Tạp chí Tin học và Điều khiển học, T.23, S.1 (2007), 27–38

KẾT HỢP ĐẶC TRƯNG THỊ GIÁC VÀ NGỮ NGHĨA TRONG TRUY VẤN VIDEO SỐ DỰA TRÊN MÔ HÌNH PHÂN CẤP DỮ LIỆU

NGUYỄN LÃM[1], LÝ QUỐC NGỌC[2], DƯƠNG ANH ĐỨC[2]

[1] *Trường Đại học Lương Thế Vinh, Nam Định*
[2] *Trường Đại học Khoa học Tự nhiên - ĐHQG - Hồ Chí Minh*

Abstract. Nowadays, digital video documents grow in both the number and the size of storing spaces. Therefore, it requires efficient management techniques and methods that allow retrieving video documents in an efficient way. This paper presents a method for automatic structural analysis of digital videos to generate the table of content and the index table of the given videos.These tables allow storing videos by a hierarchy of video, cluster of shots, shots, key-frames of shots, cluster of regions. Then a method for retrieving videos by using the input data as visual feature and semantic concepts is represented.

The video retrieval is performed by two steps in off-line and on-line modes. The off-line step consists of decomposing a video sequence into elementary segments. Then, these elementary segments are classified by a hierarchical clustering algorithm. Finally, a table of content and an index table are generated for the given video sequence. The on-line step consists of retrieving videos based on a hierachical data base using the input data as video clip, shot ,key-frames, representatives of region's clusters, then the retrieved results are filtered by semantic concepts.

The obtained results show that the proposed model is more efficient than the traditional systems which are only based on global, local visual features or keywords.

Tóm tắt. Hiện nay dữ liệu video số được lưu trữ và phát triển với số lượng ngày càng tăng, do vậy dẫn đến một nhu cầu là cần có một cách thức quản lý hữu hiệu hơn để phục vụ việc truy tìm thông tin và cách thức truy tìm. Trong bài báo này chúng tôi trình bày một phương pháp giúp phân tích tự động cấu trúc của video số nhằm tạo ra bản mục lục và chỉ mục, giúp lưu trữ nội dung đoạn video số theo cấu trúc phân cấp: video, lớp các đoạn cơ sở, đoạn cơ sở, khung hình chính, lớp các vùng và truy vấn dựa vào đặc trưng thị giác và ngữ nghĩa.

Bài toán được tiếp cận bằng việc đầu tiên là phân tích tự động video số thành các đoạn cơ sở, sau đó nhóm chúng lại theo phương pháp phân lớp phân cấp và cuối cùng là rút gọn cấu trúc phân cấp để tạo bảng mục lục và chỉ mục. Việc truy vấn được thực hiện dựa trên cấu trúc phân cấp với hai giai đoạn, trong giai đoạn đầu, kết quả truy vấn dựa vào đặc trưng thị giác, trong giai đoạn cuối, kết quả được lọc lại dựa vào ngữ nghĩa hoặc ngược lại.

Kết quả thực nghiệm cho thấy phương pháp này đạt kết quả cao hơn so với phương pháp truy vấn chỉ dựa vào đặc trưng thị giác toàn cục, cục bộ hoặc ngữ nghĩa.

1. GIỚI THIỆU

Bài báo đề xuất xây dựng một hệ thống truy vấn video số ở dạng tổng quát, đồng thời

có thể hiệu chỉnh để đáp ứng truy vấn ở dạng đặc thù nhằm đáp ứng nhu cầu có thực trong xã hội thông tin. Hiện nay người ta chưa thể xây dựng được các hệ truy vấn video đáp ứng mọi yêu cầu của người dùng do bản chất rất đa dạng của dạng dữ liệu video số, dạng dữ liệu mang yếu tố không gian và thời gian.Tuy nhiên các hệ truy vấn video số vẫn đang được nghiên cứu và phát triển với tốc độ rất nhanh nhằm đáp ứng được yêu cầu của cuộc sống. Trong bối cảnh đó, kế thừa và phát triển các kết quả nghiên cứu của cộng đồng truy vấn video số ở trong và ngoài nước, đóng góp chính của bài báo là phân tích tự động cấu trúc của video số nhằm tạo ra bảng mục lục và chỉ mục, giúp lưu trữ nội dung đoạn video số theo cấu trúc phân cấp và kết hợp đặc trưng thị giác và ngữ nghĩa trong truy vấn video số.

Hướng tiếp cận này giúp chúng tôi tận dụng được các ưu điểm của truy vấn dựa vào đặc trưng thị giác và truy vấn dựa vào ngữ nghĩa. Dùng đặc trưng thị giác, chúng tôi giảm bớt các khó khăn trong việc diễn đạt bằng lời trong câu truy vấn các yêu cầu liên quan đến thông tin thị giác. Dùng từ khóa để diễn đạt ngữ nghĩa, chúng tôi giảm bớt khó khăn trong việc diễn đạt yêu cầu truy vấn bằng thông tin thị giác. Ví dụ với yêu cầu rút trích các cảnh bình minh, hoàng hôn ở Sapa, chúng ta sẽ gặp khó khăn khi phải dùng từ khóa để mô tả Sapa. Ví dụ với yêu cầu rút trích các đoạn video có chứa cảnh thiên nhiên, nếu truy vấn dựa vào đặc trưng thị giác, chúng ta sẽ phải chuẩn bị nhiều dạng ảnh về thiên nhiên như bãi biển, núi đồi, đồng ruộng,..., nếu truy vấn dựa vào từ khóa, chúng ta chỉ cần gõ vào từ khóa thiên nhiên.

Để thiết kế hệ thống truy tìm video số dựa vào nội dung, cần giải quyết hai vấn đề chính là tổ chức, biểu diễn dữ liệu video số (theo cấu trúc thích hợp cho việc lưu trữ, truy vấn) và cách thức truy vấn (bao hàm cách thức nhập liệu khi truy vấn và chiến lược truy tìm). Trong việc tổ chức, biểu diễn dữ liệu video số để lưu trữ và truy vấn, chúng tôi chuẩn bị dữ liệu để truy vấn dựa vào đặc trưng thị giác và ngữ nghĩa.

Nhằm chuẩn bị dữ liệu để truy vấn dựa vào đặc trưng thị giác, chúng tôi ứng dụng có cải tiến một phương pháp giúp phân đoạn video thành các đoạn cơ sở, tự động thiết lập cấu trúc video dưới dạng bảng mục lục và chỉ mục dựa trên đặc trưng màu và đặc trưng chuyển động, điều này có sự mô phỏng gần giống với bảng mục lục và chỉ mục của một cuốn sách. Với dữ liệu nhập là đoạn video, chúng tôi cho kết xuất là bảng mục lục và chỉ mục của đoạn video đó. Sơ đồ tổng quát của quá trình tự động tạo bảng mục lục và chỉ mục :

Phân đoạn tuần tự → Tạo cấu trúc phân cấp → Tạo bảng mục lục và chỉ mục.

Về mặt tổ chức dữ liệu, thực hiện lưu trữ các thông tin theo sơ đồ phân cấp sau.

Đối với đặc trưng toàn cục, lưu trữ:

+ Đặc trưng của phần tử đại diện của lớp các đoạn cơ sở, đặc trưng của đoạn cơ sở trong lớp.

+ Đặc trưng của khung hình chính của phần tử đại diện của lớp, đặc trưng của khung hình chính của đoạn cơ sở trong lớp.

Đối với đặc trưng cục bộ, lưu trữ:

+ Đặc trưng của vùng tiêu biểu của đoạn cơ sở đại diện.

Dựa vào các đặc trưng toàn cục và cục bộ, có thể truy tìm các đối tượng tương tự trong ảnh, tuy nhiên các đặc trưng cục bộ chỉ thể hiện được sự tương tự về mặt cấu trúc của vùng chứ chưa thể hiện hết được cái mà vùng thể hiện ở mức ngữ nghĩa của ảnh. Vì vậy chúng tôi kết hợp thêm ngữ nghĩa vào việc truy vấn.

Nhằm chuẩn bị dữ liệu để truy vấn dựa ngữ nghĩa, trong mỗi lớp trong cây mục lục hoặc cây chỉ mục, chọn đoạn cơ sở đại diện, sau đó tiến hành phân đoạn các khung hình của cơ sở đại diện thành các vùng, các vùng này khác với các đối tượng không gian-thời gian vì các vùng này có thể không mang chuyển động mà chỉ cần tồn tại trong thời gian đủ lớn trong đoạn cơ sở đại diện. Sau đó các vùng tiêu biểu được nhóm lại thành các nhóm với phần tử đại diện nhóm (phần tử đại diện nhóm vùng), và mỗi đoạn cơ sở đại diện được đại diện bởi các phần tử đại diện nhóm vùng. Tiếp đó, tiến hành chú thích thủ công cho một số các đoạn cơ sở đại diện (tập mẫu), các đoạn cơ sở đại diện mới có thể được chú thích tự động dựa trên mô hình tương thích song môi trường ([3]).

Trong cách thức truy vấn, đối với cách nhập dữ liệu truy vấn và chiến lược truy tìm, người dùng có thể truy vấn dựa vào đặc trưng thị giác và ngữ nghĩa, cụ thể là khi truy vấn, người dùng có thể chọn một trong ba dạng sau:

+ Với dạng thứ nhất, người dùng có thể chọn thông tin thị giác toàn cục (ảnh tĩnh, đoạn cơ sở, đoạn video) hoặc thông tin thị giác cục bộ (các khung hình chính, các vùng tiêu biểu) và từ khóa thể hiện ngữ nghĩa để truy vấn. Hệ thống sẽ trả về các đoạn cơ sở (video) được sắp hạng dựa vào độ đo dị biệt giữa đoạn cơ sở (video) kết quả và đoạn video truy vấn, sau đó dựa vào từ khóa, hệ thống sẽ loại các kết quả không phù hợp ở bước trước.

+ Với dạng thứ hai, người sử dụng chọn thông tin thị giác toàn cục, cục bộ và từ khóa. Dựa vào từ, hệ thống trả về các đoạn video phù hợp về ngữ nghĩa truy vấn, sau đó hệ thống sắp hạng kết quả dựa vào độ đo dị biệt giữa đoạn video kết quả và đoạn truy vấn.

+ Với dạng thứ ba, kết quả truy vấn là giao của kết quả truy vấn chỉ dựa vào từ và kết quả truy vấn chỉ dựa vào thông tin thị giác.

Mục hai bài báo là các kết quả nghiên cứu có liên quan. Mục ba trình bày tổ chức dữ liệu video số. Mục bốn trình bày cách thức truy vấn dữ liệu video số. Mục năm trình bày kết quả thực nghiệm. Mục sáu là kết luận và hướng phát triển.

2. KẾT QUẢ NGHIÊN CỨU

Hiện nay các hệ truy vấn video số chưa thể đáp ứng mọi nhu cầu của người dùng, nhìn chung các hệ truy vấn video số đã trải qua các giai đoạn phát triển sau:

Trong giai đoạn thứ nhất, việc truy vấn dựa vào từ khóa và chú thích thủ công cho đoạn video. Cách tiếp cận này không khả thi về mặt thời gian và chi phí khi lượng dữ liệu video tăng trưởng với tốc độ nhanh chóng như hiện nay. Tuy nhiên, hiện nay các hệ thống truy vấn video được dùng rộng rãi trong thương mại vẫn đang sử dụng cách thức truy vấn này (ví dụ như các hệ thống tìm kiếm của Google, Yahoo,...).

Trong giai đoạn thứ hai, việc truy vấn được thực hiện dựa vào nội dung, dựa vào việc phân đoạn video thành các đoạn cơ sở, mỗi đoạn cơ sở được biểu diễn bởi một số khung hình chính. Truy vấn dựa vào đặc trưng thị giác toàn cục của đoạn cơ sở hoặc khung hình chính như các đặc trưng về màu sắc, vân, chuyển động ([5, 14]). Ưu điểm của cách tiếp cận này là có thể truy vấn trên cơ sở dữ liệu video tổng quát, không phụ thuộc ứng dụng, tuy nhiên khuyết điểm là chưa xét đến các đối tượng đóng vai trò quan trọng trong đoạn video, đồng thời chưa vượt qua được vấn đề về lỗ hổng ngữ nghĩa của đặc trưng thị giác, tức là sự tương đồng về đặc trưng thị giác không luôn bảo đảm sự tương đồng về ngữ nghĩa.

Giai đoạn thứ ba là thời kỳ của các hệ thống truy vấn dựa vào đối tượng ([2, 12]). Các

đối tượng này được gọi là các đối tượng không gian-thời gian. Ưu điểm của các hệ này là cho phép truy vấn dựa vào đối tượng, kết quả truy vấn phù hợp hơn về ngữ nghĩa đối với yêu cầu truy vấn. Khuyết điểm của các hệ này là phụ thuộc vào ứng dụng đặc thù, khó mở rộng cho cơ sở dữ liệu video tổng quát.

Giai đoạn thứ tư là thời kỳ truy vấn dựa vào phả hệ tri thức thị giác ([13]). Phả hệ tri thức thị giác thường gồm ba phần chính: tập các khái niệm về màu, tập các khái niệm về vân và tập các khái niệm về hình dáng. Tập các khái niệm này độc lập với các ứng dụng. Khi truy vấn, người dùng mô tả đối tượng truy vấn dựa vào phả hệ tri thức thị giác. Trong giai đoạn ngoại tuyến, hệ thống thực hiện việc liên kết các khái niệm thị giác trong phả hệ tri thức thị giác với các đặc trưng thị giác cấp thấp của đoạn cơ sở (video). Vì vậy khi truy vấn, hệ thống tiến hành đối sánh các khái niệm thị giác mà người dùng mô tả với các khái niệm thị giác được liên kết với dữ liệu video nhằm xác định kết quả truy vấn. Chúng ta không thể liên kết mọi ngữ nghĩa với dữ liệu video số, nhưng liên kết các khái niệm thị giác với dữ liệu video số là điều có thể thực hiện được. Ưu điểm của các hệ truy vấn này là không cần phải có trước đoạn video hoặc khung hình chính để truy vấn và cũng không cần phải chú thích trước nội dung video. Cách truy vấn này có thể áp dụng cho cơ sở dữ liệu video tổng quát. Tuy nhiên khuyết điểm của cách tiếp cận này là độ chính xác của kết quả truy vấn không cao, vì vậy nó được dùng như bước tiền lọc, và kèm theo kỹ thuật phản hồi từ người dùng. Tuy nhiên kỹ thuật phản hồi từ người dùng đòi hỏi người dùng phải xác nhận các kết quả truy vấn đúng để quá trình hệ thống hội tụ xảy ra nhanh hơn, lọc ra các kết quả với độ chính xác cao hơn, và điều này không phải lúc nào cũng nhận được sự đồng cảm từ người dùng.

Khảo sát qua các giai đoạn phát triển của các hệ truy vấn video, chúng tôi nhận thấy nếu truy vấn chỉ dựa vào đặc trưng thị giác hoặc chỉ dựa vào từ khóa thì kết quả truy vấn chưa đạt được như mong muốn, bởi vì bản chất của đặc trưng thị giác là chưa thể hiện được đầy đủ ngữ nghĩa của thông tin thị giác và bản chất của từ khóa là chưa lột tả hết nội dung của thông tin thị giác. Nói tóm lại, đoạn cơ sở (video) chứa đựng đồng thời hai đặc trưng cơ bản là đặc trưng thị giác và đặc trưng ngữ nghĩa. Vì vậy chúng tôi đề xuất một mô hình truy vấn video kết hợp đặc trưng thị giác và ngữ nghĩa.

3. TỔ CHỨC DỮ LIỆU VIDEO SỐ

Phần này sẽ trình bày các nét chính trong việc tổ chức, biểu diễn dữ liệu video số nhằm phục vụ việc truy vấn theo đặc trưng thị giác và ngữ nghĩa. Giai đoạn này gồm sáu công đoạn chính:

+ Phân đoạn tự động video số thành các đoạn cơ sở.
+ Tạo cấu trúc phân cấp (không chịu ràng buộc và có ràng buộc theo chiều thời gian).
+ Rút gọn cấu trúc phân cấp, gom nhóm các đoạn cơ sở để hình thành bảng mục lục (ràng buộc theo chiều thời gian) và chỉ mục (không ràng buộc theo chiều thời gian).
+ Xác định phần tử đại diện cho các lớp đoạn cơ sở (cơ sở đại diện), rút trích khung hình chính của cơ sở đại diện.
+ Phân đoạn cơ sở đại diện thành các vùng, gom nhóm các vùng, chọn phần tử đại diện nhóm vùng.
+ Chú thích ngữ nghĩa cho tập học gồm các cơ sở đại diện được tạo lập ở các bước trên.

Đóng góp chính của bài báo ở giai đoạn này là:
+ Phát biểu lại việc thiết lập bảng mục lục và chỉ mục ở dạng tổng quát hơn dựa vào giải thuật phân lớp phân cấp (Hierarchical Agglomerative Clustering (HAC)) có xét đến yếu tố thời gian.
+ Ứng dụng đặc trưng tự tương quan màu (AutoCorrelograms) trong công đoạn phân đoạn video thành các đoạn cơ sở, đối phó tốt với việc dịch chuyển đối tượng và camera.
+ Bên cạnh các đặc trưng toàn cục của đoạn cơ sở dựa vào các khung hình chính, chúng tôi dùng thêm đặc trưng cục bộ là các phần tử đại diện nhóm vùng, có thể dùng cho cơ sở dữ liệu video tổng quát đồng thời nếu kết hợp thêm với phương pháp chú thích tự động (được trình bày ở phần sau), nó cũng thích hợp cho các ứng dụng đặc thù.
+ Đồng thời có thể gán nhãn ngữ nghĩa tự động cho cơ sở đại diện, cấu trúc cây phân cấp (mục lục, chỉ mục) có thể được gán nhãn ngữ nghĩa tự động.

3.1. Phân tích tự động video số thành các đoạn cơ sở

Công đoạn này có ảnh hưởng rất quan trọng đến hiệu quả truy vấn. Ở đây, chúnh tôi đã kết hợp lược đồ tự tương quan màu và giải thuật Watershed để phân tích tự động đoạn video số thành các đoạn cơ sở ([8, 9]). Phương pháp này giúp tăng độ chính xác (mức độ tìm đúng) và độ trung thực (mức độ tìm sót).

Đa số các phương pháp truyền thống phân đoạn đường cong sai biệt dựa trên quan điểm chuyển cảnh. Đường cong được dò tìm để xác định các đỉnh dựa vào ngưỡng (tức dò tìm theo chiều dọc của đường cong sai biệt và đoạn cơ sở được xác định như là phần giữa các đỉnh này. Khuyết điểm xảy ra là nếu ngưỡng quá bé thì xảy ra tình trạng phát hiện dư và nếu ngưỡng quá lớn thì xảy ra tình trạng phát hiện thiếu. Bài báo dùng các toán tử hình thái học tác động vào đường cong sai biệt có tác dụng như lọc phi tuyến nhằm bám sát hình dáng đường cong sai biệt giúp loại bớt các đỉnh ảo để dò tìm đoạn cơ sở theo chiều ngang kết hợp với chiều dọc trên đường cong sai biệt ([8, 9]).

Ta sẽ thiết lập đường cong sai biệt với độ đo tương tự giữa các khung hình theo lược đồ tự tương quan màu ([4, 8]). Trong bước này, ta sử dụng lược đồ tự tương quan màu thay thế lược đồ màu nhằm khắc phục nhược điểm về tính cục bộ trong phân bố màu của lược đồ màu (2 ảnh có lược đồ màu tương tự nhưng có thể rất khác nhau) và làm lộ rõ sai biệt tại vị trí xảy ra chuyển cảnh, nhằm phát hiện chính xác hơn các đoạn cơ sở. Đặc trưng tự tương quan màu đối phó tốt với sự dịch chuyển của đối tượng và sự dịch chuyển của camera.

3.2. Tạo cấu trúc phân cấp

Dựa vào ý tưởng tạo bảng mục lục và chỉ mục, ta phát biểu bài toán ở dạng tổng quát hơn, dựa vào giải thuật gom nhóm phân cấp (Hierarchical Agglomerative Clustering (HAC)) ([8]) để liên kết các đoạn cơ sở theo cấu trúc cây. Trong giai đoạn gom nhóm của giải thuật HAC, việc chọn cặp đoạn cơ sở C_i, C_j để đối sánh phụ thuộc vào dạng liên kế thời gian. Ở đây, ta dùng hai dạng liên kết: liên kết phụ thuộc thời gian cho bảng mục lục và liên kết không phụ thuộc thời gian cho bảng chỉ mục. Thực chất, công đoạn này giúp chuẩn bị cho công đoạn gom nhóm các đoạn cơ sở có cùng đặc trưng thị giác.

3.3. Rút gọn cấu trúc phân cấp để hình thành bảng mục lục và chỉ mục

Trong giai đoạn này, cần loại bỏ các nút trung gian trong quá trình tạo lập cây phân cấp.

Nút cần loại bỏ dựa trên tiêu chuẩn xét độ dị biệt bé nhất và lớn nhất với nút cha ([8, 9]). Sau cùng, các nút lá có cùng nút cha sẽ được gom thành một nhóm. Kết quả ở công đoạn này là một tập các lớp, mỗi lớp chứa các đoạn cơ sở có độ dị biệt bé về đặc trưng thị giác, đồng thời các lớp này có thể được sắp theo chiều thời gian (để hình thành bảng mục lục) hoặc không phụ thuộc chiều thời gian (để hình thành bảng chỉ mục). Ở công đoạn này, giải thuật gom nhóm phân cấp được phát biểu lại ở dạng tổng quát hơn, có xét đến yếu tố thời gian.

3.4. Xác định cơ sở đại diện, rút trích khung hình chính của cơ sở đại diện

Cần xác định cơ sở đại diện cho mỗi nhóm các đoạn cơ sở ở công đoạn trước. Giả sử đoạn video $V = \{C_i, i = 1, 2, ..., n\}$, $C_i = \{S_j^i,\ j = 1, 2, ..., m_i\}$, S_j^i là các đoạn cơ sở. Trong mỗi lớp C_i, chọn cơ sở đại diện Rep_i, trong đó Rep_i là đoạn cơ sở trong lớp C_i, cơ sở đại diện Rep_i thỏa:

$$\sum_{j=1}^{m_i} d(\text{Rep}_i, S_j^i) \leqslant \sum_{j=1}^{m_i} d(S_k^i, S_j^i),\ \forall k \in [1, 2, .., m_i]$$

Dùng giải thuật HAC để xác định các khung hình chính cho cơ sở đại diện, kết quả của công đoạn này là cơ sở đại diện Repđược đại diện bởi tập các khung hình chính $\{F_{i,k}, k = 1, 2, ..., p\}$.

3.5. Phân đoạn cơ sở đại diện thành các vùng, gom nhóm các vùng, chọn phần tử đại diện nhóm vùng

Nhằm rút trích nội dung tiêu biểu cho đoạn cơ sở, ta dùng giải thuật phân đoạn ảnh cho các khung hình của cơ sở đại diện thành các vùng, gom nhóm các vùng thành các nhóm vùng với phần tử đại diện nhóm vùng ([10]). Các phần tử đại diện nhóm vùng này chính là các đặc trưng cấp cao của đoạn cơ sở, giúp người dùng có thể thể hiện nhu cầu truy vấn ở mức ngữ nghĩa sát với nội dung thực có trong video, giúp bắt thêm nhịp cầu nối giữa đặc trưng cấp thấp của video và ngữ nghĩa. Cách tiếp cận này có một số ưu điểm hơn so với cách tiếp cận xác định các đối tượng không gian-thời gian dựa vào vector chuyển động, vì xác định các đối tượng không gian-thời gian thường vấp phải khó khăn đối với các dạng video có nội dung ít chuyển động và sự thay đổi của camera cũng như khó mở rộng cho các dạng video tổng quát.

Giả sử có tập các cơ sở đại diện $\{\text{Rep}_i,\ i = 1, 2, ..., N\ \text{Re}\ p\}$, mỗi Rep_i có tập các vùng $R_i = \{r_{ij,\ j=1,..NR_i}\}$, tập vùng này được gom thành tập các nhóm vùng được biểu thị bởi $CLUS = \{clus_k,\ k = 1..NC\}$ và mỗi nhóm vùng $clus_k$ được xác định một phần tử đại diện nhóm vùng được biểu thị bởi reg_k. Các vùng của mỗi cơ sở đại diện sẽ được thay thế bởi phần tử đại diện nhóm vùng gần nhất. Kết quả của công đoạn này là mỗi cơ sở đại diện sẽ được biểu diễn bởi một tập các phần tử đại diện nhóm vùng.

3.6. Chú thích ngữ nghĩa cho tập mẫu gồm tập các cơ sở đại diện

Ta tiếp cận truy tìm video dựa vào ngữ nghĩa theo hướng chú thích tự động video. Hướng tiếp cận này giúp chú thích tri thức vào cơ sở dữ liệu video một cách tự động và hệ thống truy vấn sẽ ngày càng thông minh hơn với công cụ chú thích video tự động.

Mục đích của việc chú thích đoạn video tự động là cho trước một đoạn cơ sở S chưa được chú thích, chọn tự động tập từ khóa từ tập từ vựng cho trước nhằm mô tả S tốt nhất. Việc

chú thích này dựa vào tập học, tức tập video được chú thích trước. Tập này được xác định như sau:

$$VS = \{(S_1, W_1), ..., (S_N, W_N)\}.$$

Chọn hướng tiếp cận chú thích video dựa vào mô hình tương thích song môi trường [3] thường được dùng trong ngôn ngữ. Thay vì khảo sát việc truy vấn bằng ngôn ngữ này để truy tìm ra tài liệu được viết bởi ngôn ngữ khác, chúng tôi khảo sát việc truy vấn bằng ngôn ngữ để truy tìm ra đoạn video được viết bởi "ngôn ngữ" đặc trưng thị giác.

Sau khi có được các nhóm vùng tại bước trước, ta tiến hành chú thích thủ công tập học gồm các cơ sở đại diện để liên kết ngữ nghĩa với video. So với [3], tác giả đã thực hiện hai cải tiến chính là phân đoạn, phân nhóm vùng dựa vào giải thuật HAC và chú thích ngữ nghĩa sau khi phân đoạn, phân nhóm.

4. CÁCH THỨC TRUY VẤN DỮ LIỆU VIDEO SỐ

4.1. Phương pháp nhập liệu khi truy vấn

Bài báo trình bày phương pháp nhập liệu kết hợp đặc trưng thị giác với ngữ nghĩa.

4.1.1. Truy vấn dựa vào đặc trưng thị giác

Người dùng nhập liệu theo các cách sau:

Cách 1: Nhập đoạn cơ sở.

Đoạn đoạn cơ sở truy vấn được biểu diễn bởi tập các khung hình chính hoặc tập các phần tử đại diện nhóm (phần tử đại diện nhóm vùng của riêng đoạn cơ sở này), các khung hình chính hoặc các phần tử đại diện nhóm này sẽ được dùng trong quá trình truy vấn.

Cách 2: Chọn một số phần tử đại diện nhóm của đoạn cơ sở truy vấn.

Người dùng có thể chọn và kết hợp một số phần tử đại diện nhóm (không nhất thiết dùng toàn bộ) theo toán tử luận lý AND, OR, NOT. Dạng truy vấn tiêu biểu có dạng:

(Sel(rep$_i$) OR Sel(rep$_j$)) AND (Sel(rep$_k$) OR Sel(rep$_l$)) AND NOT (Sel(rep$_m$) OR Sel(rep$_n$))

Sel(rep) trả về giá trị True nếu vùng *rep* được chọn.

Cách 3: Chọn các phần tử đại diện nhóm vùng của cơ sở dữ liệu video, với cách nhập liệu này, người dùng không cần có trước đoạn cơ sở truy vấn.

4.1.2. Truy vấn dựa vào ngữ nghĩa

Người dùng có thể thể hiện câu truy vấn dưới dạng các từ khóa. Cho trước câu truy vấn $Q = w_1, ..., w_k$ và tập các đoạn cơ sở, mục đích của bài toán là tìm các đoạn cơ sở $S = \{reg_1 ... reg_m\}$ thích hợp với câu truy vấn Q.

4.1.3. Truy vấn dựa vào đặc trưng thị giác và ngữ nghĩa

Đoạn video mang hai đặc trưng chính là đặc trưng thị giác và ngữ nghĩa. Đồng thời như chúng ta đã biết, hình ảnh và ngôn ngữ là hai dạng thông tin quan trọng trong giao tiếp giữa người với người và giữa người với máy, vì vậy chúng tôi mong muốn kết hợp chúng trong việc truy vấn video. Dữ liệu nhập là đoạn video và câu truy vấn. Quá trình truy vấn có thể diễn ra theo các cách sau.

Cách 1: Truy vấn dựa vào đặc trưng thị giác, lọc lại kết quả dựa vào ngữ nghĩa. Nhập đoạn cơ sở truy vấn $S_Q = \{reg_1 ... reg_m\}$ và các từ truy vấn $Q = w_1, ..., w_k$.

Cách 2: Truy vấn dựa vào ngữ nghĩa, lọc lại kết quả dựa vào đặc trưng thị giác.

Nhập đoạn các từ truy vấn $Q = w_1, ..., w_k$ và cơ sở truy vấn $S_Q = \{reg_1 ... reg_m\}$.

Khi chú thích ngữ nghĩa vào tập học (tập các cơ sở đại diện), người ta thường chỉ chú trọng đến ngữ nghĩa, không chú thích chi tiết các đặc trưng về thị giác. Vì vậy cách truy vấn này giúp khắc phục các khuyết điểm của hai cách truy vấn trên, đồng thời có thể áp dụng hiệu quả vào các ứng dụng đặc thù.

4.2. Chiến lược truy tìm

4.2.1. Đối với truy vấn dựa vào đặc trưng thị giác

Với phần tử đại diện nhóm rep_q có được từ cách nhập liệu ở Mục 4.1.1, chọn nhóm có phần tử đại diện nhóm vùng gần nhất với nó, biểu thị nhóm vùng đó là $Clus_r$ và phần tử đại diện nhóm vùng là reg_r.

Biểu thị tập các đoạn cơ sở có ít nhất một vùng thuộc $Clus_r$ là $SClus_r$. Tập kết quả các đoạn cơ sở từ dạng truy vấn:

(Sel(rep$_i$) OR Sel(rep$_j$)) AND (Sel(rep$_k$) OR Sel(rep$_l$)) AND NOT (Sel(rep$_m$) OR Sel(rep$_n$)) sẽ là: *(Sclus$_i$∪ SClus$_j$∪ Clus$_l$)\ (Sclus$_m$∪SClus$_n$).*

Sau cùng kết quả được sắp hạng dựa vào độ đo dị biệt giữa hai đoạn cơ sở được đại diện bởi các phần tử đại diện nhóm vùng ([10]). Mặc dù truy vấn dựa vào phần tử đại diện nhóm vùng cũng giúp thể hiện một phần ngữ nghĩa trong truy vấn, tuy nhiên sự tương đồng về đặc trưng thị giác không nhất thiết dẫn đến sự tương đồng về ngữ nghĩa giữa đoạn cơ sở truy vấn và đoạn cơ sở kết quả. Vì vậy chúng tôi cho phép người dùng kết hợp truy vấn dựa vào đoạn cơ sở và từ khóa nhằm lọc ra kết quả vừa tương đồng về mặt thị giác vừa tương đồng về mặt ngữ nghĩa.

4.2.2. Đối với truy vấn dựa vào ngữ nghĩa

Đối với dạng truy vấn ở Mục 4.1.2, để truy tìm các đoạn cơ sở phù hợp, chúng tôi dùng phân bố xác suất $P(Q|S)$ để xác định mức độ tương thích của đoạn cơ sở với câu truy vấn, đại lượng này được xác định như sau:

$$P(Q|S) = \sum_{j=1}^{k} P(w_j|S),$$

trong đó $P(w_j|S)$ được tính theo các phương trình trong [3, 11].

Dựa vào $P(Q|S)$ ta có thể sắp hạng kết quả truy vấn.

4.2.3. Đối với truy vấn dựa vào đặc trưng thị giác và ngữ nghĩa

Cách 1: Dựa vào truy vấn theo cách 1 ở Mục 4.1.3, dùng phương pháp truy vấn video dựa vào đặc trưng thị giác ở Mục 4.2.1, ta có tập kết quả là các đoạn cơ sở RS. Dùng phương pháp truy vấn video dựa vào ngữ nghĩa ở Mục 4.2.2, sắp hạng kết quả tìm được dựa vào đại lượng $P(Q|S) = \Pi_{j=1}^{k} P(w_j|S)$

Cách 2: Dựa vào truy vấn theo cách 2 ở Mục 4.1.3, dùng phương pháp truy vấn video dựa ngữ nghĩa ở Mục 4.2.2, ta có tập kết quả các đoạn cơ sở RS. Dùng phương pháp truy vấn video dựa vào thị giác ở Mục 4.2.1, sắp hạng kết quả tìm được dựa vào độ đo dị biệt giữa đoạn cơ sở truy vấn S_Q và đoạn cơ sở thuộc tập RS ([10]).

5. KẾT QUẢ THỰC NGHIỆM

Phương pháp truy vấn đã được trình bày có thể áp dụng cho các thể loại video, không phụ thuộc ứng dụng. Tuy nhiên trong khuôn khổ một bài báo, chúng tôi trình bày một số kết quả thực nghiệm dựa trên bộ dữ liệu video số thuộc thể loại: thế giới loài vật, phim hoạt hình, thể thao, tin tức-thời sự.

Hệ thống truy vấn video số của chúng tôi có thể thực hiện các tác vụ sau:

+ Với dữ liệu nhập là một đoạn video, chúng tôi trả về cấu trúc mục lục và chỉ mục của đoạn video đó, đồng thời có thể chú thích tự động cho các đoạn cơ sở trong cấu trúc mục lục và chỉ mục vừa tạo.

+ Giúp người dùng có thể diễn đạt ý tưởng truy vấn gần với lối nghĩ của con người thông qua các phần tử đại diện nhóm vùng và các từ khóa.

Trong giai đoạn ngoại tuyến, tổ chức và biểu diễn dữ liệu video số được tiến hành cụ thể như sau:

+ Phân đoạn tự động video thành các đoạn cơ sở: Dựa vào đặc trưng màu và đặc trưng chuyển động [8].

+ Tạo cấu trúc phân cấp, rút gọn thành bảng mục lục và chỉ mục.

+ Xác định cơ sở đại diện, rút trích khung hình chính của cơ sở đại diện:

- Rút trích đặc trưng của đoạn cơ sở gồm 2 đặc trưng màu và 2 đặc trưng chuyển động [8].

- Rút trích đặc trưng của khung hình chính gồm đặc trưng về màu như lược đồ tự tương quan màu, đặc trưng về vân như lược đồ hệ số góc (Edge direction histogram) và vectơ chuyển động [4,8].

+ Phân đoạn cơ sở đại diện thành các vùng, gom nhóm các vùng, chọn phần tử đại diện nhóm vùng.

+ Rút trích đặc trưng của phần tử đại diện nhóm vùng gồm đặc trưng về màu, vân, hình dáng ([10]).

+ Chú thích ngữ nghĩa cho tập học gồm các cơ sở đại diện đã được tạo lập.

Chúng tôi tiến hành khảo sát bộ dữ liệu gồm 12 đoạn video số, được phân đoạn tự động thành 5280 đoạn cơ sở, các đoạn cơ sở này được gom nhóm thành 78 lớp (theo cấu trúc chỉ mục) với 78 cơ sở đại diện, các cơ sở đại diện này được phân tích thành 148 nhóm vùng ảnh với 148 phần tử đại diện nhóm vùng.

Để thực hiện việc truy vấn dựa vào ngữ nghĩa, chúng tôi chuẩn bị bộ từ vựng huấn luyện gồm 487 từ được tổ chức theo phả hệ tri thức (Ontology) được liệt kê một số từ đại diện như sau (bộ từ vựng này hoàn toàn có thể được mở rộng cho các ứng dụng riêng biệt):

Thiên nhiên:

Bầu trời	Núi	Sông	Đất nền	...
Mây	Đồi	Suối	Cát	...
Hoàng hôn	Cây xanh	Thác	Thảm cỏ	...
Bình minh	Rừng	Bãi biển	Nền lá	...

Động vật:

Sư tử	Chim	Cá	Người	...
Cọp	Thiên nga	Cá heo	Khỉ	...
Gấu	Vịt trời	Cá sấu	Rắn	...
Ngựa	Bồ câu	Ba ba	Khủng long	...

Thể thao:

Khán đài	Bóng đá Sân cỏ	Bóng rổ	...
Khung thành	Bóng chuyền	...	
Cầu thủ	Bóng nước	...	

Tin tức-thời sự:

Phát thanh viên	Lâm nghiệp	...
Thành phố	Thiên tai	...
Nông nghiệp	Du lịch	...
Ngư nghiệp	Hoả hoạn	...

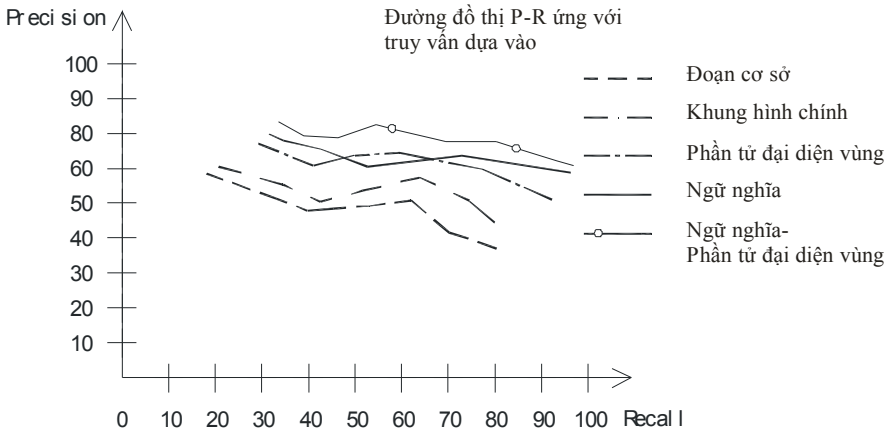
	Độ chính xác Trung bình (%)	Độ trung thực Trung bình (%)	**Mức độ hiệu quả Đối với người dùng (theo ý kiến người dùng)**
1. Phương pháp truy vấn dựa vào đặc trưng thị giác của đoạn cơ sở	58	67	Kết quả truy vấn : - Tìm được đoạn cần tìm, tuy nhiên kết quả sai còn nhiều Thời gian truy vấn : - Nhanh
2. Phương pháp truy vấn dựa vào đặc trưng thị giác của các khung hình chính	62	69	Kết quả truy vấn : - Có thể tìm được nhiều đoạn cần tìm mà phương pháp 1 không tìm được , tuy nhiên kết quả sai còn nhiều Thời gian truy vấn : - Nhanh
3. Phương pháp truy vấn dựa vào các ptđdn	74	78	Kết quả truy vấn : - Tìm được đoạn cần tìm, nội dung của đoạn cần tìm phù hợp hơn về ngữ nghĩa cần truy vấn. Dễ dàng thể hiện yêu cầu truy vấn so với phương pháp 1 và 2. Tuy nhiên một số ptđdn trong cơ sở dữ liệu không thể hiện rõ nét ngữ nghĩa. Thời gian truy vấn : - Nhanh
4. Phương pháp truy vấn dựa ngữ nghĩa	79	81	Kết quả truy vấn : - Độ chính xác cao, tuy nhiên còn bị phụ thuộc vào kho từ vựng. Thời gian truy vấn : - Nhanh
5. Phương pháp truy vấn kết hợp đặc trưng thị giác và ngữ nghĩa	83	80	Kết quả truy vấn : - Độ chính xác cao, kết quả truy vấn phù hợp về ngữ nghĩa và đặc trưng thị giác. Thời gian truy vấn : - Nhanh

Ta sử dụng 2 đại lượng: độ chính xác và độ trung thực để đánh giá tính hiệu quả của hệ thống.

Độ chính xác = (Số đoạn cơ sở tìm được đúng / Số đoạn cơ sở tìm được).

Độ trung thực = (Số đoạn cơ sở tìm được đúng / Tổng số đoạn cơ sở đúng thực có)

Thực hiện việc truy vấn dựa trên bộ dữ liệu gồm 5280 đoạn cơ sở.



Hình 1. Đường cong thể hiện mối quan hệ giữa độ chính xác và độ trung thực với các phương pháp truy vấn khác nhau

Kết quả truy vấn cho thấy mô hình kết hợp truy vấn dựa vào đặc trưng thị giác và ngữ nghĩa đem lại kết quả tốt hơn các với các phương pháp truyền thống chỉ dựa vào đặc trưng thị giác toàn cục, cục bộ hoặc chỉ dựa vào ngữ nghĩa.

6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo đã trình bày một mô hình truy tìm dữ liệu video số. Mô hình này tập trung vào hai vấn đề then chốt trong việc truy tìm, đó là tổ chức, biểu diễn dữ liệu video số và cách thức truy vấn nhằm hy vọng từng bước đạt được ba mục tiêu sau:

+ Mục tiêu thứ nhất là xây dựng mô hình truy vấn video sao cho mô hình này là sự mở rộng tự nhiên của mô hình truy vấn ảnh.

+ Mục tiêu thứ hai là mô hình có thể áp dụng cho dữ liệu video tổng quát lẫn đặc thù.

+ Mục tiêu thứ ba là kết quả truy vấn phù hợp với yêu cầu về ngữ nghĩa và thị giác của người dùng.

Kết quả thực nghiệm cho thấy mô hình có triển vọng khả quan. Việc kết hợp xây dựng phả hệ tri thức thị giác và chú thích tự động đoạn video để có thể truy vấn không cần dùng đến đoạn cơ sở đồng thời tận dụng đặc trưng mang tính thời gian như đặc trưng chuyển động trong truy vấn sẽ được trình bày trong những công trình tiếp theo.

TÀI LIỆU THAM KHẢO

[1] D. Bimbo, *Visual Information Retrieval,* Morgan Kaufmann, 1999.

[2] D. Zhong and S. F. Chang, Video object model and segmentation for content-based video indexing, *Proc.IEEE International Conference on Circuits and Systems* '97, (Hong Kong), June 1997.

[3] Jeon, V. Lawrenko, R. Mammatha, Automatic image annotation and retrieval using cross-media relevance models, *SIGIR '03: ACM* (2003).

[4] J. Huang, S. R. Kumar, M. Mitra, W. J. Zhu, R. Zabih, Image indexing using color correlograms, *IEEE International Conference on Image Processing (ICIC '01):IEEE*, 2001.

[5] Jianping Fan, Ahmed K. Elmagarmid, Xingquan Zhu, Walid G.Aref, and Lide Wu, Classview: hierarchical video shot classification, indexing, and accessing, *IEEE Transactions on Multimedia* **6** (1) (2004).

[6] M.Yeung and B. L.Yeo, Time-constrained clustering for segmentation of video into story units, *13th International Conference on Pattern Recognition*,(Vienna), August 1996 (375–380).

[7] Nguyễn Lãm, Lý Quốc Ngọc, Phân tích tự động dữ liệu video số dựa trên mô hình phân cấp dữ liệu, *Tạp chí Tin học và Điều khiển học* **21** (1) (2005).

[8] Nguyễn Lãm, Lý Quốc Ngọc, Phan Vĩnh Phước, Nguyễn Văn Kỷ Cang, Nguyễn Quốc Tuấn, Phân tích tự động dữ liệu video số hỗ trợ truy tìm thông tin thị giác dựa vào nội dung, *Tạp chí Phát Triển Khoa học và Công Nghệ Đại học Quốc Gia Tp.HCM* **8** (4) (2005).

[9] P. Salembier, L. Garrido, Binary partition tree as an efficient representation for image processing, segmentation and information retrieval, *IEEE Transactions on Image Processing* **9** (4) (2000) 561–576.

[10] Quoc Ngoc Ly, Anh Duc Duong, Hierarchical data model in content-based image retrieval, *International Journal of Information Technology, International Conference on Intelligent Computing (ICIC2005)* August 23-28, 2005 (Hefei, China).

[11] Quoc Ngoc Ly, Anh Duc Duong, Thach Thao Duong, Duc Thanh Ngo, Image retrieval based on visual information concept and automatic image annotation, *The First International Conference on Theories and Applications of Computer Science (ICTACS 2006),* Ho Chi Minh City, VietNam, August 3-5, 2006.

[12] S. F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, VideoQ-an automatic content-based video search system using visual cues, *Proc.ACM Multimedia Conf.* (Seattle, WA), November 1997.

[13] V. Mezaris, I. Kompatsiaris, M. G. Strintzis, Region-based image retrieval using an object ontology and relevance feedback, *Eurasip Journal on applied signal processing* (6) June 2004 (886–901).

[14] W. AL-Khatib, Y. F. Day, A. Ghafoor, P. B. Berra, Semantic modeling and knowledge representation in multimedia databases, *IEEE Transactions on Knowledge and Data Engineering* **11** (1) (1999) 64–80.

[15] W. Wolf, Keyframe selection by motion analysis, *ICASSP* Vol II, May 7-10 1996 (1228–1231).

Nhận bài ngày 15 - 9 - 2005
Nhận lại sau sửa ngày 5 - 3 -2007