

Truy xuất Video Dựa trên Đa đặc trưng

Mai Tiến Dũng

Trường Đại học Công nghệ thông tin, Đại học Quốc Gia TP.HCM

`dungmt@uit.edu.vn`

Tóm tắt. Truy xuất video là một trong những hướng nghiên cứu mới và có nhiều thách thức. Trong bài viết này, chúng tôi xin trình bày những khảo sát của chúng tôi về các nghiên cứu liên quan đến truy xuất video dựa trên khái niệm và định hướng nghiên cứu mà chúng tôi dự định thực hiện trong đề tài.

Từ khóa. tìm kiếm video, truy xuất video, đa đặc trưng, multimodal features, concept-based video retrieval, semantic concept.

1 Giới thiệu

Nhờ sự phát triển của công nghệ số, con người đã tạo ra một khối lượng dữ liệu đa phương tiện rất lớn, liên tục gia tăng về số lượng, đa dạng cả về nội dung, hình thức và định dạng lưu trữ. Trong đó dữ liệu video được xem như một dữ liệu đa phương tiện phức tạp vì chúng bao gồm các dữ liệu khác như văn bản, âm thanh, hình ảnh, ... Trong những năm gần đây, việc tạo ra các tập tin video càng ngày càng dễ hơn, mọi người đều có thể dùng thiết bị di động hoặc máy quay kỹ thuật số để ghi hình mà không cần phải sử dụng các phần mềm hay các thiết bị chuyên nghiệp như trước đây, đồng thời có thể chia sẻ với bạn bè thông qua các trang web chia sẻ dữ liệu trực tuyến.

Chẳng hạn theo thống kê ¹ tại thời điểm 01/2012 trung bình trong một phút có hơn 136,000 ảnh được upload lên trang web chia sẻ hình ảnh flicker, có 138MB dữ liệu các loại được upload lên facebook, và có hơn 360 giờ video được upload lên youtube, ... Tuy nhiên, để có thể sử dụng và khai thác hiệu quả khối lượng dữ liệu khổng lồ này đòi hỏi phải có những phương pháp tổ chức lưu trữ, lập chỉ mục và truy xuất thật sự hiệu quả về nhiều mặt: tài nguyên sử dụng, tốc độ thực thi và đặc biệt là khả năng đáp ứng yêu cầu của người dùng ...

Theo Jain et al. [14], video được tạo ra nhằm mục đích giải trí, thông tin, truyền thông, hoặc phân tích dữ liệu, ... điều đó cũng có nghĩa là video được tạo ra đáp ứng cho nhiều loại đối tượng người dùng khác nhau: một khách hàng xem các quảng cáo để chọn mua sản phẩm, người yêu thích bóng đá có thể xem lại cảnh ghi bàn của cầu thủ trong một trận đấu, người yêu thích phim ảnh có xem các đoạn giới thiệu phim, ... Hiện nay các website thương mại như Youtube

¹ <http://thesocialskinny.com/100-social-media-statistics-for-2012/>

², Vimeo³, Dailymotion⁴, blip.tv⁵ ... có khả năng đáp ứng tốt các nhu cầu trên và cho phép người dùng có thể xem toàn bộ hay xem lướt qua nội dung của tập tin video được lưu trữ trong hệ thống. Tuy nhiên, do có quá nhiều dữ liệu nên người dùng mong muốn truy xuất những đoạn video có chứa nội dung phù hợp với yêu cầu. Chẳng hạn người dùng đang tìm hiểu thông tin về diễn viên A nên muốn tìm những đoạn video có chứa diễn viên A, hoặc người dùng muốn xem các đoạn video có chim cánh cụt,... Đây là một trong những bài toán khó đang được cộng đồng nghiên cứu quan tâm, những kết quả bước đầu cần được tiếp tục nghiên cứu và cải tiến. Trong định hướng nghiên cứu, chúng tôi sẽ tập trung vào bài toán xác định các khái niệm có trong video để làm cơ sở cho việc truy xuất video.

2 Những Nghiên cứu Liên quan

2.1 Khái niệm trong Video

Quá trình nhận thức của con người về nội dung video là một quá trình tác động lẫn nhau giữa các khái niệm mà người đó nhận thức được, vì thế khi muốn truy xuất video, người dùng thường mô tả các nội dung cần tìm qua các từ khóa liên quan. Do đó, các hệ thống truy xuất video phải có khả năng nhận diện được các khái niệm có trong video, tổ chức lưu trữ và lập chỉ mục chúng sao cho người dùng có thể truy xuất các đoạn video thông qua các khái niệm ngữ nghĩa được mô tả như những từ khóa tìm kiếm.

Bài toán quan trọng nhất ở đây chính là làm thế nào xác định và gán các khái niệm cho tất cả các đối tượng, sự vật, hiện tượng,... xuất hiện trong video.

Có nhiều phương pháp đã được đưa ra, về cơ bản có thể chia thành hai phương pháp: thủ công và tự động.

Phương pháp Thủ công

Sau khi xem qua nội dung của video, người dùng sẽ xác định và gán các khái niệm ngữ nghĩa cho video đó. Có thể chia thành hai nhóm:

- Do các chuyên gia thực hiện [7,10], tuy nhiên công việc gán nhãn thường nhàm chán và mất nhiều thời gian, nên toàn bộ video chỉ được mô tả ngắn gọn.
- Do người dùng thực hiện, chẳng hạn trong các trang YouTube, Flickr, Facebook,... các khái niệm được gán bởi nhiều người qua các hình thức như các chú thích, các trò chơi nhận diện và gán khái niệm cho đối tượng [1,2] hay công cụ cho phép người dùng gán nhãn [20],... Vì có nhiều người cùng tham gia nên có thể dẫn đến sự nhập nhằng giữa các khái niệm được xác định, các khái niệm thường mang tính cá nhân, và có thể không đầy đủ,...

² <http://www.youtube.com>

³ <http://www.vimeo.com>

⁴ <http://www.dailymotion.com>

⁵ <http://blip.tv>

Tuy nhiên, với khối lượng dữ liệu lớn và không ngừng gia tăng, việc thực hiện thủ công là không khả thi, các khái niệm được gán không thể phản ánh hết các nội dung trong video làm cho việc truy xuất các đoạn video sẽ không hiệu quả [22].

Phương pháp Tự động

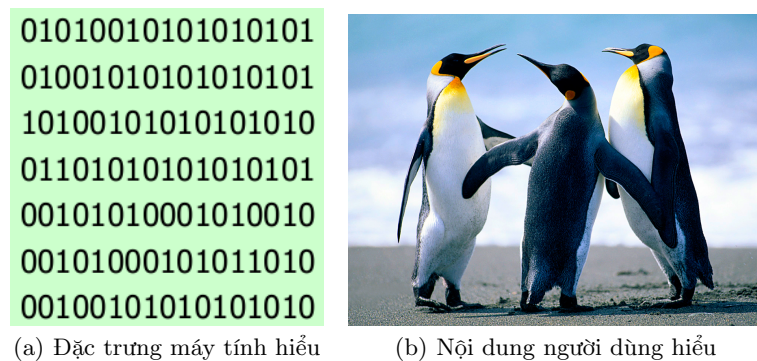
Mục đích của phương pháp này là tự động xác định và gán các khái niệm có trong video thông qua các bộ mô tả (descriptors). Trong các hệ thống tìm kiếm trên video hiện nay như Youtube, Video, Baidu, Blinkx,... để trả về các video kết quả dựa trên các khái niệm do người dùng nhập vào dưới dạng văn bản. Các bộ mô tả chủ yếu dựa vào tên file video, các thông tin mô tả về video khi người dùng upload lên hệ thống, các văn bản trong tài liệu có nhúng video vào (surrounding text), social tags, văn bản chú thích trong video (closed captions), hay sử dụng kỹ thuật nhận dạng tiếng nói để chuyển âm thanh thành văn bản (speech transcript), các metadata trong tập tin video,... Tuy nhiên, người dùng thường upload tập tin video lên hệ thống và chia sẻ video thông qua địa chỉ liên kết của video đó, vì thế họ rất ít khi mô tả nội dung của video khi upload; ngoài ra, tên file của video thường không mô tả về video, có nhiều nguyên nhân nhưng chủ yếu do người dùng upload trực tiếp các file được ghi hình bằng các máy kỹ thuật số hay điện thoại di động. Việc sử dụng văn bản chú thích có những thành công đáng kể [3] nhưng hạn chế lớn nhất khi các văn bản chú thích không liên quan đến nội dung video (chẳng hạn các dòng quảng cáo hay các dòng tin ngắn). Các hệ thống truy xuất video dựa trên văn bản được rút trích từ các phần mềm nhận dạng tiếng nói đặc biệt hiệu quả trên dữ liệu tin tức, phỏng vấn, bài thuyết trình,... nhưng trường hợp tiếng nói trong video không sử dụng ngôn ngữ Tiếng Anh như Trung Quốc, Arap,... là một thách thức và kết quả tìm kiếm thường có độ chính xác không cao.

Dữ liệu video là một dữ liệu đa phương tiện phức tạp và mang yếu tố giác quan (ảnh, âm thanh, video), nên cần thiết phải áp dụng các kỹ thuật xử lý ảnh, âm thanh và thị giác máy tính để xác định sự hiện diện các khái niệm trong video. Thông thường người ta chia video thành hai dòng dữ liệu là dòng âm thanh (audio stream) và dòng hình ảnh (visual stream). Vì thế ngoài việc rút trích các thông tin từ dòng âm thanh, thì các thông tin từ việc xử lý hình ảnh rất quan trọng. Vì thế một số lượng lớn các nghiên cứu đặt vai trò quan trọng của nội dung hình ảnh (visual content) lên hàng đầu chứ không phải văn bản hay các đặc trưng khác được rút trích từ dòng âm thanh.

Việc phân tích xác định nội dung của dữ liệu hình ảnh có một lịch sử lâu đời [19], xuất phát từ những năm 1960s. Với một vài thành công bước đầu, các nhà nghiên cứu trong những năm 1970s cho rằng bài toán hiểu nội dung dữ liệu hình ảnh sẽ sớm được giải quyết hoàn toàn. Tuy nhiên, đến những năm 1980s người ta cho rằng những dự đoán này quá lạc quan. Trong những năm 1990s, truy xuất ảnh dựa trên nội dung là một trong những hướng nghiên cứu mới và nhận được nhiều sự quan tâm của các nhóm nghiên cứu trên thế giới, mục tiêu là phát triển các phương pháp tìm kiếm ảnh trong tập dữ liệu lớn dựa trên nội dung.

Đầu những năm 2000s, nghiên cứu trong lĩnh vực truy xuất dựa trên nội dung đã có một sự chuyển đổi từ ảnh sang video [8] [4] [5] [32] [34]. Đặc điểm chung của những phương pháp này là dựa trên các đặc trưng mức thấp như màu sắc, cấu trúc văn, hình dáng và không gian - thời gian. Hầu hết những hệ thống dựa trên truy vấn theo mẫu (query-by-example), người sử dụng sẽ vẽ phác thảo hay cung cấp ảnh mẫu cần truy vấn, chỉ những ảnh có sẵn trong hệ thống và có một sự khác nhau không đáng kể so với ảnh truy vấn được chọn làm kết quả. Nếu ảnh chưa có trong hệ thống thì các phương pháp truy vấn ảnh dựa trên nội dung sẽ không hiệu quả. Ngoài ra, để diễn đạt nội dung cần tìm dựa trên các đặc trưng mức thấp là rất khó khăn đối với người dùng, trong một số trường hợp người dùng không nhớ chính xác nội dung cần tìm. Vì thế nhu cầu cấp thiết phải phát triển các hệ thống cho phép người dùng nhập vào các văn bản mô tả về nội dung cần tìm và hệ thống trả sẽ về kết quả tương ứng, nói cách khác, người dùng mong muốn tìm kiếm dựa trên sự tương đồng về ngữ nghĩa giữa những từ khóa nhập vào với các đặc trưng mức thấp mà máy tính lưu trữ và có thể xử lý được. Đây là một bài toán khó và có nhiều thách thức. Có hai khó khăn lớn nhất:

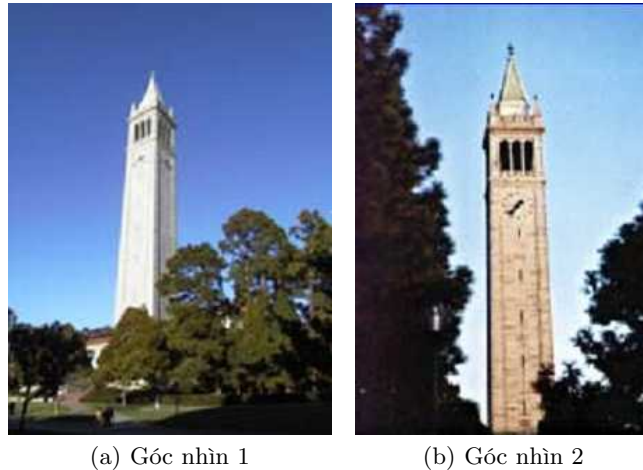
- Có sự khác biệt khá lớn giữa những đặc trưng mức thấp mà máy tính rút trích từ video với những khái niệm mức cao mà con người hiểu về video đó, sự khác biệt này được gọi là sự khác biệt về ngữ nghĩa (semantic gap) [23]. Chẳng hạn trong hình 1, con người có thể nhìn và hiểu bức ảnh hình chim cánh cụt nhưng với máy tính thì chúng chỉ là những bit 0 và 1



Hình 1. Hình ảnh minh họa sự khác biệt về ngữ nghĩa

- Một trong những nguyên nhân gây ra sự khác biệt về ngữ nghĩa đó là cùng một đoạn video nhưng những người khác nhau có thể có những cách hiểu khác nhau, đặc biệt là các video về cảm xúc hay tình cảm [9]. Thông thường những video có nhiều cách hiểu khác nhau hay có thể hiểu theo sự chủ quan của từng người ít được xem xét. Tuy nhiên, để xác định các khái niệm mang tính khách quan có hiện diện trong video hay không là một bài toán

khó, chẳng hạn chúng ta muốn xác định tòa nhà có trong video hay không. Một trong những khó khăn chính đó là cùng một khái niệm nhưng có nhiều thể hiện khác nhau: góc nhìn khác nhau, điều kiện ánh sáng hay tỉ lệ khác nhau, ... sự khác biệt này được gọi là sự khác biệt về giác quan (sensory gap). Chẳng hạn trong hình 2, cùng một tòa tháp nhưng hai hình khác biệt nhau do góc nhìn khác nhau.



Hình 2. Minh họa sự khác biệt về giác quan của cùng một khái niệm [27]

Tóm lại, có hai bài toán mà chúng ta cần giải quyết:

- Xác định sự hiện diện của các khái niệm mang tính khách quan c_i có trong các đoạn video s_i hay không.
- Xác định khái niệm c_i trong những điều kiện khác biệt về giác quan.

2.2 Những Nghiên cứu Liên quan đến Khái niệm trong Video

Xác định các khái niệm ngữ nghĩa trong video là một trong những hướng nghiên cứu mới, nhiều thách thức trong lĩnh vực truy xuất video. Cho đến thời điểm hiện nay, theo tìm hiểu của chúng tôi thì chưa có một hệ thống nào có thể xác định được tất cả các khái niệm có trong thế giới thực, các hội thảo hàng đầu trong lĩnh vực này thường cho trước danh sách các khái niệm và tập dữ liệu huấn luyện có liên quan. Mục tiêu đặt ra cho các nhóm nghiên cứu là xác định các khái niệm có hiện diện trong tập kiểm tra. Chẳng hạn: trong TRECVID 2012⁶, số khái niệm cần xác định là 500; imageclef2012⁷ sử dụng 94 khái niệm; PASCAL-VOC 2012⁸ sử dụng 20 khái niệm.

⁶ <http://www-nlpir.nist.gov/projects/tv2012/tv2012.html>

⁷ <http://www.imageclef.org>

⁸ <http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2012/index.html>

Một số công trình nghiên cứu ban đầu dựa trên các luật được xây dựng bằng tay để nhận diện các khái niệm riêng lẻ [30,31,36]. Tuy nhiên, khi kích thước dữ liệu tăng, số lượng khái niệm cần xác định lớn hơn thì những phương pháp này không khả thi vì cần quá nhiều luật. Khi đó chúng ta cần phải có các phương pháp ở mức tổng quát hơn, cho phép xác định được nhiều khái niệm và có thể áp dụng trên tập dữ liệu lớn. Để giải bài toán này, thông thường ta áp dụng các phương pháp máy học để xác định mối liên quan giữa các đặc trưng mức thấp với các khái niệm ngữ nghĩa ở mức cao [17,24,28,29]. Ta có thể tổng quát thành bài toán như sau: cho trước một vector đặc trưng x_i biểu diễn đặc trưng cho một đoạn video, mục tiêu là dùng x_i để xác định khái niệm c_i có hiện diện trong đoạn video đó hay không. Có nhiều phương pháp rút trích đặc trưng x_i và nhiều phương pháp máy học được áp dụng để xác định mối quan hệ giữa x_i và c_i .

Chọn Đặc trưng

Các đặc trưng được xây dựng phải hướng đến mục tiêu giảm khoảng cách khác biệt về ngữ nghĩa và có thể nhận diện được khái niệm trong điều kiện có sự khác biệt về giác quan. Do video là một dữ liệu phức tạp, chúng có thể chứa văn bản, âm thanh, hình ảnh, ... nhưng sự khác biệt giữa video với các loại dữ liệu khác đó là video gồm một dãy hình liên tục theo thời gian để tạo sự chuyển động, vì thế, so với các đặc trưng khác (văn bản, âm thanh, ...), các đặc trưng liên quan đến hình ảnh được quan tâm hơn cả.

Khó khăn liên quan đến đặc trưng hình ảnh là chúng phải được mô hình hóa sao cho có thể xác định chính xác các khái niệm trong các môi trường khác nhau, nghĩa là chúng có thể làm giảm khoảng cách khác biệt về ngữ nghĩa trong những điều kiện giác quan khác nhau (chẳng hạn ánh sáng thay đổi, góc nhìn khác nhau, tỉ lệ khác nhau, ...). Vì thế chúng ta cần phải kết hợp nhiều đặc trưng với nhau. Sau đây chúng tôi sẽ trình bày một số đặc trưng được sử dụng của các nhóm nghiên cứu đã có nhiều kết quả cao trong các hội thảo liên quan đến lĩnh vực truy xuất video.

- Nhóm nghiên cứu của Tokyo Institute of Technology, đạt kết quả cao nhất trong TRECVID 2011, họ sử dụng 5 loại đặc trưng về hình ảnh và 1 loại đặc trưng âm thanh: SIFT features with Harris-Affine detector, SIFT features with Hessian-Affine detector, SIFT and hue histogram with dense sampling, HOG with dense sampling, HOG from temporal subtraction images, MFCC audio features. Vector đặc trưng gồm 6 loại đặc trưng trên được gọi là Gaussian mixture model (GMM) supervectors và các đặc trưng được rút trích bằng cách áp dụng tree-structured GMMs [11,12].
- Nhóm nghiên cứu của University of Amsterdam(UvA)⁹, sử dụng bộ mô tả dựa trên màu sắc (color descriptors)bằng cách áp dụng đặc trưng SIFT cho các hệ màu khác nhau [21].

⁹ <http://www.ceessnoek.info>

- Nhóm nghiên cứu Vireo¹⁰ sử dụng hai đặc trưng toàn cục grid-based color moments (CM), grid-based wavelet texture (WT) và hai đặc trưng cục bộ: SIFT-DoG, SIFT-HA; phương pháp máy học là SVM.
- Nhóm nghiên cứu của Carnegie Mellon University¹¹, sử dụng 5 đặc trưng SIFT-HL (Harris-Laplace), CSIFT (Color SIFT)-HL, SIFT-DS, CSIFT-DS, và MoSIFT. Phương pháp máy học được áp dụng là Sequential Boosting SVM.

Chúng ta có thể thấy, đặc trưng SIFT và các biến thể của nó đóng vai trò rất quan trọng trong việc xác định các khái niệm. Ngoài ra, trong các nhóm đặc trưng trên thì đặc trưng liên quan đến âm thanh chiếm một tỉ lệ rất nhỏ so với các đặc trưng liên quan đến hình ảnh.

Phương pháp Máy học

Do khối lượng dữ liệu video rất lớn, có nhiều khái niệm cần xác định, vì thế sau khi biểu diễn các đoạn video bằng vector đặc trưng, chúng ta sử dụng các phương pháp máy học để xác định mối liên quan giữa các khái niệm và vector đặc trưng. Phương pháp máy học thường được sử dụng là học giám sát (supervised learning) dựa trên tập mẫu chứa các khái niệm cho trước.

Có hai thách thức mà các phương pháp học phải đối mặt, đó là trường hợp số vec-tor đặc trưng quá nhiều có thể gây nhiễu và trường hợp có nhiều tham số được dùng trong mô hình (over-fitting) [13]. Nói cách khác, cần phải xử lý trường hợp mất cân bằng giữa các mẫu dương so với mẫu âm và trường hợp dữ liệu bị nhiễu hoặc lỗi. Để đáp ứng các yêu cầu trên, phương pháp SVM là phù hợp hơn cả. Việc chọn hàm nhân (kernel) và tham số C để cân bằng giữa lỗi huấn luyện và độ phức tạp của mô hình ảnh hưởng đáng kể đến kết quả [16, 29]. Hàm nhân thường được chọn dựa trên hàm Gaussian RBF hoặc χ^2 . Để khắc phục hiện tượng over-fitting, các tham số được đánh giá dựa trên phương pháp đánh giá chéo (cross-validation).

Kết hợp Đặc trưng

Như chúng ta biết, trong một đoạn video có thể biểu diễn bằng nhiều đặc trưng khác nhau dựa trên các mô hình khác nhau, vì thế việc chỉ sử dụng một mô hình (unimodal) hay nhiều mô hình (multimodal) cần được xem xét.

- Trong phương pháp sử dụng một mô hình, có nhiều phương pháp xác định các đặc trưng dựa vào hình ảnh (màu sắc, cấu trúc văn, hình dạng,...) hoặc các biến thể của chúng, phạm vi rút trích đặc trưng có thể áp dụng trong các vùng khác nhau như toàn cục, cục bộ hay ở mức các điểm keypoints [15, 25, 33, 35]. Tuy nhiên, sử dụng cùng một mô hình có thể không phù hợp cho tất cả các khái niệm khác nhau, vì thế ta thường kết hợp nhiều mô hình.

¹⁰ <http://vireo.cs.cityu.edu.hk/>

¹¹ <http://www.lti.cs.cmu.edu/>

- Trong phương pháp dựa trên nhiều mô hình, các đặc trưng được rút trích từ nhiều kênh khác nhau như âm thanh, hình ảnh [28].

Sau khi các đặc trưng được rút trích, chúng được kết hợp thành một vector đặc trưng mô tả cho video. Mặc dù ưu điểm của việc kết hợp này giúp ta tiết kiệm chi phí cho quá trình học, tuy nhiên, chúng ta cần quan tâm đến một số yếu tố như số thành phần của vector sẽ gia tăng, mức độ liên quan với các khái niệm của các đặc trưng khác nhau, số chiều và miền giá trị của các đặc trưng khác nhau, và thứ tự kết hợp các đặc trưng,...

Mặc dù đây là một trong những hướng mới trong truy xuất video, nhưng dựa trên một số kết quả được công bố trong [18,26] và một số công trình nghiên cứu có liên quan [6,27] thì những kết quả này cần được cải tiến nâng cao hơn.

3 Một số Kết quả và Định hướng Nghiên cứu

Trong thời gian qua, chúng tôi có thực hiện 2 đề tài khoa học cấp ĐHQG không trọng điểm và công bố hai bài báo:

- Một bài tại hội thảo ICCCI 2012, đã được chấp nhận, tên bài báo “Caption text and keyframe similarity based video retrieval system”. Trong bài báo này chúng tôi trình bày về kết quả xây dựng một hệ thống truy vấn video dựa trên việc rút trích các phụ đề và so sánh độ tương tự giữa ảnh truy vấn với ảnh chính trong các đoạn video. Đây chỉ là một công trình sơ khởi, làm nền tảng để phát triển các phương pháp khác trong hệ thống.
- Một kết quả khác liên quan đến việc thực hiện so sánh các bộ mô tả đặc trưng về hình dáng của ảnh được chúng tôi công bố tại hội thảo ITEDU@2010 (“A comparative study of shape descriptors in content based image retrieval”).

Trong đề tài này, chúng tôi dự định tập trung nghiên cứu các nội dung sau:

- Sử dụng phương pháp đa mô hình, kết hợp nhiều đặc trưng dựa trên dữ liệu văn bản, âm thanh và hình ảnh.
- Đề xuất các biến thể của các đặc trưng để nâng cao hiệu quả xác định các khái niệm trong đoạn video
- Nghiên cứu các phương pháp máy học, sự kết hợp các đặc trưng và đề xuất mô hình cải tiến khả năng nhận diện các khái niệm trong video.
- Xây dựng một hệ thống áp dụng các phương pháp và mô hình được đề xuất để đánh giá hiệu quả và cho phép người dùng truy xuất dữ liệu video dựa trên các khái niệm.

4 Kết luận

Trong bài viết này, chúng tôi đã trình bày tổng quan về đề tài này, những nghiên cứu có liên quan và đặc biệt là định hướng nghiên cứu của chúng tôi. Tuy nhiên, những định hướng này mang nhiều tính chủ quan và phiến diện. Chúng tôi rất mong nhận được sự đóng góp ý của quý độc giả để chúng tôi có thể hoàn thành nghiên cứu của mình.

Tài liệu tham khảo

1. Ahn, L.v.: Games with a purpose. *Computer* 39(6), 92–94 (Jun 2006), <http://dx.doi.org/10.1109/MC.2006.196>
2. Ahn, L.v., Dabbish, L.: Designing games with a purpose. *Commun. ACM* 51(8), 58–67 (Aug 2008), <http://doi.acm.org/10.1145/1378704.1378719>
3. Brown, M.G., Foote, J.T., Jones, G.J.F., Sparck Jones, K., Young, S.J.: Automatic content-based retrieval of broadcast news. In: *Proceedings of the third ACM international conference on Multimedia*. pp. 35–43. MULTIMEDIA '95, ACM, New York, NY, USA (1995), <http://doi.acm.org/10.1145/217279.215080>
4. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Trans. Pattern Anal. Mach. Intell.* 24(8), 1026–1038 (Aug 2002), <http://dx.doi.org/10.1109/TPAMI.2002.1023800>
5. Chang, S.F., Chen, W., Meng, H.J., Sundaram, H., Zhong, D.: A fully automated content-based video search engine supporting spatio-temporal queries. *IEEE Trans. Cir. and Sys. for Video Technol.* 8(5), 602–615 (Sep 1998), <http://dx.doi.org/10.1109/76.718507>
6. Chong-Wah Ngo, S. A. Zhu, W.Z.C.C.T.Y.L.P., Tan, H.K.: Vireo @ trecvid 2011: Instance search, semantic indexing, multimedia event detection and known-item search (December 2011), *nIST TRECVID Workshop (TRECVID'11)*, Gaithersburg, USA,
7. Davis, M.: Editing out video editing. *IEEE MultiMedia* 10(2), 54–64 (Mar 2003), <http://dx.doi.org/10.1109/MMUL.2003.1195161>
8. Del Bimbo, A., Pala, P.: Visual image retrieval by elastic matching of user sketches. *IEEE Trans. Pattern Anal. Mach. Intell.* 19(2), 121–132 (Feb 1997), <http://dx.doi.org/10.1109/34.574790>
9. Hanjalic, A., Xu, L.Q.: Affective video content representation and modeling. *IEEE Transactions on Multimedia* 7(1), 143–154 (2005)
10. Huurnink, B., Hofmann, K., de Rijke, M.: Assessing concept selection for video retrieval. In: *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. pp. 459–466. MIR '08, ACM, New York, NY, USA (2008), <http://doi.acm.org/10.1145/1460096.1460170>
11. Inoue, N., Saito, T., Shinoda, K., Furui, S.: High-level feature extraction using sift gmms and audio models. In: *Proceedings of the 2010 20th International Conference on Pattern Recognition*. pp. 3220–3223. ICPR '10, IEEE Computer Society, Washington, DC, USA (2010), <http://dx.doi.org/10.1109/ICPR.2010.787>
12. Inoue, N., Shinoda, K.: A fast map adaptation technique for gmm-supervector-based video semantic indexing systems. In: *Proceedings of the 19th ACM international conference on Multimedia*. pp. 1357–1360. MM '11, ACM, New York, NY, USA (2011), <http://doi.acm.org/10.1145/2072298.2072014>
13. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* 22(1), 4–37 (Jan 2000), <http://dx.doi.org/10.1109/34.824819>
14. Jain, R., Hampapur, A.: Metadata in video databases. *SIGMOD Rec.* 23(4), 27–33 (Dec 1994), <http://doi.acm.org/10.1145/190627.190638>
15. Jiang, Y.G., Yang, J., Ngo, C.W., Hauptmann, A.G.: Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Transactions on Multimedia* 12, 42–53 (2010)

16. Naphade, M.R.: On supervision and statistical learning for semantic multimedia analysis. *J. Vis. Commun. Image Represent.* 15(3), 348–369 (Sep 2004), <http://dx.doi.org/10.1016/j.jvcir.2004.04.010>
17. Naphade, M.R., Huang, T.S.: A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Transactions on Multimedia* pp. 141–151 (2001)
18. Quenot, G.: Trecvid 2011 semantic indexing task (2011), <http://www-nlpir.nist.gov/projects/tvpubs/tv11.slides/tv11.sin.slides.pdf>
19. Rosenfeld, A.: Picture processing by computer. *ACM Comput. Surv.* 1(3), 147–176 (Sep 1969), <http://doi.acm.org/10.1145/356551.356554>
20. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision* 77(1-3), 157–173 (May 2008), <http://dx.doi.org/10.1007/s11263-007-0090-8>
21. van de Sande, K.E., Gevers, T., Snoek, C.G.: Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(9), 1582–1596 (September 2010)
22. Seinstra, F.J., Geusebroek, J.M., Koelma, D., Snoek, C.G.M., Worring, M., Smeulders, A.W.M.: High-performance distributed image and video content analysis with parallel-horus. *IEEE Multimedia* 14(4), 64–75 (October–December 2007)
23. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.* 22(12), 1349–1380 (Dec 2000), <http://dx.doi.org/10.1109/34.895972>
24. Snoek, C.G.M., Gemert, J.C.V., Gevers, T., Huurnink, B., Koelma, D.C., Liempt, M.V., Rooij, O.D., Seinstra, F.J., Smeulders, A.W.M., Thean, A.H.C., Veenman, C.J., Worring, M.: The mediamill trecvid 2006 semantic video search engine. In: *Proceedings of the 4th TRECVID Workshop* (2006)
25. Snoek, C.G.M., van de Sande, K.E.A., Li, X., Mazloom, M., Jiang, Y.G., Koelma, D.C., Smeulders, A.W.M.: The MediaMill TRECVID 2011 semantic video search engine. In: *Proceedings of the 9th TRECVID Workshop*. Gaithersburg, USA (December 2011)
26. Snoek, C.G.M., Smeulders, A.W.M.: Visual-concept search solved ? *IEEE Computer* 43(6), 76–78 (June 2010)
27. Snoek, C.G.M., Worring, M.: Concept-based video retrieval. *Found. Trends Inf. Retr.* 2(4), 215–322 (Apr 2009), <http://dx.doi.org/10.1561/15000000014>
28. Snoek, C.G.M., Worring, M., van Gemert, J.C., Geusebroek, J.M., Smeulders, A.W.M.: The challenge problem for automated detection of 101 semantic concepts in multimedia. In: *Proceedings of the 14th annual ACM international conference on Multimedia*. pp. 421–430. MULTIMEDIA '06, ACM, New York, NY, USA (2006), <http://doi.acm.org/10.1145/1180639.1180727>
29. Snoek, C.G.M., Worring, M., Geusebroek, J.M., Koelma, D.C., Seinstra, F.J., Smeulders, A.W.M.: The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(10), 1678–1689 (October 2006)
30. Szummer, M., Picard, R.W.: Indoor-outdoor image classification. In: *Proceedings of the 1998 International Workshop on Content-Based Access of Image and Video Databases (CAIVD '98)*. pp. 42–51. CAIVD '98, IEEE Computer Society, Washington, DC, USA (1998), <http://dl.acm.org/citation.cfm?id=791220.791680>
31. Vailaya, A., Jain, A., Zhang, H.J.: On image classification: City vs. landscape. In: *Proceedings of the IEEE Workshop on Content - Based Access of Image and Video Libraries*. pp. 3–8. CBAIVL '98, IEEE Computer Society, Washington, DC, USA (1998), <http://dl.acm.org/citation.cfm?id=521772.791885>

32. Wang, J.Z., Li, J., Wiederhold, G.: Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Trans. Pattern Anal. Mach. Intell.* 23(9), 947–963 (Sep 2001), <http://dx.doi.org/10.1109/34.955109>
33. Wang, M., Hua, X.S., Yuan, X., Song, Y., Dai, L.R.: Optimizing multi-graph learning: towards a unified video annotation scheme. In: *Proceedings of the 15th international conference on Multimedia*. pp. 862–871. MULTIMEDIA '07, ACM, New York, NY, USA (2007), <http://doi.acm.org/10.1145/1291233.1291431>
34. Westerveld, T.H.W., de Vries, A.P.: Multimedia retrieval using multiple images. In: *Proceedings of the International Conference on Image and Video Retrieval* (3). pp. 344 – 352. Springer (July 2004), <http://oai.cwi.nl/oai/asset/18058/18058A.pdf>, received Best Paper Award,
35. Wilkins, P., Adamek, T., O'Connor, N.E., Smeaton, A.F.: Inexpensive fusion methods for enhancing feature detection. *Image Commun.* 22(7-8), 635–650 (Aug 2007), <http://dx.doi.org/10.1016/j.image.2007.05.012>
36. Zhang, H., Tan, S.Y., Smoliar, S.W., Yihong, G.: Automatic parsing and indexing of news video. *Multimedia Syst.* 2(6), 256–266 (Jan 1995), <http://dx.doi.org/10.1007/BF01225243>