SO SÁNH PHƯƠNG PHÁP NHẬN DẠNG HÀNH ĐỘNG CON NGƯỜI TRONG ĐOẠN VIDEO QUAY BẰNG MỘT CAMERA DÙNG DTW VÀ HMM

COMPARISON OF HUMAN ACTION RECOGNITIONS IN MONOCULAR VIDEOS USING DTW AND HMM

Hoàng Lê Uyên Thục[1], **Phạm Văn Tuấn**[1], **Shian-Ru Ke**[2]

[1]*Trường Đại học Bách khoa, Đại học Đà Nẵng; Email: hluthuc@dut.udn.vn, pvtuan@dut.udn.vn*
[2]*Trường Đại học Washington; Email: srke@uw.edu*

Tóm tắt – Trong bài báo này, chúng tôi tìm hiểu và so sánh hai thuật toán nhận dạng Dynamic Time Warping (DTW) và mô hình Markov ẩn HMM. Trước tiên, từ mỗi khung video, chúng tôi dùng kỹ thuật mô hình hóa cơ thể 3D để ước lượng tư thế người 3D, bao gồm tọa độ 3D của các điểm đặc trưng; rồi chuyển các tọa độ này sang thuộc tính quan hệ hình học GRF, mô tả quan hệ hình học giữa các điểm trong một tư thế nhằm giảm số hướng và gia tăng sự khác biệt giữa các tư thế. Tiếp đến, nhằm giảm số hướng hơn nữa, chúng tôi áp dụng kỹ thuật k-means clustering vào các GRF để tạo ra các vector thuộc tính. Cuối cùng, chúng tôi lần lượt sử dụng DTW và HMM để nhận dạng hành động và so sánh hiệu quả nhận dạng của chúng. Trong hệ thống, để nhận dạng các hành động lặp lại, chúng tôi sử dụng một biến thể của HMM gốc là HMM tuần hoàn CHMM. Các kết quả thực nghiệm trên cơ sở dữ liệu IXMAS cho thấy CHMM nổi trội hơn nhiều so với DTW.

Từ khóa – nhận dạng hành động con người; mô hình hóa người 3D; thuộc tính quan hệ hình học; dynamic time warping (DTW); mô hình Markov ẩn tuần hoàn.

Abstract – In this paper, the use of two well-known recognition algorithms which are Dynamic Time Warping (DTW) and Hidden Markov Model (HMM) are studied and compared. From each frame in monocular videos, we first estimate the 3D human pose which consists of 3D coordinates of specific human joints using an efficient 3D human modeling technique; then convert them into a set of geometrical relational features (GRF), which describe the geometric relations among body joints of a pose for dimensionality reduction and discrimination increase. Next, the k-means clustering technique is applied to those GRFs to generate feature vectors for further dimensionality reduction. Finally, we use DTW and HMM in succession for recognition of actions and then compare their recognition performance. In our system, in order to recognize the repeated actions we use a variation of original HMM which is cyclic HMM (CHMM). The experiment results on IXMAS dataset show that CHMM stands out DTW in terms of recognition rate.

Key words – human action recognition; 3D human modeling; geometrical relational feature; dynamic time warping; cyclic hidden Markov model.

1. Đặt vấn đề

Nhận dạng hành động con người liên quan đến việc phân loại các hành động của con người từ tín hiệu video. Đây là một lĩnh vực nghiên cứu theo hướng "hiểu tín hiệu video" đã được áp dụng khá nhiều trên thế giới như: hệ thống giám sát an ninh thông minh, hệ thống chăm sóc sức khỏe, hệ thống giao thông thông minh, ...

Một hệ thống nhận dạng hành động điển hình gồm hai bước xử lý chính: một là trích thuộc tính và hai là nhận dạng hành động. Bước một tương đương với biến đổi mỗi khung video vào thành một vector thuộc tính đa hướng. Trong bước hai, ta cần xác định (một cách thống kê) chuỗi thuộc tính trích được thuộc vào hành động nào trong các hành động đã biết.

Nhận dạng hành động là một công việc khó khăn và phức tạp do tư thế con người khác nhau tùy thuộc vào góc quay của camera, độ chiếu sáng, nền, quần áo, tốc độ chuyển động, sự che khuất, ... Để nhận dạng chính xác, các thuộc tính cần phải đối phó được với sự thay đổi thang không gian-thời gian, cũng như phải chứa đựng các đặc tính duy nhất của cùng một hành động nhưng thực hiện bởi nhiều người. Vấn đề quan trọng tiếp theo là cần một chiến lược nhận dạng hiệu quả trong không gian thuộc tính có được, nghĩa là, xây dựng việc học có ý nghĩa chỉ với một số lượng mẫu huấn luyện hữu hạn.

Có thể phân loại các thuật toán nhận dạng thành nhận dạng tĩnh và nhận dạng động. Nhận dạng tĩnh không quan tâm đến thông tin thời gian trong tín hiệu, nó được thực hiện dựa vào các khung trọng yếu (key frames) trích ra từ chuỗi các khung video vào theo một tiêu chí nào đó. Ngược lại, nhận dạng động có quan tâm đến thông tin thời gian trong tín hiệu video. Nhận dạng động bao gồm phương pháp so khớp mẫu và dùng mô hình không gian trạng thái. Trong phương pháp so khớp mẫu, chuỗi vector thuộc tính vào được so sánh theo từng khung với chuỗi vector thuộc tính huấn luyện để tìm ra sự tương tự. Phương pháp dùng mô hình không gian trạng thái biểu diễn mỗi hành động bằng một mô hình gồm nhiều trạng thái, mỗi trạng thái tương đương một tư thế trong hành động. Để nhận dạng hành động, ta tính likelihood giữa mô hình và hành động quan sát rồi quyết định hành động nhận dạng được chính là hành động tương ứng với mô hình cho likelihood cao nhất.

Nhiều phương pháp mới về nhận dạng hành động con người từ tín hiệu video đề xuất trong những năm gần đây đã cho những kết quả rất khả quan. Chẳng hạn, trong phương pháp [7], D. Weinland và cộng sự thực hiện mô hình hóa các hành động bằng lưới 3D xây dựng từ các ảnh quay từ nhiều camera. Sau đó, các mẫu 3D này được dùng để tạo ra các khung hình bóng 2D dùng cho nhận dạng. Phương pháp này bị phụ thuộc vào góc quay của camera. Trong phương pháp [8], I. N. Junejo và cộng sự đã đề xuất dùng ma trận tự tương tự (self-similarity matrix). Ma trận này được tính từ khoảng cách giữa các thuộc tính trích từ từng cặp khung trong chuỗi hành động theo thời gian. Ma trận này đã được chứng minh là ổn định đối với sự thay đổi góc quay của camera, tuy nhiên vấn đề che khuất chưa được giải quyết tốt.

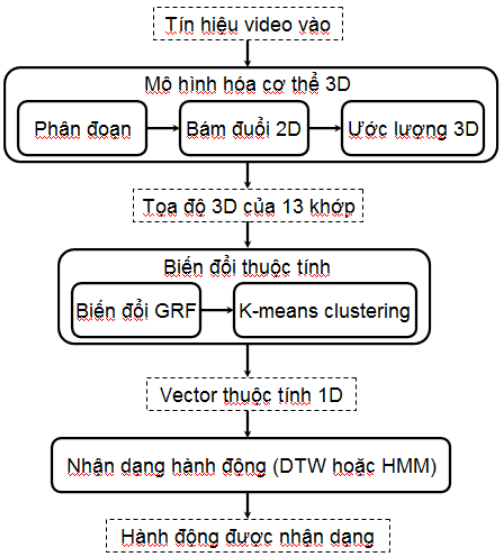
Trong bài báo này, chúng tôi đề xuất mới một hệ thống nhận dạng hành động con người trong đoạn video quay bằng một camera. Để đối phó với sự thay đổi độ chiếu sáng, quần

áo, góc quay và sự che khuất, chúng tôi lựa chọn một kỹ thuật mô hình hóa cơ thể 3D hiệu quả, giúp ước lượng tốt các tọa độ 3D của các điểm đặc trưng; sau đó biến đổi các tọa độ 3D này thành tập thuộc tính quan hệ hình học (GRF) rồi phân nhóm dùng thuật toán k-means clustering. Trong khâu nhận dạng, chúng tôi chọn hai thuật toán tiêu biểu cho phương pháp so khớp mẫu là Dynamic Time Warping (DTW) và tiêu biểu cho mô hình không gian trạng thái là Hidden Markov Model (HMM).

Nội dung chính phần tiếp theo của bài báo gồm: Mục 2 trình bày hệ thống do chúng tôi đề xuất, Mục 3 báo cáo các thí nghiệm và đánh giá kết quả, cuối cùng là kết luận ở Mục 4.

2. Hệ thống nhận dạng hành động đề xuất

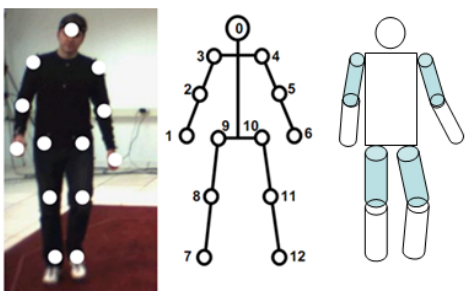
Hình 1 mô tả tổng quan hệ thống đề xuất, bao gồm 3 khối chức năng chính: mô hình hóa cơ thể 3D, biến đổi thuộc tính và nhận dạng hành động. Các mục sau sẽ lần lượt trình bày chi tiết các bước trên.



Hình 1: *Tổng quan hệ thống đề xuất*

2.1. Mô hình hóa cơ thể 3D

Mô hình hóa cơ thể 3D dùng trong hệ thống được thực hiện theo phương pháp [1] do các ưu điểm nổi trội của nó. Mô hình 3D bao gồm phần đầu, mình và tứ chi. Đầu được biểu diễn bằng hình tròn, mình được biểu diễn bằng hình chữ nhật, mỗi chi được biểu diễn bằng hai hình trụ: một cho phần trên và một cho phần dưới của chi như Hình 2.

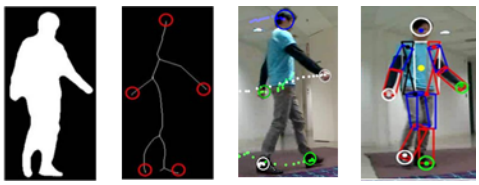


Hình 2: *Từ trái sang: ảnh gốc, 13 điểm đặc trưng, mô hình cơ thể 3D.*

Kết quả đầu ra của khối mô hình hóa 3D là tọa độ 3D của 13 điểm đặc trưng trong cơ thể gồm đầu, hai tay, hai khuỷu tay, hai vai, hai chân, hai đầu gối và hai hông như Hình 2, được ước lượng từ mỗi khung hình trong chuỗi khung video vào.

Trong khối mô hình cơ thể 3D có 3 bước chính là: phân đoạn đối tượng nhằm trích con người chuyển động ra khỏi nền, cùng với các thuộc tính 2D gồm ảnh gốc, ảnh bóng, ảnh viền và ảnh chuyển động như trên Hình 3; bám đuổi 2D nhằm xác định và bám theo vị trí của 5 điểm là đầu, 2 bàn tay và 2 bàn chân qua từng khung hình (Hình 3); ước lượng 3D nhằm ước lượng thô mô hình 3D ban đầu, sau đó ước lượng tinh nhằm tìm mô hình 3D tốt nhất dựa vào các vị trí của 5 điểm đặc trưng nói trên (Hình 4).

Hình 3: *Từ trái sang: ảnh gốc, ảnh bóng, ảnh viền và ảnh chuyển động.*



Hình 4: *Từ trái sang: ảnh gốc, vị trí của 5 điểm, quỹ đạo của 5 điểm, mô hình ước lượng 3D .*

2.2. Biến đổi thuộc tính

Bước xây dựng cơ sở dữ liệu rất quan trọng, ảnh hưởng lớn đến toàn bộ quá trình nhận dạng sau này. Trong bước này, ta tiếp tục biến đổi tập tọa độ 3D của 13 điểm đặc trưng ước lượng từ mỗi khung video nói trên thành một vector thuộc tính. Hai bước biến đổi được thực hiện tại đây bao gồm biến đổi thành thuộc tính quan hệ hình học (GRF) 15 hướng và thực hiện k-means clustering. Mục đích của bước này là giảm số hướng của vector thuộc tính. Cụ thể nếu dùng trực tiếp tọa độ 3D của 13 điểm thì vector thuộc tính sẽ là 13 x 3 = 39 hướng, còn nếu biến đổi GRF sẽ giảm còn 15 hướng. Tuy số hướng giảm nhưng GRF đã được chứng minh là gia tăng sự khác biệt giữa các tư thế của cơ thể, dẫn đến tăng khả năng nhận dạng [2].

Thuộc tính GRF mô tả quan hệ vị trí giữa các điểm đặc trưng của cơ thể. Bộ mô tả thuộc tính GRF sử dụng trong hệ thống gồm 15 thuộc tính như trình bày trong Bảng 1. Thuộc tính GRF gồm hai loại là thuộc tính khoảng cách (F1 $\sim$ F9) và thuộc tính góc (F10 $\sim$ F15).

Xét thuộc tính khoảng cách F_1 làm ví dụ: dấu của F_1 cho biết tay phải ở trước hay sau so với mặt phẳng tạo bởi vai phải, hông phải và hông trái; độ lớn của F_1 cho biết khoảng cách xa gần giữa tay phải và mặt phẳng này.

Bảng 1: *Chi tiết thuộc tính GRF 15 hướng.*

Thuộc tính	Mô tả
$F_{1,2}$	Khoảng cách có dấu giữa tay phải / trái và mặt phẳng xác định bởi vai phải / trái, hông phải và hông trái
$F_{3,4}$	Khoảng cách có dấu giữa chân phải / trái và mặt phẳng xác định bởi vai phải, vai trái và hông phải/trái
$F_{5,6}$	Khoảng cách dấu giữa tay phải / trái và mặt phẳng xác định bởi vai phải / trái và pháp vector đầu – điểm giữa hai hông
F_7	Khoảng cách giữa trọng tâm cơ thể và chân thấp nhất theo hướng Y
F_8	Khoảng cách giữa hai bàn chân theo hướng Y
F_9	Khoảng cách tích lũy giữa trọng tâm cơ thể ở khung hiện tại và khung đầu tiên
$F_{10,11}$	Góc giữa cẳng tay và cánh tay phải / trái
$F_{12,13}$	Góc giữa đùi và bắp chân phải / trái
F_{14}	Góc gập của cơ thể dọc theo hướng X
F_{15}	Sự thay đổi của góc quay ngang của cơ thể giữa khung hiện tại và khung trước đó

Tiếp theo, các vector GRF 15 hướng được phân nhóm dùng thuật toán k-means clustering [3]. Mỗi vector GRF được chuyển thành một từ mã trong số k từ mã, hay còn gọi là một ký hiệu, dựa trên cơ sở lân cận gần nhất. Như vậy, mỗi khung video vào được chuyển thành một ký hiệu trong số k ký hiệu, và chuỗi khung video vào lúc này được biểu diễn bằng một chuỗi vector thuộc tính 1 hướng.

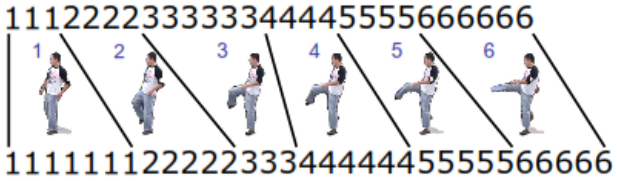
2.3. Nhận dạng hành động

Như đã nói trên, trong khâu nhận dạng, chúng tôi sử dụng hai thuật toán nhận dạng phổ biến là: (1) Dynamic Time Warping (DTW) tiêu biểu cho nhóm phương pháp so khớp mẫu và (2) mô hình Markov ẩn (HMM) tiêu biểu cho mô hình không gian trạng thái nhằm so sánh hiệu quả nhận dạng giữa chúng.

2.3.1. Dynamic Time Warping (DTW)

DTW là một phương pháp so khớp mẫu điển hình. Thường thì con người thực hiện hành động với các tốc độ nhanh chậm khác nhau. Do vậy, việc đánh giá sự tương tự giữa hành động mẫu có sẵn với hành động cần nhận dạng cần phải xem xét đến sự khác biệt này. Trước tiên, DTW biểu diễn chuỗi khung video hành động mẫu có sẵn bằng một chuỗi các vector thuộc tính tham chiếu. Khi chuỗi khung video chứa hành động cần nhận dạng đưa vào thì chuỗi vector thuộc tính trích được từ đây sẽ được so sánh với chuỗi vector thuộc tính tham chiếu để xác định độ tương tự. Độ tương tự cao nhất (hay là khoảng cách nhỏ nhất) được chọn làm tiêu chuẩn để nhận dạng hành động.

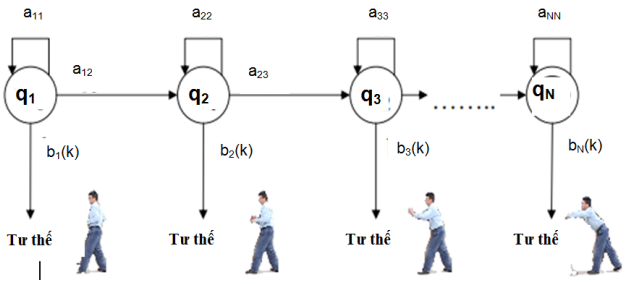
DTW giải quyết sự sai khác tốc độ giữa hai chuỗi bằng các phép toán xóa-chèn, nén-giãn, và thay thế. Ưu điểm của DTW là đơn giản và hiệu quả chấp nhận được với điều kiện thứ tự thời gian của chuỗi cần nhận dạng giống với thứ tự thời gian của chuỗi tham chiếu. Hình 5 minh họa sự so khớp giữa hai chuỗi có tốc độ thực hiện khác nhau [4]. Mỗi con số ở đây biểu diễn một tư thế.



Hình 5: *Ví dụ về so khớp hai chuỗi "đá" với hai tốc độ thực hiện khác nhau [4].*

2.3.2. Mô hình Markov ẩn (HMM)

HMM là một mô hình không gian trạng thái điển hình, vốn rất phổ biến trong nhận dạng tiếng nói [5]. Cấu trúc của một HMM gồm một chuỗi Markov ẩn và một tập hữu hạn các phân bố xác suất đầu ra. Cụ thể là, một HMM được xác định bởi một tập 3 ma trận $\lambda = \{A, B, \pi\}$, trong đó A = ma trận chuyển tiếp = $\{a_{ij}\}$, với a_{ij} là xác suất chuyển từ trạng thái q_i sang q_j, $(i, j) \in [1 : N)$; B = ma trận quan sát = $\{b_j(k)\}$, với $b_j(k)$ là xác suất ký hiệu ra v_k (rời rạc) quan sát được tại trạng thái $q_j, k \in [1 : M); \pi = \{\pi_i\}$, với π_i là xác suất trạng thái khởi đầu. Để nhận dạng hành động, ta cần huấn luyện một HMM cho mỗi hành động. Trong giai đoạn huấn luyện, cần xác định số trạng thái của một HMM, tối ưu hóa xác suất chuyển đổi trạng thái và xác suất ký hiệu quan sát để các ký hiệu tạo ra có thể tương ứng với các vector thuộc tính của chuỗi khung video huấn luyện. Trong giai đoạn kiểm tra, ta tính xác suất mà một HMM cụ thể có thể tạo ra chuỗi ký hiệu kiểm tra tương ứng với vector thuộc tính trích từ khung video kiểm tra, để đo likelihood giữa mô hình và chuỗi khung video kiểm tra. Likelihood cực đại được chọn làm tiêu chuẩn để nhận dạng các hành động.

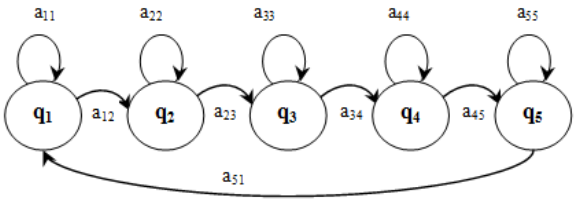


Hình 6: *Ví dụ về HMM trái-phải biểu diễn hành động "với tay"[4].*

Hình 6 đưa ra một ví dụ về HMM gốc [4]. Mỗi ảnh người trong hình biểu diễn một tư thế mà xác suất quan sát được tư thế đó $-b_j(k)$là cao nhất trong mỗi trạng thái q_j.

2.3.3. Mô hình Markov ẩn tuần hoàn (CHMM)

Trong các hành động cần nhận dạng có thể có các hành động có tính lặp đi lặp lại gần theo chu kỳ như đi bộ, vẫy tay,... Để nhận dạng các hành động này, thay vì sử dụng HMM gốc, chúng tôi đề xuất sử dụng HMM tuần hoàn CHMM – một biến thể của HMM gốc [6]. HMM tuần hoàn là HMM gốc 5 trạng thái có thêm chuyển tiếp từ trạng thái cuối về trạng thái đầu tiên như trong Hình 7, tức là xác suất $a_{51} \neq 0$ (trong HMM gốc thì $a_{51} = 0$) Chuyển tiếp này biểu diễn kết thúc của một chu kỳ và bắt đầu một chu kỳ mới trong một hành động lặp lại.



Hình 7: *Mô hình CHMM dùng trong hệ thống đề xuất*

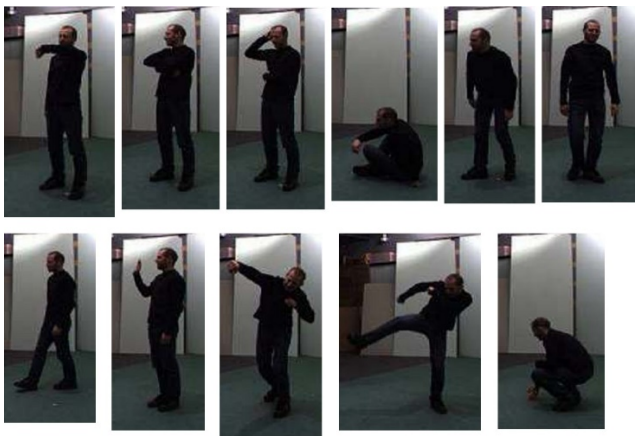
3. Thí nghiệm và đánh giá kết quả

Để đánh giá hệ thống đề xuất và so sánh hiệu quả nhận dạng của hai phương pháp DTW và CHMM, chúng tôi tiến hành hai thí nghiệm trên cơ sở dữ liệu IXMAS [7]. Cả hai thí nghiệm đều dùng chung phương pháp trích vector thuộc tính như đã trình bày trong các mục 2.1, 2.2 nhưng khác phương pháp nhận dạng, một thí nghiệm dùng DTW và một thí nghiệm dùng CHMM.

3.1. Cơ sở dữ liệu IXMAS

Cơ sở dữ liệu IXMAS được thực hiện bởi 12 người, mỗi người thực hiện 11 hành động là: xem giờ (check_watch), tréo tay (cross_arm), gãi đầu (scratch_head), ngồi xuống (sit_down), đứng lên (get_up), xoay người (turn_around), đi bộ (walk), vẫy tay (wave), đấm (punch), đá (kick), và cúi nhặt đồ vật (pick_up). Tín hiệu video thu được từ 5 camera. Hệ thống đề xuất nhằm nhận dạng hành động từ tín hiệu video quay bằng một camera nên chúng tôi chỉ chọn một camera là camera 3 cho tất cả các thí nghiệm.

Hình 8 là ảnh của 11 hành động trong cơ sở IXMAS.



Hình 8: *Các ảnh trong cơ sở IXMAS*

3.2. Thí nghiệm

Như đã trình bày ở trên, trước tiên, mỗi khung video vào được chuyển thành một vector 39 hướng (là tọa độ 3D của 13 điểm), rồi thành một vector GRF 15 hướng. Sau đó, chúng tôi thực hiện k-means clustering với k = 64 để chuyển vector GRF này thành một trong số 64 vector 1 hướng.

Trong thí nghiệm 1, chúng tôi sử dụng thuật toán nhận dạng là DTW. Chuỗi tham chiếu/nhận dạng là chuỗi các vector 1 hướng trong số 64 vector có thể có. Tiêu chuẩn nhận dạng là khoảng cách nhỏ nhất giữa chuỗi tham chiếu và chuỗi cần nhận dạng. Kết quả thí nghiệm được thể hiện trên ma trận trong Bảng 2.

Trong thí nghiệm 2, chúng tôi sử dụng cùng vector thuộc tính với thí nghiệm 1 nhưng thuật toán nhận dạng là CHMM. Chúng tôi chia các đoạn video trong cơ sở dữ liệu thành 5 phần, đánh số từ 1 đến 5; sau đó dùng phần 2-5 cho huấn luyện và phần 1 để kiểm tra; và làm như thế cho đến hết. Kết quả thí nghiệm được thể hiện trên ma trận trong Bảng 3.

Bảng 2: *Ma trận kết quả thí nghiệm với DTW.*

Single_Action_DTW (%)	check_watch	cross_arm	scratch_head	sit_down	get_up	turn_around	walk	wave	punch	kick	pick_up
check_watch	83	0	0	0	0	0	0	17	0	0	0
cross_arm	0	75	17	0	0	0	0	8	0	0	0
scratch_head	8	8	50	0	0	0	0	8	8	17	0
sit_down	8	0	0	83	8	0	0	0	0	0	0
get_up	0	0	0	0	100	0	0	0	0	0	0
turn_around	0	8	0	0	0	75	8	0	0	0	8
walk	8	0	0	0	0	8	75	0	0	8	0
wave	8	0	17	0	0	0	0	75	0	0	0
punch	8	8	8	0	0	0	8	17	42	8	0
kick	8	8	17	0	0	0	8	17	25	17	0
pick_up	0	0	0	0	8	8	8	0	0	0	75

Bảng 3: *Ma trận kết quả thí nghiệm với CHMM.*

Single_Action (%)	check_watch	cross_arm	scratch_head	sit_down	get_up	turn_around	walk	wave	punch	kick	pick_up
check_watch	92	0	0	0	0	0	0	0	8	0	0
cross_arm	0	92	0	0	0	0	0	0	8	0	0
scratch_head	0	0	92	0	0	0	0	0	8	0	0
sit_down	0	0	0	100	0	0	0	0	0	0	0
get_up	0	0	0	0	100	0	0	0	0	0	0
turn_around	0	0	0	0	0	75	25	0	0	0	0
walk	0	0	0	0	0	0	100	0	0	0	0
wave	8	0	8	0	0	0	0	83	0	0	0
punch	0	0	0	0	0	0	0	0	92	8	0
kick	0	0	0	0	0	0	0	0	0	100	0
pick_up	0	0	0	17	0	0	0	0	0	0	83

3.3. So sánh và đánh giá

Từ kết quả thí nghiệm trong Bảng 2 và 3 ta thấy: với cùng thuộc tính và thí nghiệm trên cùng cơ sở dữ liệu thì tỷ lệ nhận dạng trung bình của CHMM là 91.7% và của DTW là 68.2%.

Như vậy, mô hình Markov ẩn tuần hoàn (CHMM) nổi trội hơn hẳn Dynamic Time Warping (DTW) cho nhận dạng hành động.

Ngoài ra, để đánh giá hệ thống đề xuất, chúng tôi cũng đã tiến hành so sánh với một vài hệ thống nhận dạng gần đây [7], [8] trên cùng cơ sở dữ liệu.

Hệ thống [7] có tỷ lệ nhận dạng là 80.5%, hệ thống [8] cho tỷ lệ nhận dạng là 71.2%; trong khi hệ thống đề xuất (kết hợp thuộc tính GRF và nhận dạng dùng CHMM) cho tỷ lệ nhận dạng là 91.7%. Điều này chứng tỏ tỷ lệ nhận dạng của hệ thống đề xuất cao hơn hẳn.

4. Kết luận

Kỹ thuật nhận dạng hành động từ tín hiệu video được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau của cuộc sống hiện đại. Trong bài báo này, chúng tôi đã phân tích, lựa chọn và kết hợp hiệu quả các kỹ thuật mô hình hóa cơ thể 3D, chuyển đổi thuộc tính quan hệ hình học GRF, phân nhóm k-means và mô hình Markov ẩn tuần hoàn CHMM với nhau, tạo nên hệ thống nhận dạng có kết quả rất khả quan. Ngoài ra, bài báo cũng đã thực hiện so sánh tỷ lệ nhận dạng của hệ thống đề xuất với các hệ thống mới khác, cho thấy ưu điểm hơn hẳn của hệ thống đề xuất.

Tài liệu tham khảo

[1] Shian-Ru Ke, Jenq-Neng Hwang, Kung-Ming Lan, and Shen-Zheng Wang, *"View-Invariant 3D Human Body Pose Reconstruction using a Monocular Video Camera,"* Proc. IEEE ICDSC, 2011, pp. 1-6.

[2] Hoang Le Uyen Thuc, Pham Van Tuan, and Jenq-Neng Hwang, *"An Effective 3D Geometric Relational Feature Descriptor for Human Action Recognition,"* Proc. IEEE RIVF, 2012, pp. 270-275.

[3] John A. Hartigan and Manchek A. Wong, *"Algorithm AS 136: A k-means clustering algorithm,"* Applied statistics, 1979, pp. 100-108.

[4] J. K. Aggarwal and M. S. Ryoo, *"Human Activity Analysis: A Review,"* ACM Computing Surveys, vol. 43 (3), 2011.

[5] Lawrence R. Rabiner, *"A tutorial on hidden Markov models and selected applications in speech recognition,"* Proc. IEEE, vol. 77(2), 1989, pp. 257-286.

[6] Hoang Le Uyen Thuc, Shian-Ru Ke, Jenq-Neng Hwang, Pham Van Tuan, Truong Ngoc Chau, *"Quasi-Periodic Action Recognition from Monocular Videos via 3D Human Models and Cyclic HMMs,"* Proc. IEEE ATC, 2012, pp. 110-113.

[7] D. Weinland, E. Boyer, R. Ronfard, *"Action Recognition from Arbitrary Views using 3D Exemplars,"* Proc. IEEE ICCV, 2007, pp. 1-7.

[8] Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, *"Learning Realistic Human Actions from Movies,"* Proc. IEEE CS Conf. Computer Vision and Pattern Recognition, 2008, pp. 1-8.

[9] I.N. Junejo, E. Dexter, I. Laptev, P. Perez, *"View-Independent Action Recognition from Temporal Self-Similarities"*, IEEE Transactions on PAMI, vol. 33, no. 1, 2011, pp. 172-185.

(BBT nhận bài: 15/12/2013, phản biện xong: 29/12/2013)