NHẬN DIỆN HÀNH ĐỘNG NGƯỜI BẰNG CẬP NHẬT KẾT QUẢ PHÂN LỚP

ThS. Lê Viết Tuấn
Trường Đại học Yersin Đà Lạt

Tóm tắt

Bài báo khai thác mối tương quan giữa hành động và cảnh trong nhận diện hành động của người trong video. Để khai thác mối tương quan này, bài báo kết hợp đặc trưng hành động và đặc trưng ngữ cảnh bằng phương pháp cập nhật kết quả phân lớp. Qua kết quả thử nghiệm thấy được tính hiệu quả và tính tối ưu của thuật toán kết hợp.

Abstract

Human action recognition by updating the classification scores

In this paper, we exploit the co-occurrence between scenes and actions to recognize a human action in video. To exploit this co-occurrence, the paper combines action and scenes featured by the method of updating classification scores. The experimental results have proved the efficiency and effectiveness of the combinative method.

1. Giới thiệu

Hiện nay dữ liệu video dễ dàng được tạo ra bởi các thiết bị như: Máy ảnh kỹ thuật số, máy tính xách tay, điện thoại di động, bên cạnh đó các trang web chia sẻ video tăng trưởng không ngừng. Bài toán nhận diện hành động người trong video, đóng góp một phần tự động hóa khai thác tài nguyên dữ liệu nhiều thông tin này. Các ứng dụng liên quan đến bài toán nhận diện hành động như:

An ninh và các hệ thống giám sát truyền thống gồm mạng lưới các camera và được giám sát bởi con người. Với sự tăng lên của camera cũng như các hệ thống này được triển khai nhiều địa điểm, dẫn đến vấn đề hiệu quả và độ chính xác của người giám sát khi phải bao quát hết toàn bộ hệ thống. Nhiệm vụ của thị giác máy tính là tìm ra giải pháp có thể thay thế hoặc hỗ trợ người giám sát. Tự động nhận ra các bất thường từ các hệ thống giám sát là vấn đề thu hút khá nhiều nghiên cứu. Một ứng dụng liên quan nữa đó là tìm kiếm đoạn video hành động "quan tâm" từ cơ sở dữ liệu video lớn được lưu trữ bởi các hệ thống giám sát.

Tương tác giữa người và máy vẫn còn nhiều thách thức, những dấu hiệu thị giác là phương thức quan trọng nhất giao tiếp phi ngôn ngữ. Khai thác hiệu quả phương thức giao tiếp này: Thông qua cử chỉ, hành động, hoạt động hứa hẹn tạo ra thế hệ máy tính tương tác chính xác và tự nhiên hơn với con người. Một ứng dụng điển hình trong lĩnh vực này là "ngôi nhà thông minh", phản ứng thông minh với cử chỉ, hành động của người sử dụng. Tuy nhiên những ứng dụng này vẫn chưa đủ hoàn chỉnh và vẫn đang thu hút nhiều nghiên cứu.

Bên cạnh đó nhận diện hành động người trong video còn được ứng dụng trong tóm tắt, truy vấn video, phân tích thể thao.

Bài báo trình bày phương pháp kết hợp đặc trưng Histograms of 3D gradient orientations (HOG3D) - biểu diễn đặc trưng hành động với đặc trưng Color SIFT - biểu diễn đặc trưng ngữ cảnh của hành động. Có thể tìm thấy nhiều phương pháp kết hợp trong các công trình như 3, 10, tuy nhiên hầu hết các phương pháp sau khi kết hợp đều phải biểu diễn đặc trưng trên một không gian lớn hơn. Do phải biểu diễn trên một không gian lớn hơn nên khả năng phân lớp sẽ kém chính xác hơn so với trên một không gian nhỏ hơn. Bài báo trình bày phương pháp kết hợp cập nhật kết quả phân lớp nhằm tránh phải biểu diễn đặc trưng trên một không gian lớn hơn đồng thời khai thác mối tương quan giữa hành động và cảnh xảy ra.

Bài báo được tổ chức như sau: Phần 2 trình bày các công trình liên quan đến bài toán nhận diện hành động, phần 3 phương pháp đề xuất để biểu diễn hành động người, phần 4 là kết quả thực nghiệm, cuối cùng là phần kết luận.

2. Các công trình liên quan

Có hai hướng tiếp cận chính, đó là phương pháp toàn cục (holistic methods) và phương pháp đặc trưng cục bộ (local feature methods).

Phương pháp toàn cục: Cấu trúc toàn cục và chuyển động của cơ thể được sử dụng để biểu diễn hành động người. Hầu hết các phương pháp toàn cục cho kết quả tốt hơn với dữ liệu video trong ngữ cảnh ràng buộc trước. Do cần mã hóa một lượng lớn thông tin thị giác, các phương pháp toàn cục nhạy cảm với nhiễu, che khuất và thay đổi góc nhìn. Hầu

hết các phương pháp cho thấy chi phí tính toán cao cho tiền xử lý dữ liệu đầu vào như trừ nền, phân đoạn, theo vết đối tượng.

Phương pháp đặc trưng cục bộ lưu giữ đặc trưng dáng và chuyển động cho một vùng cục bộ trong video, Ít nhạy cảm với nhiễu, che khuất một phần, không bị ảnh hưởng bởi các bước tiền xử lý.Tuy nhiên phụ thuộc vào hiệu quả và độ chính của việc phát hiện điểm đặc trưng. Đặc trưng được trích xuất trực tiếp từ video, nên có thể tránh được các sai sót của các phương pháp trong giai đoạn tiền xử lý như phân đoạn, trừ nền hay phát hiện người.

Miêu tả đặc trưng HOG3D 6 được đề xuất bởi Kläser dựa trên lược đồ hướng gradien 3D. Gradient trong video được tính hiệu quả nhờ phép tính tổng trong video (integral video). Một khối 3D được chia thành lưới gồm nhiều ô, lược đồ miêu tả cuối cùng là nối của tất cả lược đồ gradient của các ô trong khối.

Laptev và các đồng nghiệp 10 giới thiệu bộ miêu tả HoG và HoF. Để miêu tả thông tin dáng và chuyển động, tác giả kết hợp histograms of oriented spatial gradients (HoG) và histograms of optical flow (HoF). Lược đồ được tích lũy trong vùng lân cận không-thời gian tại điểm đặc trưng được phát hiện. Mỗi vùng cục bộ xung quanh điểm đặc trưng được chia thành N x N x M ô lưới, mỗi ô gồm 4 bin cho lược đồ HoG và 5 bin cho lược đồ HoF. Lược đồ tại mỗi ô được chuẩn hóa, và nối lại thành bộ miêu tả HoG và HoF cuối cùng.

Marcin Marszałek và các đồng nghiệp 12 khai thác ngữ cảnh là cảnh tự nhiên khi nhận diện hành động người trong video. Thông thường hành động người luôn có một mối ràng buộc với thông tin cảnh xảy ra hành động. Công trình của các tác giả dựa trên mô hình túi từ và phân lớp bằng SVM. Để biểu diễn hành động người, Harris3D được sử dụng để phát hiện điểm đặc trưng, miêu tả đặc trưng HoG và HoF được tính cho vùng xung quanh điểm đặc trưng phát hiện bằng Harris3D. Để biểu diễn cho cảnh, các tác giả sử dụng Harris2D để phát hiện các vị trí nổi trội trong từng frame độc lập được trích xuất từ video; miêu tả đặc trưng SIFT cho vùng xung quanh các điểm đặc trưng Harris2D được phát hiện.

Nazli Ikizler-Cinbis và Stan Sclaroff 5 đề xuất phương pháp cho nhận diện hành động người mà kết hợp nhiều đặc trưng của các thực thể như: Đối tượng, cảnh và con người. Cụ thể, tác giả sử dụng lược đồ hướng gradient (Histogram of Oriented Gradients) mã hóa thông tin dáng, optical flow mã hóa thông tin chuyển động, để miêu tả đặc trưng ngữ cảnh tác giả trích xuất đặc trưng Gist từ 5 frames trích ngẫu nhiên của video, để miêu tả đặc trưng màu, tác giả sử dụng lược đồ màu. Các đặc trưng được kết hợp với nhau thông qua mô hình học đa thực thể.

Lamberto Ballan và các đồng nghiệp 3 đã đề xuất một phương pháp mới cho phân lớp hành động người trong video khi kết hợp miêu tả đặc trưng gradient 3D (3D gradeint descriptor) và miêu tả dòng quang học (optical flow descriptor). Tác giả đề nghị 2 phương pháp kết hợp, phương pháp đầu tiên là hai miêu tả đặc trưng này được nối với nhau, và vector miêu tả đặc trưng kết hợp này dùng để xây dựng từ điển hành động của người. Phương pháp thứ hai, từ điển cho mỗi loại đặc trưng được xây dựng độc lập, sau đó lược đồ tần suất xuất hiện của hai từ điển được kết hợp với nhau.

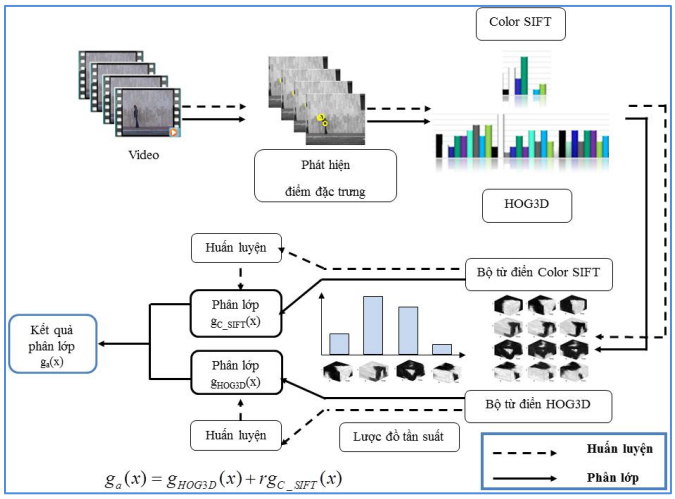
3. Phương pháp đề xuất

Để nhận diện hành động người trong dữ liệu video thực tế, bài báo khai thác mối tương quan giữa hành động (action) và cảnh (scene). Hành động người hầu như luôn có một mối ràng buộc với cảnh xảy ra hành động. Giả sử nếu có hình ảnh hồ bơi trong cảnh thì khả năng đó là hành động bơi lặn (diving), nếu trong cảnh có sân bóng rổ thì xác suất đó là hành động bơi lặn sẽ giảm xuống, mà khả năng hành động ở đây là hành động chơi bóng rổ.

Trong bài toán nhận diện đối tượng và cảnh 14,

Hình 1: Mối tương quan giữa hành động và cảnh

đặc trưng SIFT màu cho thấy ưu thế trong biểu diễn thông tin ngữ cảnh. SIFT màu là mở rộng của đặc trưng SIFT, về cơ bản đặc trưng SIFT miêu tả phân bố cạnh độ xám, SIFT màu mở rộng bằng miêu tả cạnh màu cục bộ xung quanh điểm đặc trưng. Điều này đạt được bằng cách thay vì chỉ tính biên cạnh dựa trên biến thiên cường độ độ xám thì tính trên màu.



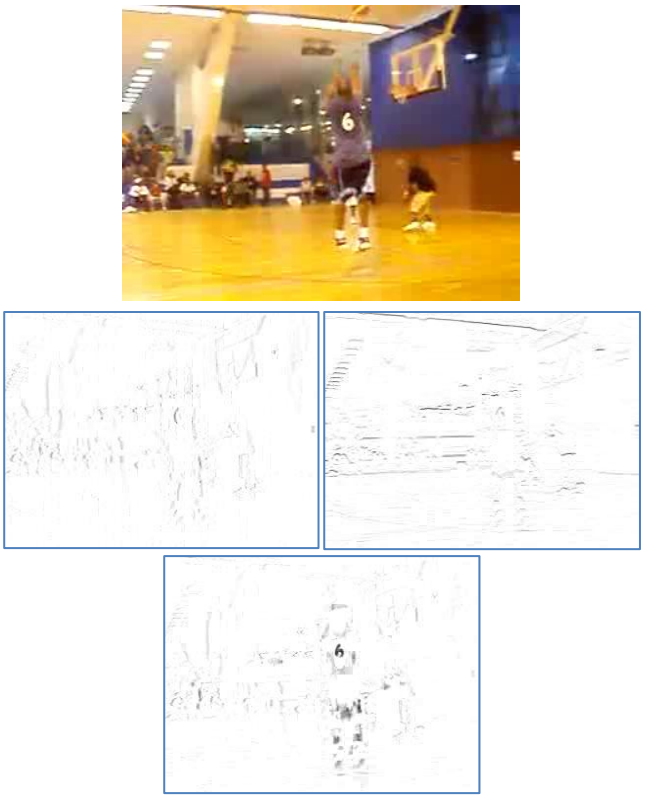
Hình 2: Mô hình nhận diện hành động người trong dữ liệu video thực tế

Biểu diễn đặc trưng hành động:

Phát hiện điểm đặc trưng: Điểm đặc trưng được xem như những vị trí duy nhất trong không-thời gian của hành động. Điểm đặc trưng Harris3D 9 là điểm mà tại đó không chỉ có sự thay giá trị điểm ảnh vượt trội trên miền không gian mà cả trong miền thời gian. Điểm đặc trưng xuất hiện tại những vị trí mà có sự thay đổi đột ngột về hướng chuyển động, hay nói cách khác có sự gián đoạn trong luồng chuyển động, chuyển động trơn tru thì không có góc xuất hiện.

Hình 3: Điểm đặc trưng Harris3D

Miêu tả đặc trưng: Kläser và các đồng nghiệp đã đề xuất miêu tả đặc trưng HOG3D 6, dựa trên lược đồ hướng biến thiên độ sáng ba chiều (Histograms of 3D gradient orientations). Miêu tả đặc trưng HOG3D là sự kết hợp thông tin dáng và chuyển động tại cùng một thời điểm, đây là điểm khác biệt lớn so với các công trình kết hợp thông tin dáng và chuyển động trước đó. Một khối 3 chiều được chia thành $n_x \times n_y \times n_t$ ô lưới. Miêu tả đặc trưng tương ứng là nối của các lược đồ gradient sau khi đã được lượng tử hóa. Gradient 3 chiều được tính hiệu quả nhờ phép tính tổng trong video (integral video). So với các miêu tả đặc trưng trước đó như 4, 8, các tác giả chủ yếu tính toán gradient và dòng quang học để biểu diễn hành động. Hầu hết các miêu tả đặc trưng này dựa trên thành phần độ lớn gradient, nên có một nhược điểm chung đó là khá nhạy cảm với điều kiện chiếu sáng thay đổi. Miêu tả đặc trưng HOG3D dựa trên hướng gradient, nên khá ổn định với điều kiện chiếu sáng.



Hình 4: tính gradient theo x, y và t

Biểu diễn đặc trưng cảnh:

Phát hiện điểm đặc trưng: Difference-of-Gaussian được sử dụng để phát hiện điểm trọng yếu (keypoint) mà bất biến với tỉ lệ và hướng, bằng cách trừ 2 ảnh L (là ảnh tích chập ảnh Gaussian với ảnh đầu vào) kế nhau trong octave.

Miêu tả đặc trưng: Opponent SIFT 14 miêu tả các kênh màu trong không gian màu opponent sử dụng miêu tả đặc trưng SIFT. Miêu tả đặc trưng là kết hợp của 3 vector miêu tả đặc trưng một chiều dựa trên từng kênh màu của không gian màu opponent:

Kênh O3 chứa thông tin cường độ, 2 kênh còn lại miêu tả thông tin màu của ảnh, tuy nhiên 2 kênh này vẫn chứa 1 ít thông tin cường độ, vì vậy không bất biến với sự thay đổi cường độ sáng.

Túi đặc trưng:

Lĩnh vực nhận dạng đối tượng trên ảnh và tìm kiếm đa phương tiện đã ghi nhận những thành công

lớn của phương pháp dựa trên từ điển thị giác. Hạn chế của hướng tiếp cận này mắc phải là sự mất mát thông tin về không gian và hình dạng khi gom nhóm thành các từ thị giác. Hầu hết các phương pháp dựa trên mô hình từ điển thị giác đều có một bước giống nhau, đó là biểu diễn tất cả các mẫu học hay mẫu kiểm nghiệm bằng các từ thị giác. Các từ thị giác chỉ có khả năng duy nhất tổng quát hóa thông tin tại một vùng cục bộ, mà bỏ qua mọi thông tin về vị trí cũng như vị trí tương đối giữa các từ thị giác.

Từ 2 tập vector miêu tả biểu diễn cho đặc trưng hành động và ngữ cảnh, thuật toán gom nhóm không giám sát k-means++ (k-means++ clustering) được áp dụng để phân tập vector về K nhóm. Hầu hết các hệ thống nhận diện hành động người dựa trên mô hình túi đặc trưng như 10, 12 đều sử sụng thuật toán góm nhóm k-means bởi tính đơn giản và tốc độ hội tụ của nó, tuy nhiên k-mean chỉ có thể đạt được tối ưu cục bộ, do các trọng tâm được sinh ngẫu nhiên trong bước khởi tạo. Bài báo sử dụng thuật toán k-means++ để cải thiện tốc độ cũng như độ chính xác khi xây dựng mô hình túi đặc trưng.

Khoảng cách Euclid được dùng để tính khoảng cách giữa hai điểm trong không gian đặc trưng. Mỗi nhóm (cluster) được xem như là 1 từ thị giác (visual word), với số nhóm k = 4000.Tập tất cả các từ vựng tạo thành bộ từ điển thị giác (visual vocabulary) hay còn gọi là codebook cho mô hình BoF.

Huấn luyện và phân lớp:

Bài báo huấn luyện bộ phân lớp chuyên biệt cho mỗi hành động sử dụng chiến lược one-vs-all. Mẫu dương là tập các video của hành động cần huấn luyện và mẫu âm là tập các video của các hành động còn lại.

Cập nhật kết quả phân lớp:

Gọi $A' = v_1, \dots, v_k$ tập con của tập dữ liệu huấn luyện $A = v_1, \dots, v_n$, $k \leq n$.

Ta có kết quả phân lớp với đặc trưng HOG3D và HoF trên tập dữ liệu A' lần lượt là:

$$S_{HOG3D} = (s_{HOG3D_1}, \dots, s_{HOG3D_k})$$

$$S_{HoF} = (s_{HoF_1}, \dots, s_{HoF_k})$$

Chuẩn hóa L2-norm cho S_{HOG3D}, S_{HoF}: S_{HOG3D}, S_{HoF}

$$S_{HOG3D} \rightarrow \frac{S_{HOG3D}}{S_{HOG3D}\ _2^2 + \epsilon^2}; S_{HoF} \rightarrow \frac{S_{HoF}}{S_{HoF}\ _2^2 + \epsilon^2}$$

Hệ số r được xác định dựa vào công thức: (e: hằng số giá trị nhỏ)

$$r = \frac{1}{\frac{1}{k}\sum_{i=1}^{k} \frac{S_{HOG3D_i}}{S_{HoF_i}+e}} \quad (1.1)$$

4. Thử nghiệm và kết quả

Thử nghiệm:

Tập dữ liệu hành động Youtube 11 được giới thiệu bởi Liu và các đồng nghiệp, bao gồm 11 lớp hành động: ném bóng rổ (basketball shooting), đi xe đạp (biking), lặn (diving), đánh golf (golf swinging), cưỡi ngựa (horseback riding), đá bóng (soccer juggling), đu xích đu (swinging), đánh quần vợt (tennis swinging), nhảy (trampoline jumping), bóng chuyền(volleyball spiking), đi bộ với chó (walking with a dog). Mỗi đoạn video dài khoảng 3-10 giây. Kích thước mỗi frame ảnh là: 320 x 240.

Tập dữ liệu bao gồm 1600 đoạn video. Với mỗi lớp hành động, các đoạn video được nhóm thành 25 nhóm, mỗi nhóm có thể nhiều hơn 4 đoạn video. Theo đề nghị của tác giả, với mỗi lớp hành động sẽ chọn 4 video đầu tiên trong mỗi nhóm cho quá trình huấn luyện và phân lớp. Đây là một tập dữ liệu với nhiều thách thức: sự chuyển động của camera, hình dạng và tư thế của đối tượng, tỉ lệ của đối tượng, điểm quan sát, nền hỗn loạn, điều kiện chiếu sáng.

Kết quả:

Bảng 1: So sánh các biểu diễn đặc trưng thông dụng trên tập Youtube

	Đặc trưng	Tập dữ liệu Youtube
Bài báo	Cập nhật HOG3D + OSIFT	**82%**
Luận án tiến sĩ của Klaser năm 2010, trang 49 7.	HOG3D (tối ưu tham số)	68.3%
	HOG3D	68.1%
	HOGHOF	71.2%
	HOG	68.0%
	HOF	63.9%

Bảng 2: Độ chính xác trên tập dữ liệu Youtube

	Tham khảo	Phương pháp	Độ chính xác
BOF	Bài báo	HOG3D-OSIFT	**82%**
	Heng Wang 15	Dense Trajectories + HoG-HoF-MBH	84.2%
	Klaser's thesis (chương 5) 7	Feature trajectories + HoG-HoF-MBH	79.8%
	Klaser's thesis (chương 3) 7	Harris3D + HOG3D	68.36 %
KHÁC	Liu 11	motion/static features + pruning + grouping + AdaBoost	71.2%

5. Kết luận

Bài báo đề xuất phương pháp cập nhật kết quả phân lớp khi kết hợp đặc trưng HOG3D (biểu diễn đặc trưng dáng, chuyển động) và Opponent SIFT (biểu diễn đặc trưng ngữ cảnh) để khai thác mối tương quan giữa hành động và cảnh trong dữ liệu video thực tế. Kết quả cho thấy, đặc trưng ngữ cảnh góp phần nâng cao hiệu quả phân lớp cho đặc trưng hành động, đặc biệt đặc trưng ngữ cảnh sẽ làm việc tốt đối với những hành động xảy ra trong những không gian nhất định như: hồ bơi, sân bóng rổ.

Tài liệu tham khảo

1. *Tuan Le-Viet, Ngoc Ly-Quoc (2012) "Human Action Recognition on Simple and Complex Background in Video",* IEEE International Conference on Control Automation and Information Sciences (ICCAIS).

2. D. Arthur, S. Vassilvitskii, K-means++ (2007), *The Advantages of Careful Seeing, In Proc. of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms,* 1027-1035.

3. L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra (2012), *"Effective Codebooks for Human Action Representation and Classification in Unconstrained Videos",* IEEE Transactions on Multimedia, vol. in press.

4. P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie (2005) *Behavior recognition via sparse spatio-temporal features.* In VS-PETS.

5. Nazli Ikizler-Cinbis and Stan Sclaroff. Object, Scene and Actions (2010), *Combining Multiple Features for Human Action Recognition*, In ECCV.

6. A. Klaser, M. Marsza lek, and C. Schmid (2008) *A spatio-temporal descriptor based on 3D-gradients,* In BMVC.

7. A. Klaser (2010), *Learning human actions in videos, Thesis (PhD),* University DE GRENOBLE.

8. I. Laptev and T. Lindeberg (2004), *Local descriptors for spatio-temporal recognition,* In SCVMA.

9. I. Laptev (2005), *On space-time interest points,* IJCV, 64:107-123.

10. I. Laptev, M. Marsza lek, C. Schmid, and B. Rozenfeld (2008), *Learning realistic human actions from movies,* In CVPR.

11. J. Liu, J. Luo, and M. Shah (2009), *Recognizing realistic actions from videos "in the wild",* In CVPR.

12. M. Marszalek, I. Laptev, and C. Schmid, (2009), Actions in context, In CVPR.

13. T. B. Moeslund, A. Hilton, V. Krűger (2006), *A survey of advances in vision-based human motion capture and analysis,*. CVIU, 104(2-3):90-126, 2006.

14. K. E. A. van de Sande, T. Gevers, and C. G. M. (2008), *Snoek. Evaluation of color descriptors for object and scene recognition,* In IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska.

15. H. Wang, A. Klaser, c. Schmid, CL. Liu (2011), Action recognition by dense trajectories, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3169–3176.