*Research Article*

INTERNATIONAL JOURNAL OF ADVANCED ROBOTIC SYSTEMS

*International Journal of Advanced Robotic Systems*
May-June 2017: 1–8

DOI: 10.1177/1729881417704544
journals.sagepub.com/home/arx
SAGE

# Golf video tracking based on recognition with HOG and spatial–temporal vector

**Li Weixian[1,2], Lou Xiaoping[1,2], Dong Mingli[1,2] and Zhu Lianqing[1,2]**

**Abstract**

The hand and club movements contain golfer's swing information, which can be obtained to provide good visualization to be shared on Internet and be summarized in golf studying. In this article, a hand and club tracking framework based on recognition with a complex descriptor combining histograms of oriented gradients and spatial–temporal vector is proposed to obtain their movement trajectories in golf video. After the hand and club are recognized in initial windows defined by the body region, a boosted classifier trained by the proposed descriptor is utilized for recognition and tracking in a searching window predicted by trajectory fitting with previous four object positions. Experiments show that the boosted classifier can have a precision and recall rate both better than 97%, and the hand and club tracking are basically correct in our testing videos.

**Keywords**

Tracking, object recognition, trajectory prediction, golf, HOG

Date received: 27 December 2016; accepted: 26 February 2017

Topic: Special Issue - Multi-Modal Fusion for Robotics
Topic Editor: Huaping Liu
Associate Editor: Huaping Liu

## Introduction

With the development of the mobile communication and smart devices, sharing of sports experience on Internet is becoming very popular these days, such as football or basketball videos uploaded on Facebook or WeChat. Not just a lifestyle, those short videos also contain movement information of both players' arms/legs and driving implements, like a club, bat, cue, racket, hand, or foot, which can be used to coach beginners as to how to use their physical strength to improve their grades in the sports, which has a wild prospect in the sports training market.[1–6] In this article, we focus on obtaining swing movements by object tracking from golf videos to provide good visualization to be shared on Internet and meaningful data for study in golf training.

Object tracking is one of the most researched topics in computer vision and machine learning. In golf videos, the hand and club, showing movements of players' strength and swing, are our interested objects and they often move extremely fast and deform drastically, which are fundamental obstacles that all object tracking methods face. To be clear, the imaging quality in our golf videos is limited by many factors:

1. Compared with the whole players' body, the hand and club appear very small. With a video resolution of 720 × 1028 pixels (as shown in Figure 1 (with shielded face)), the hand patch is imaged at a

[1] Beijing Information Science & Technology University, Beijing Key Laboratory of Optoelectronic Measurement Technology, Beijing, China
[2] Beijing Information Science & Technology University, School of Instrumentation Science and Optoelectronic Engineering, Beijing, China

**Corresponding author:**
Lou Xiaoping, Beijing Information Science & Technology University, Beijing Key Laboratory of Optoelectronic Measurement Technology, 12 Qinghe Xiaoying East Rd, Haidian District, Beijing 100192, China.
Email: louxiaoping@bistu.edu.cn

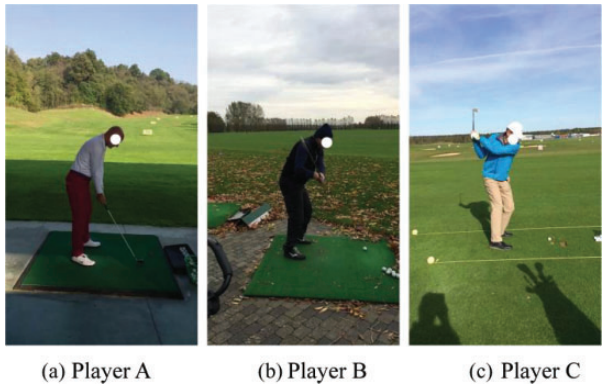


**Figure 1.** Frames of three players from videos with a resolution of 720 × 1280 pixels shot from the back side of the players.

resolution of 63 × 63 pixels (Figure 2) and the club patch at 48 × 48 pixels (Figure 3), both are short of details.

2. The appearances of the hand and club in one swing are deformed in sequence. Each row in Figures 2 and 3 are typical positions of one player along the backswing and downswing.
3. The swing speed can be as fast as 150 miles/h and the club can be hardly seen on the fast downswing. The club of player B is blurred in Figure 1(b) and on the last two columns on second row in Figure 3.
4. The camera in mobile phone has low performance in dynamic shooting. All these videos in our application are shot by mobile phones from the back side of the players.

**Figure 2.** Sequent appearances of the hands of the three players on a golf swing. The hand patch is sized at 63 × 63 pixels. The first row is the hands of player A, the second is player B, and the last is player C. The first four columns in every row are with the hands at the very beginning, above the knee, over the shoulder, and at the peak of a backswing, respectively. The last four columns are with the hands at the beginning, at the peak, over the shoulder, and down to the knee of a faster downswing, respectively.

**Figure 3.** Sequent appearances of the clubs of the three players on a golf swing. The club patch is sized at 48 × 48 pixels. The first row is the clubs of player A, the second is player B, and the last is player C. The first four columns on every row are with the clubs at the very beginning, above the knee, over the shoulder, and at the peak of a backswing, respectively. The last four columns are with the clubs at the beginning, at the peak, over the shoulder, and down to the knee of a faster downswing, respectively. Each one here is from the same frame with the one in the corresponding position in Figure 2.

Therefore, in this article, we aim to find the characteristics of the concerned objects and propose a tracking framework for golf video based on object recognition using machine learning with histograms of oriented gradients (HOG) and spatial–temporal vector.

The tracking framework is introduced in the second section, and performance evaluation about recognition and tracking is reported in the third sections. In the fourth section, we summarize our method and discuss problems.

## Tracking framework

The tracking framework for golf video consists of three main parts: *initialization*, *object trajectory prediction*, and *object recognition*. Initialization finds the initial positions of the hand and club in the video, object trajectory prediction estimates the possible object position in the current frame by trajectory fitting with their previous recognized positions, and object recognition recognizes the object in a searching window centered on that predicted position using machining learning. Detailed algorithms are listed below.

### *Initialization*

At initialization, the main task is to give candidate windows for the objects and then object recognition can be applied to detect the initial hand and club just in the candidate windows. Since the initial objects (hand and club) are usually in fixed positions relatively to the player's body, we designed an initialization strategy as *player's body–object window–object*.

First, the player's body is detected automatically according to Dollar et al.[7] with aggregated channel feature (ACF), which is not described in detail in this article.

Second, if $\mathrm{Rect}\{x_b, y_b, w_b, h_b\}$ is the denotation of the player's body position, the hand window denoted as $\mathrm{Rect}\{x_h, y_h, w_h, h_h\}$ and the club window denoted as $\mathrm{Rect}\{x_c, y_c, w_c, h_c\}$ are defined as follows from experience.

$$\begin{cases} x_h = x_b + 0.35w_b \\ y_h = y_b + 0.4h_b \\ w_h = 0.5w_b + 64 \\ h_h = 0.4h_b \end{cases} \tag{1}$$

$$\begin{cases} x_c = x_b + 0.65w_b \\ y_c = y_b + 0.8h_b \\ w_c = 1.45w_b \\ h_c = 0.3h_b \end{cases} \tag{2}$$

Third, the hand and club patch are recognized by object recognition in the above defined object windows.

### *Object trajectory prediction*

Since the third part object recognition is time-consuming and resource consuming, this part estimates object possible position on the current frame by their previous trajectory to decrease processing area in object recognition. Because the golf swing is an approximate circle movement centered on shoulder joint, the trajectories of the hand and club are not irregular. Figure 4 shows trajectories of hand and club with two totally different players. The whole trajectories are complicate curves hard to be expressed, but the neighbor positions in the trajectories are regular.

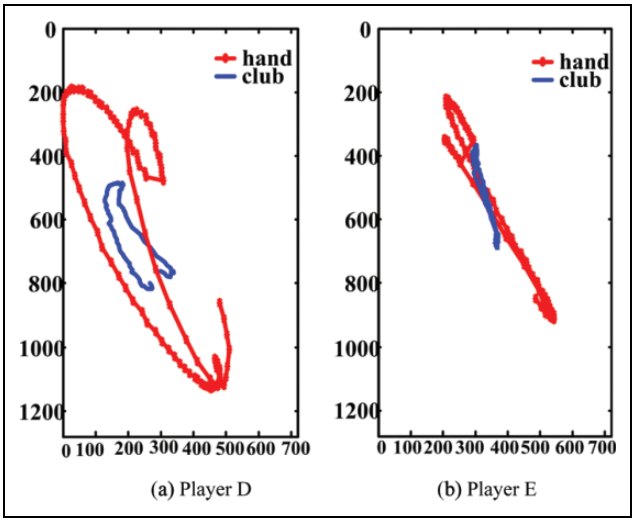


**Figure 4.** Hand and club trajectories of two players.

To simplify the prediction progress, we assume that the object's local trajectories are quadratic, that is to say the coordinates in $x$ and $y$ axes are quadratic polynomial in frame $i$ (time $t$), respectively. Suppose $(a_i, b_i, c_i, k_i, l_i, m_i)$ are the quadratic coefficients on $i$th frame, the object predicted position $(x_i, y_i)$ is denoted as

$$\begin{cases} x_i = a_i i^2 + b_i i + c_i \\ y_i = k_i i^2 + l_i i + m_i \end{cases} \tag{3}$$

The coefficients $(a_i, b_i, c_i, k_i, l_i, m_i)$ can be estimated by the local four previous tracked positions $(x_{i-1}, y_{i-1}), \ldots,$ and $(x_{i-4}, y_{i-4})$ with $i > 4$.

$$\begin{bmatrix} (i-1)^2 & i-1 & 1 \\ (i-2)^2 & i-2 & 1 \\ (i-3)^2 & i-3 & 1 \\ (i-4)^2 & i-4 & 1 \end{bmatrix} \begin{bmatrix} a_i \\ b_i \\ c_i \end{bmatrix} = \begin{bmatrix} x_{i-1} \\ x_{i-2} \\ x_{i-3} \\ x_{i-4} \end{bmatrix} \tag{4}$$

$$\begin{bmatrix} (i-1)^2 & i-1 & 1 \\ (i-2)^2 & i-2 & 1 \\ (i-3)^2 & i-3 & 1 \\ (i-4)^2 & i-4 & 1 \end{bmatrix} \begin{bmatrix} k_i \\ l_i \\ m_i \end{bmatrix} = \begin{bmatrix} y_{i-1} \\ y_{i-2} \\ y_{i-3} \\ y_{i-4} \end{bmatrix} \tag{5}$$

### *Object recognition*

*Feature descriptor.* Many methods are used to find and identify objects in an image or video with the fact that the

objects may vary in different illuminations, scales, or viewpoints, such as the popular feature-based approach, which is also our choice in this article. What's more, we prefer simple feature descriptor than a complex one to make it possible to transplant the tracking framework to mobile phone. On the other hand, the complicate descriptors for rotation and scale problem, such as scale-invariant feature transform (SIFT)[8,9] and speeded up robust feature,[10] are not the case in our golf videos with barely any scale and rotation change. Because simple descriptors, Haar[11,12] and local binary pattern),[13] are more suitable to feature objects with rich texture, HOG,[14] sensitive to outline, is adopted here.

However, as mentioned before, image quality of the objects is limited and object appearances are projected differently in one swing. We found learning only by HOG is not applicable and need to fuse other useful information. There are two other unignorable spatial and temporal clues that are useful in golf videos. The spatial one is the appearances of the hand and club which are relatively similar to the position of one golf swing. Figure 2 shows in the early of backswing and in the late of downswing (columns 1, 2, and 8), the fist back is shot, and in the late of backswing and in the early of downswing (columns 4, 5, and 6) the fist side is projected. On the other hand, the temporal clue is the appearance sequences of the hand and club are similar in every video, and their appearances can be implied by the previous frame. Figures 2 and 3 show the different players' hand or club appearance changes similarly in sequence.

Therefore, the contribution of this article is to combine HOG's sensitivity with object's outline and relationship of object appearance with space and time. Unlike the fusion of different channels, such as visual–tactile information or visual–audio information,[15–21] we turn the object spatial and temporal information into feature vectors like HOG, and a complex feature descriptor is proposed with HOG and spatial–temporal vector, which is referred to as $[HOG_i \quad HOG_i - HOG_{i-1} \quad \mathbf{X}_i - \mathbf{X}_0]$. $HOG_i$ is the HOG vector of the object patch in $i$th frame with a dimension of $N$. $\mathbf{X}_0 = [x_0 \quad y_0]$ are the initial position coordinates of the object and $\mathbf{X}_i = [x_i \quad y_i]$ are the position coordinates of the object in $i$th frame. Our proposed feature descriptor is in a dimension of $2N + 2$.

*Training.* Based on the above feature descriptor, we follow the adaptive boosting algorithm with the help of OpenCV and a boosted classifier is trained.[22,23] The training configuration and classifier performance will be shown in the second section.

*Recognition.* Since the real object may be not on the predicted position $(x_i, y_i)$, recognition according to the scores calculated based on the boosted classifier is carried out in the sliding window which is centered on $(x_i, y_i)$ and has a larger width and height than the patch. In this application, we check the score of the patch on $(x_i, y_i)$ at first, and if the score is not qualified, every patch in the sliding window will be checked until the ideal score is found.

**Table 1.** Configuration of the training database.

| Properties | Hand | Club |
|---|---|---|
| Number of videos | 99 | 99 |
| Number of positive sample | 13,287 | 13,287 |
| Number of negative sample | 147,671 | 147,671 |
| Patch size (pixel) | 63 × 63 | 48 × 48 |
| HOG dimension of a patch | 1296 | 900 |
| Proposed descriptor dimension of a patch | 2594 | 1802 |

HOG: histograms of oriented gradients.

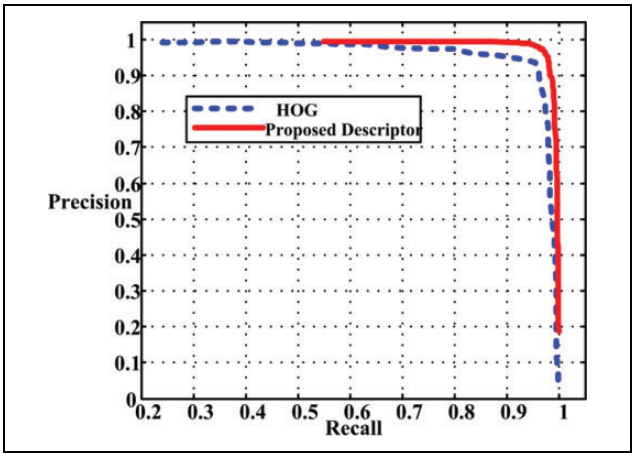


**Figure 5.** Precision–recall performance.

## Results

The tracking based on the proposed descriptor with HOG and spatial–temporal vector are carried out, and their performances are analyzed below.

### *Training performance*

As shown in Table 1, since the hand and club are sized as 63 × 63 and 48 × 48, respectively, the dimension of the proposed descriptor of the hand and club is 2594 and 1802, respectively. Our training database consists of 13,287 positive samples and 147,671 negative samples from 99 videos. These samples are divided randomly into two parts: 80% of the positive and negative samples are used only for training an adaptive boosted classifier, and the remaining 20% for testing the performance of the classifier.

A score for every test sample can be obtained using the boosted classifier and is recognized as object or not by comparing it with a threshold. Figure 5 gives the precision and recall curves of classifiers based on conventional HOG and the proposed descriptor. Each point responds to a precision value (vertical) and recall value (horizontal) calculated with a score threshold and a curve is drawn when the score threshold changes from −15.0 to 20.0 with a step of

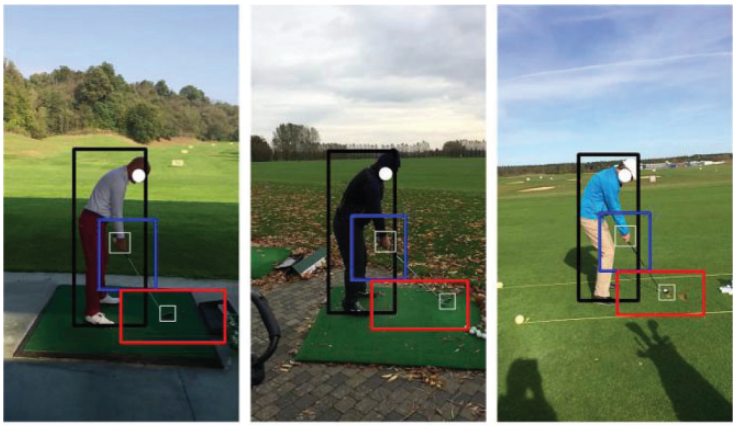


**Figure 6.** *Initialization* results: player's body, object window, and object.

**Figure 7.** Hand and club tracking of player A. The body position is in black box. The initial hand and club are in white patch, the current hand and club in the current frame are in blue and red patch, respectively. The first row is from backswing: The blue dots and red dots constitute trajectories of the hand and club, respectively. The second row is from downswing: The yellow dots and green dots constitute trajectories of the hand and club, respectively.

0.5. The red curve shows sharper descending when recall converges toward 1, which means both a higher precision rate and a higher recall rate can be designed with this specific score threshold, and is supposed to have better performance than the former one. That is the reason we propose this complex descriptor based on HOG and spatial–temporal vector in this article. The score threshold corresponding to the point (0.9706, 0.9706) on the red

**Figure 8.** Hand and club tracking of player B. The illustrations are the same as Figure 7.

curve in Figure 5 is 7.5 and will be applied in the following recognition and tracking.

### *Tracking performance*

Figure 6 is the initialization results with the strategy of player's body–object window–object. The black box is the detected player's body using ACF; the blue and red box is the object windows that are defined by Equations (1) and (2); the small white boxes are the hand and club patch recognized by the trained boosted classifier in the above defined object windows. Results show that the initial hand and club can be correctly found.

The tracking framework is carried out for videos of players A, B, and C with our proposed method. Because the hands have smaller movements, the hands tracking of players A, B, and C are totally correct in blue and yellow trajectories in Figures 7–9, respectively. However, the clubs can move very fast and the tracking are just basically correct. Player A is imaged well and the club tracking is totally correct. Player B is imaged at dawn and the club tracking failed in 4 frames of the video with 280 frames. As shown in the seventh image of Figure 8, the fast club is seriously blurred in the late of the downswing and is not recognized. Player C is imaged very well but the club tracking failed in 4 frames of the video with 140 frames. As shown in the seventh image of Figure 9, the club can be hardly seen by human when it is moving in front of the leg and the club cannot be tracked.

In our application, if a patch with ideal score is found in the sliding window, we just believe that the object doesn't exist and the program skips to the next frame. This will avoid misrecognition and make the tracking trajectories reliable when the object is blurred or blocked. The hand trajectories in blue and yellow and the club trajectories in red and green are achieved by our proposed method in Figures 7–9.

## Conclusion

In this article, a hand and club tracking framework using machine learning based on a descriptor combining HOG and spatial–temporal vector is proposed to improve tracking performance for golf video. After the hand and club are recognized in initial windows positioned by the body region, the boosted classifier trained by the proposed complex descriptor is used for recognition and tracking in a searching window predicted by trajectory fitting with previous four object positions. The boosted classifier has a precision and recall rate both better than 97% and the hand

**Figure 9.** Hand and club tracking of player C. The illustrations are the same as Figure 7.

and club tracking are basically correct in our testing videos, which have been shot in the different outdoors. The tracking results provide visualization of object movements and can be utilized to other desired information.

Since our golf video database is not large enough, the popular deep learning has not been applied in our framework. In the future, as we get more videos, more work can be done to further improve tracking performance with deep learning when videos are shot in the night, in the overcast day, or in other bad situations.

## Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article: This work is supported by the National Natural Science Fund of China (51475047), Scientific Research Project of Beijing Educational Committee (KM201711232003), Key Projects of Science and Technology Program of Beijing Municipal Education Commission (KZ201611232031), and Zepp Labs.

## References

1. Morariu V, Harwood D and Davis LS. Tracking people's hands and feet using mixed network AND/OR search. *IEEE Trans Pattern Anal Mach Intell* 2013; 35(5): 1248–1262.
2. Li Y, Dore A and Orwell J. Evaluating the performance of systems for tracking football players and ball. In: *Advanced video and signal based surveillance*, Como, Italy, 15–16 September 2005, pp. 632–637. IEEE.
3. Kumada K, Usui Y and Kondo K. Golf swing tracking and evaluation using Kinect sensor and particle filter. In: *International symposium on intelligent signal processsing & communications systems*, Okinawa, Japan, 12–15 November 2013, Vol. 112, pp. 698–703, IEEE.
4. Choi A, Joo SB, Oh E, et al. Kinematic evaluation of movement smoothness in golf: relationship between the normalized jerk cost of body joints and the clubhead. *Biomed Eng Online* 2014; 13(1): 1–12.
5. Cao NKN, Kang HJ and Suh YS. Golf swing motion tracking using inertial sensors and a stereo camera. *IEEE Trans Instrum Meas* 2014; 63(4): 943–952.
6. Gehrig N, Lepetit V and Fua P. Visual golf club tracking for enhanced swing analysis. In: Harvey R and Bangham A (eds) *Proceedings of the British machine vision conference*, Norwich, UK, September 2003, Vol. 47, pp. 1–10, BMVA Press.

7. Dollar P, Appel R, Belongie S, et al. Fast feature pyramids for object detection. *IEEE Trans Pattern Anal Mach Intell* 2004; 36: 1532–1545.
8. Lowe D. Object recognition from local scale-invariant features. In: *Proceedings of the international conference on computer vision*, Corfu, Greece, 20–25 September 1999, Vol. 2, pp. 1150–1157, IEEE.
9. Lowe D. Distinctive image features from scale-invariant keypoints. *Int J Comp Vis* 2004; 60(2): 91–110.
10. Bay H, Ess A, Tuytelaars T, et al. SURF: speeded up robust features. *Comp Vis Image Underst* 2008; 110(3): 346–359.
11. Papageorgiou CP, Oren M, and Poggio T. A general framework for object detection. In: *International conference on computer vision*, Bombay, India, 4–7 January 1998, pp. 555–562, IEEE.
12. Viola P and Jones M. Rapid object detection using a boosted cascade of simple features. *Comp Vis Pattern Recognit* 2001; 1: l511–l518.
13. Wang L and He DC. Texture classification using texture spectrum. *Pattern Recognit* 1990; 23(8): 905–910.
14. Dalal N and Triggs B. Histograms of oriented gradients for human detection. In: *IEEE computer society conference on computer vision and pattern recognition*, 20–25 June 2005, Vol. 1, pp. 886–893. IEEE.
15. Liu H, Yu Y, Sun F, et al. Visual–tactile fusion for object recognition. *IEEE Trans Autom Sci Eng* 2017; 14(2): 996–1008.
16. Dobrisek S, Gajsek R, Mihelic F, et al. Towards efficient multi-modal emotion recognition. *Int J Adv Robot Sys* 2013; 10(1): 257–271.
17. Liu H, Liu Y, and Sun F. Robust exemplar extraction using structured sparse coding. *IEEE Trans Neural Netw Learn Syst* 2014; 26(8): 1816–1821.
18. Zhang E, Chen B, Wang X, et al. On the design of a wearable multi-sensor system for recognizing motion modes and sit-to-stand transition. *Int J Adv Robot Sys* 2014; 11(1): 1.
19. Liu H, Sun F, Fang B, et al. Robotic room-level localization using multiple sets of sonar measurements. *IEEE Trans Instrum Meas* 2017; 66(1): 2–13.
20. Lahat D, Adali T and Jutten C. Multimodal data fusion: an overview of methods, challenges and prospects. *Proc IEEE* 2015; 103(9): 1449–1477.
21. Liu H, Guo D, and Sun F. Object recognition using tactile measurements: kernel sparse coding methods. *IEEE Trans Instrum Meas* 2016; 65(3): 1–10.
22. Schapire R and Singer Y. Improved boosting algorithms using confidence-rated predictions. *Mach Learn* 1999; 37(3): 297–336.
23. Webpage: http://opencv.org/ (accessed 17 April 2016)