2019 International Workshop on Smart Info-Media Systems in Asia (SISA 2019), Sep. 4-6, 2019

Body Part Categorization and Occlusion Detection Based Volleyball Players' Spike Height Analysis

Ziken Li*, Xina Cheng*† and Takeshi Ikenaga*
*Graduate School of Information, Production and Systems, Waseda University, Japan
†School of Artificial Intelligence, Xidian University, China

***Abstract*—Volleyball players' spike height is very important in sports analysis of volleyball as it provides a quantitative criterion to judge spike motion. Spike height is the highest point of trajectory of spike hand, which is obtained by 3D tracking. There are two problems in acquiring the position of hand: the deformation of hand and the occlusion of different camera views. This paper proposes body part categorization and occlusion detection based observation model to solve these problems. Body part categorization is a detection based observation model, which tracks the target not by image similarity but by category, so it is robust to deformation. Occlusion detection solves the problem that unknown number of occluded views by detecting occluded views and eliminating their influence. The experiments are based on the videos of final game and semi-final game of 2014 Japan Inter High School Mens Volleyball in Tokyo. The experiment results of proposed method are that: 92.86% of test sequences can be successfully detected the spike height, and in which the average error of spike height is 7.45cm.**

I. Introduction

As the developing of sports analysis, more and more motion data is needed, especially the 3D data. Volleyball players spike height, the highest point of spike hand, is important in 3D information, because it provides a quantitative judgement criterion of spike for sports analysis, and it also improves the entertainment experience for live sports by showing the spike height information during the highlight retrospection. Spike height is obtained by 3D tracking of hands, in which there are problems to solve: deformation and occlusion of hands.

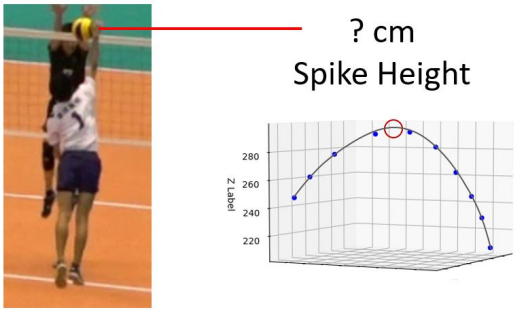


Fig. 1. Spike height frame and hand trajectory

The conventional method of tracking volleyball players body parts focus on the block motion, which is not suitable for spike motion. Xie [1] builds an articulate model of human, which consists of head, shoulders, elbows and hands for upper body, then uses this model to track each body part one by one. To reduce the degrees of freedom of articulate human model, some specially designed restraints for block motion are used. However, such restraints cannot be applied to spike, as spike has more complex movement and much higher moving speed. Without the restraints of human model, it is very hard to handle exponentially increasing computational complex when tracking each part sequentially. So the method that tracks each part one by one is not suitable for our target. Besides, there are no deformation problem when block, also no occlusion problem because block hands are always on the top.

Both deformation and occlusion are difficult to be solved by conventional tracking algorithms. For deformation problem: tracking algorithms such as correlation filter [2] and tracking by histograms of color or texture [3] use the image similarity to the initial frame for finding the current target position, which is not suitable for deformable target, as it could have very low similarity to itself in the initial frame. For occlusion problem: There are conventional works focus on occlusion in mono-camera tracking, such as tracking partly occluded objects [4] and predict the occluded joints of human by relative joints [5]. In multi-view tracking, the situation is different, as some views are occluded while other views are not occluded. So how to detect the occluded views and eliminate the negative influence of them is the problem. There is another conventional work solves the occlusion problem in multi-view tracking of volleyball [6], but this method suppose there is always one occluded view, which is the camera view with the lowest likelihood, and directly delete it. This method does not solve the unknown occluded camera views problem.

To solve these problems, we combine the idea of detection and classifying with tracking, and propose body part categorization for solving deformation problem; propose occlusion detection for solving unknown number of occluded views. Compared to conventional tracking algorithm, body part categorization observation model abandons the image similarity but focus on detection and perception level, which can be concluded as a hand is always a hand even with deformation. Body part categorization observation model will detect the category of each particle, and find the best candidate particle. Occlusion detection aims to reduce the influence of occlusion by classifying camera views into two classes: occluded views and non-occluded views, which is achieved by a specially trained classifier and a non-linear filter. Both of proposals will be introduced in section III.

The rest of this paper is organized as follows. The whole

system is introduced in section II. Details of the proposed method is explained in section III. Finally, the experiments and conclusions are in section IV and section V, respectively.

II. Framework

To obtain the spike height, 3D tracking results of hand are needed. Particle filter is chosen as the key tracking algorithm, because of the complexity and unpredictability of hand motion. As shown in Fig. 2, particle filter consists of 3 key steps [7]: prediction, observation and resampling.

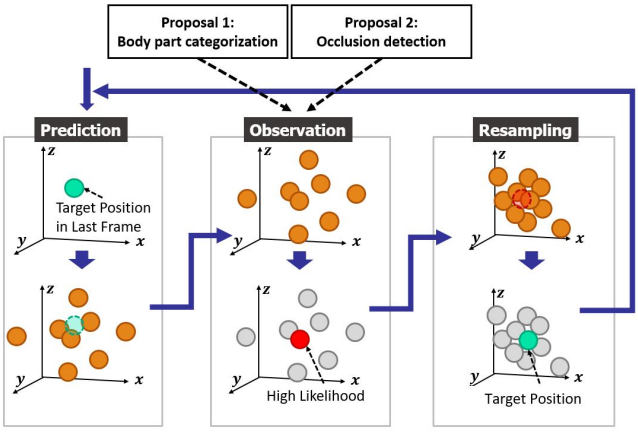


Fig. 2. Framework and proposals

A. Prediction

The state vector S_k in frame k consists of three-dimensional coordinate $X_k = [x_k, y_k, z_k]^T$ and velocity $V_k = [v_{x_k}, v_{y_k}, v_{z_k}]^T$:

$$S_k = [X_k^T, V_k^T]^T \tag{1}$$

Based on target position in frame $k-1$, a set of possible particles in frame k is distributed by following rules: for each particle, we predict the V_k by adding a Gaussian noise ΔV_k to V_{k-1}.

$$V_k = V_{k-1} + \Delta V_k \tag{2}$$

Then the state vector will will be updated:

$$S_k = [X_{k-1}^T + V_k^T, V_k^T]^T \tag{3}$$

B. Observation

The observation space w_k is a set of images of camera m in frame k, and M is the total number of cameras.

$$w_k = [w_k^1, w_k^2, ..., w_k^m, ..., w_k^M] \tag{4}$$

For particle i, its 3D coordinate X_k^i is project to 2D images, the observation space, and likelihood is:

$$L(S_k^i|w_k) = \prod_{j=1}^{M} L(S_k^i|w_k^j) \tag{5}$$

The calculation of $L(S_k^i|w_k^j)$ will be introduced in section III.

C. Resampling

In this step, particles will be redistributed around the high like hood particles by adding Gaussian noise ΔX_k to coordinates:

$$S_k^R = [X_k^T + \Delta X_k^T, V_k^T]^T \tag{6}$$

Then do observation again and pick the best candidate particle as the tracking result.

III. Proposed Method

In observation part of particle filter, how to determine which particle is the best candidate particle is an intractable problem because of deformation and occlusion. To solve these problems, body part categorization and occlusion detection are proposed.

A. Body Part Categorization

Conventional tracking methods focus on the image similarity, the key idea of which can be concluded as find the most similar one. For example, using correlation filter [2] to track, the region which has the most similar color histogram to the initial frame or to the last frame will be regarded as the tracking result. However, during the spike, hand shape is always in fast changing, so actually the image similarity between two frames is low, which is the reason why traditional tracking methods such as using color histogram are hard to track deformable objects.

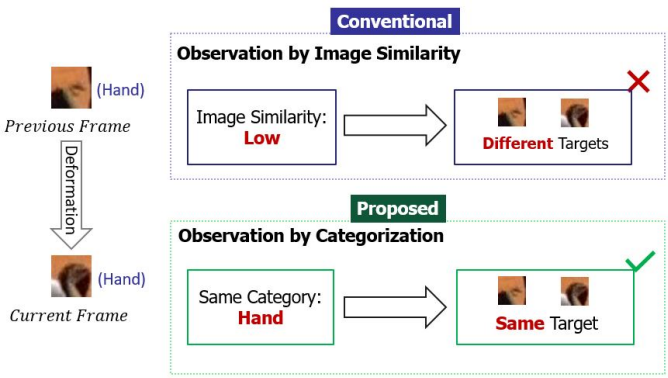


Fig. 3. Conceptual difference of body part categorization

Body part categorization is proposed based on the hypothesis that if two targets belong to the same body part category, and close to each other spatially, they are the same one. Obviously, this hypothesis is another way of expression for the continuity of motion, which is an axiom in real world.

The categorization result is determined by heatmap H_k^j, which is a probability distribution map of each pixel for a certain body part generated by OpenPose [8]. Following is an example of heatmap, the brighter part means high probability

Fig. 4. Heatmap of right hand

to belong to the hand category, while the darker part means low probability.

To improve the performance, categories of all nearby pixels are taken into consideration to by doing convolution with a Gaussian Kernel K. The size of Gaussian Kernel is set to 15 and σ is set to 1. So for particle i in frame k, the body part categorization likelihood $L_B(S_k^i|w_k)$ is:

$$L_B(S_k^i|w_k) = \prod_{j=1}^{M} H_k^{i,j} * K \tag{7}$$

B. Occlusion Detection

To reduce the influence of occlusion, this method aims to detect the occluded views and eliminate the influence of them.

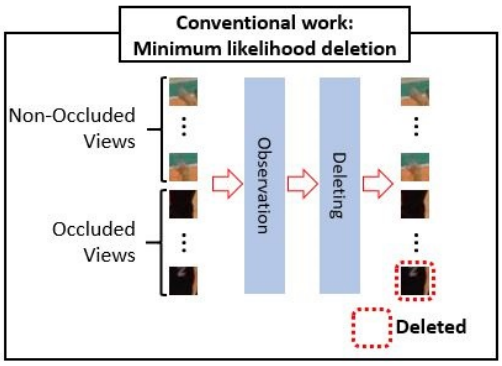
Cheng's method [6] to deal with occlusion is that suppose there is always one occluded view, which is the camera view with the lowest likelihood, and directly delete it. However, this method does not suit for the situation that multiple camera views are occluded, or no occlusion situation. If there are 2 occluded views, but only 1 of them is deleted, the other one will lead to a great negative influence on the tracking result, as occluded views have extremely low likelihoods.

The proposed method has 2 key steps, classifying and filtering.

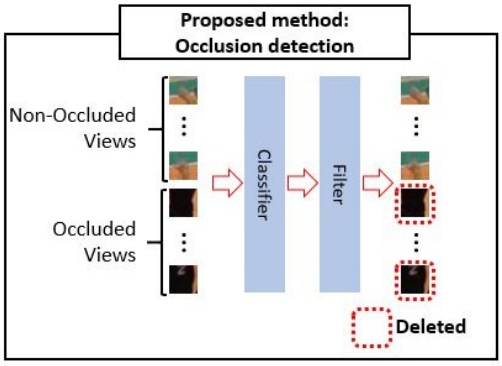
1) Classifying: In this step, as shown in Fig. 6 particles are projected to 2D image of each camera view. Take each projected position as the center of each small image. These small images are the input of a specially trained classifier, which classifies whether the image in camera view j is a hand or not by probability P_k^j.

The key idea of this classifier is that the probability of being a hand indicates the probability of not being occluded. So once we got the probability outputs from this classifier, the occluded views can be determined.

The reason why another neural network is specially trained is that the output of OpenPose is not reliable enough for this task. As shown in Fig. 4, there are some undetected hands, because OpenPose is a general-purpose pose estimation tool which is not suitable for unnormal posture, like arms up. So we choose to train a classifier to classify volleyball players hands specially. Training data include 10,000 labeled images.



(a) Conventional work



(b) Proposed method

Fig. 5. Conceptual difference of occlusion detection

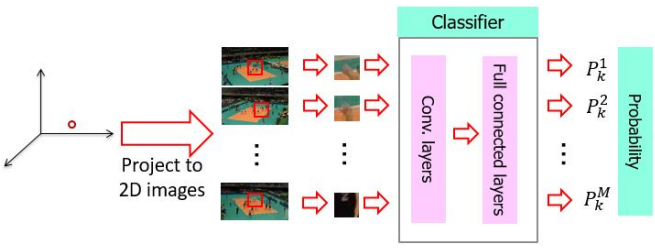
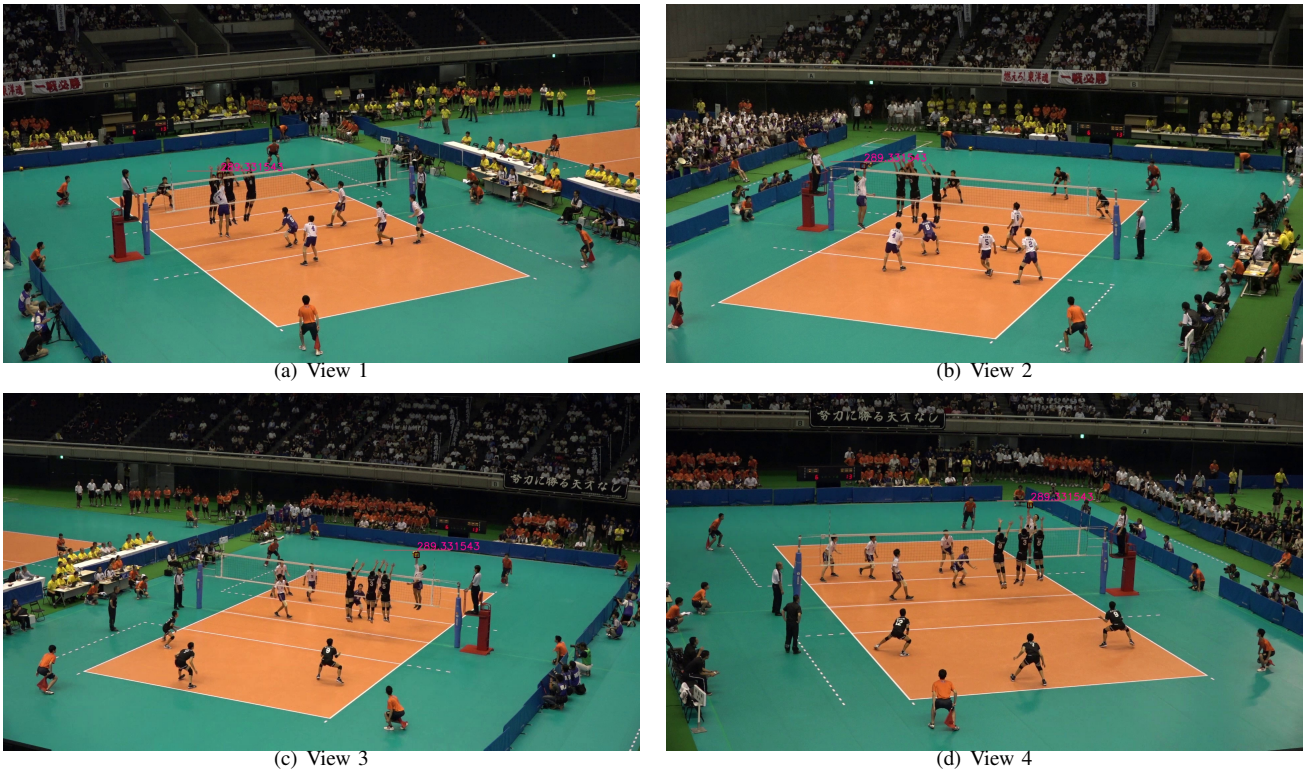


Fig. 6. Flowchart of classifying

2) Filtering: By using a non-linear filter to the probability outputs of the first step, the influence of occlusion is eliminated. This non-linear filter divides the probability outputs to 2 classes, occluded and non-occluded. If the probability of being a hand is larger than the threshold T, the view will be regarded as a non-occluded view, also an encouragement factor ξ will be multiplied to it; If the probability is lower than the threshold T, this view will be regarded as an occluded view, and the output will be set to 1 which will have no influence on the following processes. In proposed method, the T is set as 0.8, and the ξ is set as 100. In frame k, for a given particle i, the likelihood in given camera view j is:

$$L_C(S_k^i|w_k^j) = \begin{cases} \xi P_k^j, & P_k^j \geq T \\ 1, & P_k^j < T \end{cases} \tag{8}$$



(a) View 1 (b) View 2 (c) View 3 (d) View 4

Fig. 7. Examples of experiment results

Then multiply the outputs of the non-linear filter, the classifying likelihood $L(S_k^i|w_k)$ can be obtained.

$$L_C(S_k^i|w_k) = \prod_{j=1}^{M} L_C(S_k^i|w_k^j) \tag{9}$$

Finally, we combine the likelihoods of two proposed methods, the likelihood of particle i can be obtained:

$$L(S_k^i|w_k) = L_A(S_k^i|w_k) \times L_B(S_k^i|w_k) \times L_C(S_k^i|w_k) \tag{10}$$

Where $L_A(S_k^i|w_k)$ is a combination of color likelihoods and edge likelihoods used in previous particle filter based tracking system [6].

IV. Experiment

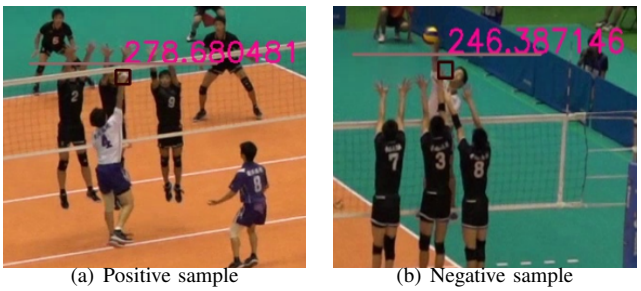
The experiments are run on the videos of final game and semi-final game of 2014 Japan Inter High School Mens Volleyball in Tokyo Metropolitan Gymnasium. The test videos are 4 corner views of volleyball yard. The resolution of videos is 1920 times 1080, and the frame rate is 60 frames per second (fps). The test data set contains 2 volleyball games, in which the total number of spikes is 196, 115 for game 1 and 81 for game 2. For the software environment, our proposed method is implemented on C++ and OpenCV 3.4.1.

The performance is evaluated by success rate and average height error.

Success rate is defined as:

$$SuccessRate = \frac{\#SuccessSequence}{\#TotalSequence} \tag{11}$$

The definition of success sequence is that if the tracking box cover or partly cover the true hand position when it reaches the highest position, it is a success sequence, as shown in Fig. 8.



(a) Positive sample (b) Negative sample

Fig. 8. Definition of success sequence

Average height error is average value of height error of all success sequences, and height error is defined as:

$$h_{err} = |z - z_{ref}| \tag{12}$$

Where z is the detected spike height, the highest position of hand trajectory, and z_{ref} is the reference spike height which is reconstructed by 2D human labelled ground truth of hand position in each camera view. The test result are

(a) View 1 (b) View 2 (c) View 3 (d) View 4

(e) View 1 (f) View 2 (g) View 3 (h) View 4

Fig. 9. Enlarged Examples of experiment results

TABLE I
EVALUATION RESULT

	Success Rate	Average Height Error
Basic Framework: L_a	34.69%	16.39cm
P1: $L_a \times L_b$	75.51%	11.73cm
P1 + P2: $L_a \times L_b \times L_c$	92.86%	7.45cm

shown in Table I, where the basic framework means the combination of color likelihoods and edge likelihoods used in previous particle filter based tracking system, P1 and P2 corresponding to the proposal of body part categorization and occlusion detection respectively, both of which contain the basic framework. Demonstration are shown in Fig.7 and Fig. 9.

The method proposed in this paper still has some problems unsolved. According to this table, there are 7.14% test sequences failed, which means in all 196 test sequences 14 are failed. The main reason of these failed cases is the noise of other hands, such as rival blockers' hands and stuffs' hands. For the future work, a trajectory re-check system will be employed to detect the abnormal trajectory change caused by noise of other hands. Besides, the average height error is 7.45cm, which is still too large to be used in precise sports analysis. The reason is unsmooth tracking results of hands, and the solution could be adding smoothing filters.

V. CONCLUSIONS

We have developed a system to extract spike height of volleyball by 3D tracking of hand, which detect the spike height by finding the highest position of hand trajectory. For solving deformation and occlusion problems in tracking part, body part categorization and occlusion detection are proposed. The success rate of has reached to 92.86%, and the average height error is 7.45cm. For future work, we plan to implement an abnormal trajectory check to increase the success rate, and implement smooth filters to reduce the average height error. And after these modification, this system can be used in volleyball analysis system or improving the entertainment experiment of highlight retrospection in TV content by showing the spike height information.

VI. ACKNOWLEDGEMENT

This work was supported Waseda University Grant for Special Research Projects (2019Q-055).

REFERENCES

[1] F. Xie, X. Cheng, and T. Ikenaga, "Motion State Detection Based Prediction Model for Body Parts Tracking of Volleyball Players," *Pacific Rim Conference on Multimedia,* Springer, Cham, 2017, pp. 280-289.

[2] D. S. Bolme, et al, "Visual object tracking using adaptive correlation filters," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* IEEE, 2010, pp. 2544-2550.

[3] J. Ning, et al, "Robust object tracking using joint color-texture histogram," *International Journal of Pattern Recognition and Artificial Intelligence,* pp. 1245-1263, July 2009.

[4] X. Mei, et al, "Minimum error bounded efficient ℓ 1 tracker with occlusion detection," *CVPR 2011,* IEEE, 2011, pp. 1257-1264.

[5] G. Shu, et al, "Part-based multiple-person tracking with partial occlusion handling," *2012 IEEE Conference on Computer Vision and Pattern Recognition,* IEEE, 2012, pp. 1815-1821.

[6] X. Cheng, et al, "Anti-occlusion observation model and automatic recovery for multi-view ball tracking in sports analysis," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* IEEE, 2016, pp. 1501-1505.

[7] F. Gustafsson, "Particle filter theory and practice with positioning applications," *IEEE Aerospace and Electronic Systems Magazine,* pp. 53-82, July 2010.

[8] Z. Cao, et al, "Realtime multi-person 2d pose estimation using part affinity fields," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2017, pp. 1865-1870.