 80% đến 98% các từ, tùy thuộc vào chất lượng tín hiệu, ngôn ngữ cho phép, độ dài của mỗi đầu vào và sự thay đổi của người nói. Nhận dạng giọng nói thật dễ dàng khi có micrô tốt, vốn từ vựng nhỏ, mô hình ngôn ngữ mạnh mẽ có thể dự đoán những từ nào có thể xảy ra tiếp theo, giới hạn số lần phát âm một từ (hoặc yêu cầu tạm dừng giữa các từ) và khi hệ thống có thể được đào tạo đặc biệt cho một diễn giả duy nhất. Hệ thống không chính xác bằng đường dây điện thoại, khi có tiếng ồn trong phòng, khi có lượng từ vựng lớn không bị hạn chế, khi các từ trong một câu nói chạy cùng nhau và khi người nói mới sử dụng hệ thống.

# **Robotics**

## **Khái niệm robot là gì?**

Robot là một bộ điều khiển đa chức năng, có thể lập trình được thiết kế để di chuyển vật liệu, bộ phận, công cụ hoặc thiết bị cụ thể thông qua các chuyển động được lập trình thay đổi để thực hiện nhiều nhiệm vụ khác nhau.

Robot học là 1 nhánh của AI cái mà tập trung vào thiết kế và phát triển robot thông minh. Robot đó phục vụ cho các hoạt động nhân tạo cái tương tác với môi trường thực tiễn. Những mục tiêu chính là điều khiển robot bằng nhận thức, cầm nắm, chuyển động và làm thay đổi đặc tính vật lý của chúng. Điều này cho phép robot thực hiện nhiệm vụ lặp đi lặp lại và đòi hỏi thể lực, giải phóng người lao động không tốn sức đồng thời đảm bảo tính nhất quán và tin cậy trong vận hành.

## **Robot làm được những gì ?**

Trong khi con người làm nhiều việc khác nhau bằng cách sử dụng ít nhiều cùng một cơ thể, thì thiết kế của robot rất khác nhau tùy thuộc vào nhiệm vụ mà chúng dự định thực hiện.

- **Sản xuất và xử lý vật liệu**

Sản xuất được coi là lĩnh vực truyền thống của robot. Các nhiệm vụ lặp đi lặp lại trên dây chuyền sản xuất là mục tiêu tự nhiên của tự động hóa  
 Tuy nhiên, hầu hết các robot được sử dụng ngày nay đều rất hạn chế về khả năng cảm nhận và thích ứng. Robot tự động vẫn đang đấu tranh để được chấp nhận.

- **Robot kẹp gắp và vận chuyển**

Robot di động (mobot) cũng đang trở nên hữu ích rộng rãi. Hai ứng dụng chính là chuyển phát nhanh trong các tòa nhà, đặc biệt là bệnh viện và làm nhân viên bảo vệ.

Vd: Một công ty đã bán được hơn 3000 chiếc "điện thoại di động gửi thư". Những robot này đáp ứng các yêu cầu từ thiết bị đầu cuối máy tính và mang tài liệu hoặc vật tư đến đích ở một nơi khác trong tòa nhà. Robot di chuyển qua hành lang và thang máy, tránh va chạm với các chướng ngại vật khác như con người và đồ đạc.

-Robot phù hợp với nhiệm vụ này vì nó có tính sẵn sàng cao (có thể làm việc 24 giờ một ngày), độ tin cậy cao để nguồn cung cấp không bị mất hoặc thất lạc và tiến trình của nó có thể được theo dõi để cho phép chuẩn bị cho việc đến nơi hoặc phát hiện sự cố. một robot có thể hoạt động liên tục trong nhiều tháng, thực hiện các quan sát khoa học, báo cáo về hoạt động di chuyển của tàu ngầm đối phương hoặc khắc phục các sự cố trong cáp xuyên đại dương.

Chúng tôi đã thấy (trang 586) rằng các phương tiện tự động đang di chuyển trên đường cao tốc của chúng ta và AUV (phương tiện tự hành dưới nước) đang di chuyển trên biển. Việc gửi một robot xuống đáy đại dương sẽ rẻ hơn nhiều so với việc gửi một chiếc tàu ngầm có người lái.

- **Môi trường nguy hiểm**

Robot di động là một công nghệ quan trọng giúp giảm thiểu rủi ro đến tính mạng con người trong môi trường nguy hiểm

Vd: Chúng cũng có thể được sử dụng cho các nhiệm vụ thường ngày trong các tình huống nguy hiểm bao gồm dọn dẹp chất thải độc hại, thu hồi biển sâu, thăm dò, khai thác và xử lý các vật liệu nguy hiểm sinh học. Người điều khiển thường sẵn sàng hướng dẫn robot, nhưng nó cũng cần có quyền tự chủ để nhận biết và ứng phó với các mối nguy hiểm đối với sức khỏe của chính nó và của người khác.

- **Telepresence and thực tế ảo**

Khi máy tính lan rộng từ các ứng dụng khoa học/kinh doanh sang công nghệ tiêu dùng, khả năng hiện diện từ xa và thực tế ảo tiếp tục thu hút trí tưởng tượng của công chúng. Ý tưởng ở trong nhà của một người và có thể cảm nhận được những môi trường kỳ lạ, thực tế (điều khiển từ xa) hoặc tưởng tượng (thực tế ảo) thực sự rất hấp dẫn và là động lực thúc đẩy một số bước chuyển lớn trong ngành công nghiệp máy tính và giải trí.

- **Tăng khả năng của con người**

Không kém phần hấp dẫn là nỗ lực tái tạo các tác nhân đã mất của con người. Khi một cánh tay hoặc chân bị cắt cụt, các cơ ở phần cụt vẫn phản ứng với tín hiệu từ não, tạo ra dòng điện cơ. Li,mbs giả có thể thu các tín hiệu này và khuếch đại chúng để uốn cong các khớp và di chuyển các ngón tay nhân tạo. Một số bộ phận giả thậm chí còn có phản hồi điện da mang lại cảm giác chạm vào. Cũng đã có công việc cung cấp cho con người những cảm biến nhân tạo. Ví dụ, một dự án MITI của Nhật Bản đã chế tạo một nguyên mẫu robot dẫn đường cho người mù. Nó sử dụng sóng siêu âm để đảm bảo rằng chủ nhân của nó luôn ở trong khu vực an toàn khi họ đi cùng nhau. Võng mạc và ốc tai nhân tạo, chủ yếu dựa trên VLSI tương tự, hiện đang là chủ đề của nghiên cứu chuyên sâu.

## **Robot được làm từ những gì?**

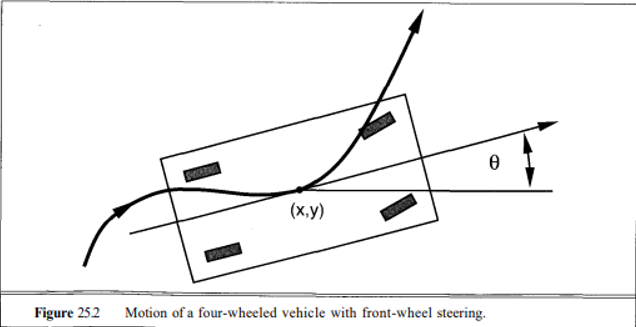
Các robot được phân biệt với nhau bởi các bộ phận tác động và cảm biến mà chúng được trang bị. Một số robot có các đầu nối đặc biệt ở liên kết cuối cùng cho phép chúng nhanh chóng loại bỏ một bộ phận tác động đầu cuối và gắn một bộ phận tác động đầu cuối khác. Robot được trang bị tốt cũng có một hoặc nhiều cảm biến, có thể bao gồm máy ảnh, cảm biến hồng ngoại, radar, sóng âm và gia tốc kế.

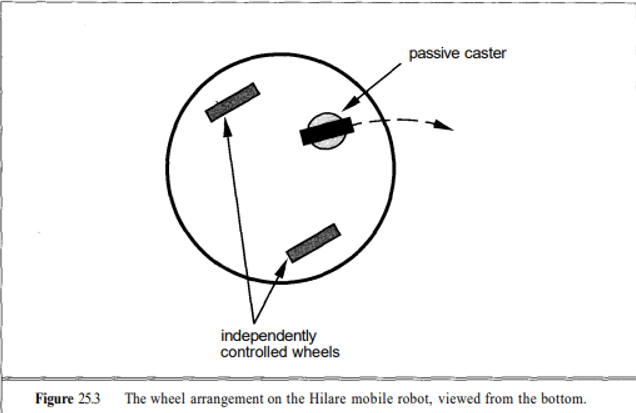
**3.1. Tác nhân: công cụ hành động**

Bộ tác động được sử dụng theo hai cách chính: thay đổi vị trí của robot trong môi trường của nó (vận động) và di chuyển các vật thể khác trong môi trường (thao tác). Công dụng thứ ba, để thay đổi hình dạng hoặc các đặc tính vật lý khác của vật thể, thuộc lĩnh vực kỹ thuật cơ khí hơn là robot, vì vậy chúng tôi sẽ không đề cập đến nó.

**Vận động**

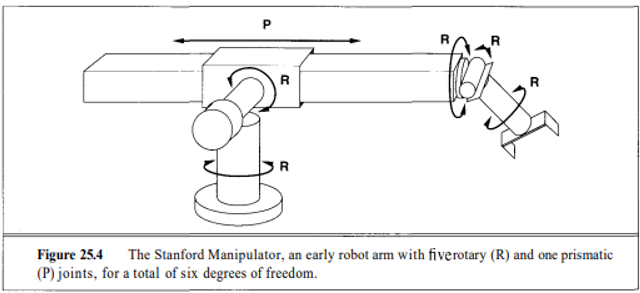
Có thể chế tạo các robot di động thực sự có hình ảnh ba chiều, nhưng với cái giá phải trả là độ phức tạp cơ học cao. Cần phải thiết kế các bánh xe hoặc mặt lốp sao cho chuyển động theo hướng lái được kiểm soát nhưng chuyển động ngang được tự do. Bộ truyền động ba chiều thường thay thế lốp hoặc mặt lốp bằng một loạt con lăn thẳng hàng với hướng truyền động. Mặc dù những thiết kế này làm cho cuộc sống của kiến ​​trúc sư điều khiển trở nên dễ dàng hơn, nhưng những thiết kế phổ biến của robot di động phi giao diện (Hình 25.2 và 25.3) không khó điều khiển và tính đơn giản về mặt cơ học của chúng khiến chúng trở thành lựa chọn tốt nhất trong hầu hết các tình huống.





**Thao tác**

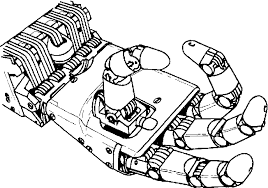
Hầu hết các bộ điều khiển đều cho phép chuyển động quay (quay quanh một trục cố định) hoặc chuyển động lăng trụ (chuyển động tuyến tính, như với một pít-tông bên trong một hình trụ).



**3.2. Cảm biến: công cụ về nhận thức**

**Quyền sở hữu**

Giống như con người, robot có giác quan cảm thụ bản thể để cho chúng biết khớp của chúng ở đâu. Bộ mã hóa được lắp vào các khớp cung cấp dữ liệu rất chính xác về góc hoặc độ giãn của khớp. Nếu đầu ra của bộ mã hóa được đưa trở lại cơ cấu điều khiển trong quá trình chuyển động, robot có thể có độ chính xác định vị cao hơn nhiều so với con người. Đối với một bộ điều khiển, điều này thường có nghĩa là độ chính xác khoảng vài phần nghìn (phần nghìn inch) ở vị trí bộ phận tác động cuối của nó. Ngược lại, con người chỉ có thể quản lý được một hoặc hai cm. Để tự kiểm tra điều này, hãy đặt ngón tay của bạn vào một đầu của thước kẻ (hoặc một vật nào khác có chiều dài quen thuộc).



**Cảm biến lực**

ví dụ: nhiệm vụ cạo sơn khỏi kính cửa sổ bằng lưỡi dao cạo. Để có được toàn bộ lớp sơn đòi hỏi phải định vị chính xác khoảng một micron theo hướng vuông góc với kính. Một sai số nhỏ đến từng milimet sẽ khiến robot bị mất lớp sơn hoàn toàn hoặc làm vỡ kính. Rõ ràng, con người không làm điều này chỉ bằng cách sử dụng điều khiển vị trí.

Chuyển động phù hợp với cảm biến lực. Nhiều nhiệm vụ khác liên quan đến tiếp xúc, chẳng hạn như viết, mở cửa và lắp ráp ô tô, đòi hỏi phải kiểm soát lực chính xác. Lực có thể được điều chỉnh ở một mức độ nào đó bằng cách điều khiển dòng điện của động cơ điện, nhưng việc điều khiển chính xác cần có cảm biến lực. Những cảm biến này thường được đặt giữa bộ điều khiển và bộ phận tác động cuối và có thể cảm nhận được lực và mô men xoắn theo sáu hướng. Bằng cách sử dụng điều khiển lực, robot có thể di chuyển dọc theo một bề mặt trong khi vẫn duy trì tiếp xúc với một áp suất cố định. Những chuyển động như vậy được gọi là chuyển động tuân thủ và cực kỳ quan trọng trong nhiều ứng dụng robot.

**Cảm biến xúc giác**

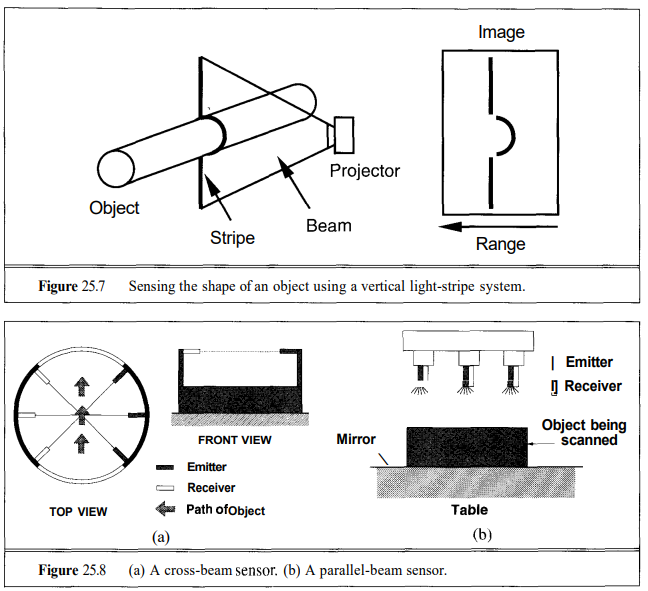
Việc nhặt một tách cà phê bằng giấy hay thao tác với một chiếc vít nhỏ đòi hỏi nhiều điều hơn là khả năng cảm nhận bản thân. Lực tác dụng lên chiếc cốc phải vừa đủ để nó không bị trượt nhưng không đủ để làm nó bị nát. Việc thao tác vít đòi hỏi thông tin về vị trí chính xác của nó so với các ngón tay mà CẢM BIẾN TACTILE nó tiếp xúc. Trong cả hai trường hợp, cảm biến xúc giác (hoặc cảm biến xúc giác) có thể cung cấp thông tin cần thiết. Cảm biến xúc giác là phiên bản robot của xúc giác con người. Cảm biến xúc giác của robot sử dụng vật liệu đàn hồi và sơ đồ cảm biến để đo độ biến dạng của vật liệu khi tiếp xúc. Cảm biến có thể cung cấp dữ liệu tại một loạt điểm trên bề mặt đàn hồi, tạo ra hình ảnh tương tự như hình ảnh camera nhưng có sự biến dạng hơn là cường độ ánh sáng. Bằng cách hiểu tính chất vật lý của quá trình biến dạng, có thể rút ra các thuật toán tương tự như thuật toán thị giác và có thể tính toán thông tin vị trí cho các vật thể mà cảm biến chạm vào. Cảm biến xúc giác cũng có thể cảm nhận được độ rung, giúp phát hiện tách cà phê sắp thoát ra khỏi tay cầm của người giữ. Con người sử dụng sơ đồ này với vòng lặp servo 5 rất nhanh để phát hiện trượt và kiểm soát lực bám đến gần mức tối thiểu cần thiết để chống trượt

**Cảm biến sonar**

Sonar là định hướng âm thanh và phạm vi. Sonar cung cấp thông tin hữu ích về các vật thể ở rất gần robot và thường được sử dụng để tránh va chạm khẩn cấp nhanh chóng. Đôi khi nó được sử dụng để lập bản đồ môi trường của robot trên một khu vực lớn hơn. Trong trường hợp thứ hai, một dãy gồm hàng chục cảm biến siêu âm trở lên được lắp xung quanh chu vi của robot, mỗi cảm biến hướng về một hướng khác nhau. Mỗi cảm biến lý tưởng sẽ đo khoảng cách đến chướng ngại vật gần nhất theo hướng nó chỉ. Sonar hoạt động bằng cách đo thời gian truyền đi của xung âm thanh do cảm biến tạo ra để chạm tới một vật thể và bị phản xạ trở lại. Xung hoặc "chirp" thường có tần số khoảng 50 kHz. Con số này cao hơn gấp đôi giới hạn trên đối với con người là 20 kHz. Âm thanh ở tần số đó có bước sóng khoảng 7 mm. Tốc độ của âm thanh là khoảng 330 m/giây, do đó thời gian trễ của một vật cách xa 1 m là khoảng 6 x 10 ~ 3 giây. Sonar rất hiệu quả trong việc tránh chướng ngại vật và theo dõi mục tiêu ở gần, chẳng hạn như một robot di động khác. Nhưng mặc dù có thể đo độ trễ thời gian rất chính xác, sóng siêu âm hiếm khi có thể tạo ra dữ liệu chính xác và đáng tin cậy để lập bản đồ. Vấn đề đầu tiên là độ rộng chùm tia. Thay vì tạo ra chùm âm thanh hẹp, cảm biến thông thường tạo ra chùm âm thanh hình nón có độ lan truyền từ 10° trở lên. Vấn đề thứ hai xuất phát từ bước sóng tương đối dài (7 mm) của âm thanh sonar. Các vật thể rất trơn tru so với bước sóng này trông sáng bóng hoặc "gương" đối với cảm biến. Những vật thể như vậy phản chiếu âm thanh giống như một tấm gương hoàn hảo. Âm thanh sẽ chỉ được nhận lại từ các mảng bề mặt vuông góc với chùm tia. Các đồ vật có bề mặt phẳng và cạnh sắc sẽ phản xạ rất ít âm thanh theo hầu hết các hướng. (Quan sát tương tự được sử dụng để thiết kế máy bay và tàu tàng hình tránh radar.) Sau khi bị phản xạ từ bề mặt, âm thanh có thể chạm vào bề mặt gồ ghề và bị phản xạ trở lại cảm biến. Thời gian trễ sẽ không tương ứng với một vật thể vật lý mà tương ứng với một "con ma", thứ có thể biến mất một cách bí ẩn khi robot di chuyển. Cảm biến nhiễu có thể được xử lý bằng cách trước tiên xây dựng mô hình xác suất của cảm biến, sau đó sử dụng cập nhật Bayesian để tích hợp thông tin thu được theo thời gian khi robot di chuyển xung quanh. Cuối cùng, những bản đồ có độ chính xác hợp lý có thể được xây dựng và những hình ảnh ma quái có thể được loại bỏ.

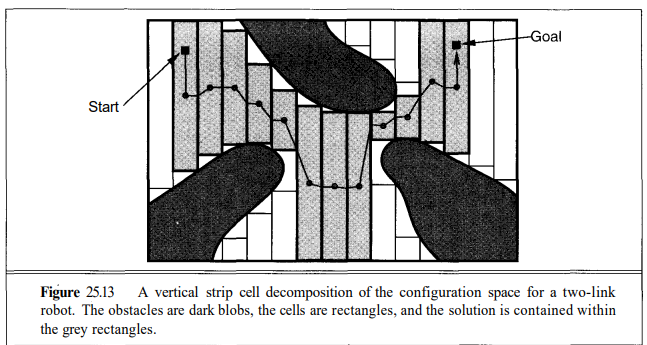
**Camera data**

Bằng cách di chuyển sọc hoặc bằng cách sử dụng một số raster sọc ở các khoảng cách khác nhau, có thể tạo ra bản đồ ba chiều rất dày đặc của vật thể trong một khoảng thời gian ngắn. Hiện nay có một số thiết bị bao gồm nguồn laser, điều khiển sọc, máy ảnh và tất cả quá trình xử lý hình ảnh cần thiết để tính toán bản đồ khoảng cách đến các điểm trong hình ảnh. Theo quan điểm của người dùng, những công cụ tìm phạm vi laser này thực sự là cảm biến độ sâu, cung cấp hình ảnh độ sâu được cập nhật nhanh chóng, có thể vài lần một giây.

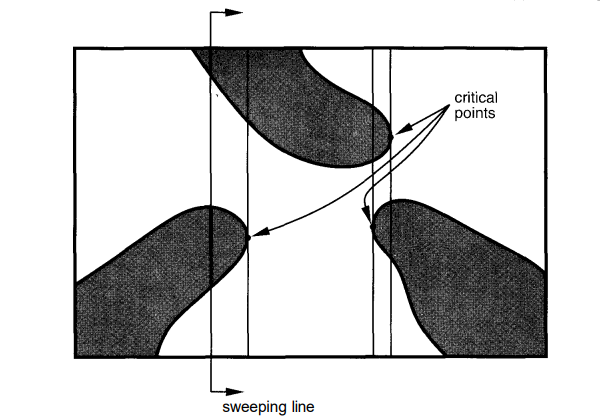


## **Thu nhập thông tin 3 chiều sử dụng hình ảnh hoặc dữ liệu**

**Thu nhập thông tin 3 chiều sử dụng hình ảnh hoặc dữ liệu1. Phân chia ô** chia không gian liên tục thành một số lượng ô hữu hạn, dẫn đến vấn đề tìm kiếm rời rạc



**2. Các phương pháp khung hóa** tính toán một "khung" một chiều của không gian cấu hình, mang lại một vấn đề tìm kiếm đồ thị tương đương**.**

****

**3. Lập kế hoạch sai số giới hạn**

giả định các giới hạn về độ không chắc chắn của cảm biến và bộ truyền động, và trong một số trường hợp có thể tính toán các kế hoạch được đảm bảo thành công ngay cả khi gặp phải lỗi nghiêm trọng của bộ truyền động

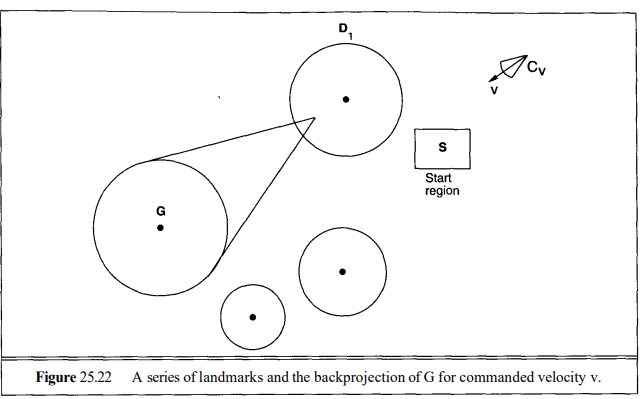
**4. Điều hướng dựa trên điểm mốc**

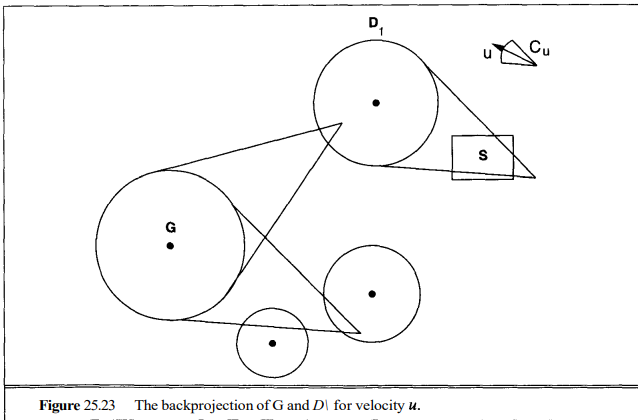
giả định rằng có một số khu vực trong đó vị trí của robot có thể được xác định chính xác bằng cách sử dụng các điểm mốc, trong khi ở bên ngoài các khu vực đó, robot có thể chỉ có thông tin định hướng

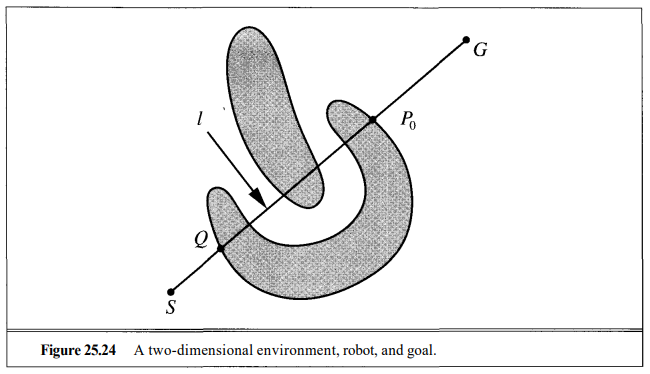
**5. Thuật toán online**

Chúng ta giả định rằng môi trường ban đầu hoàn toàn không xác định được, mặc dù hầu hết đều giả định một số dạng cảm biến vị trí chính xác

Kiến trúc cổ điển: Trước hết, nó giới thiệu về các kiến trúc cổ điển của robot, bắt đầu từ những năm 1960 với các công cụ đầu tiên cho robot thông minh như hệ thống thị giác và các thuật toán lập kế hoạch. Shakey, một trong những dự án robot thông minh đầu tiên, đã đặt nền móng cho nhiều dự án sau này. Tuy nhiên, nhận thấy rằng kế hoạch ban đầu của Shakey thường gặp khó khăn khi thực thi do lỗi bánh xe trượt, sai số đo lường và lỗi khác.







\_Hết\_