Vol.2016-CVIM-202 No.5
2016/5/12

Human Body Pose Estimation for Team Sport Videos with Poselets-Regressor and Head Detector

Masaki Hayashi[1,a)] Yoshimitsu Aoki[1,b)]

Abstract: We propose a novel human pose estimation framework for team sports videos captured by a fixed monocular camera. This framework is integrated with the standard tracking-by-detection approach and is able to estimate most of the poses of the athletes, even with the side-view poses that frequently appear in team sport videos. The per-frame pose estimator, i.e., the poselets-regressor, used in the present study estimates the relative joint position (e.g., the pelvis center), which is also the local coordinate center of the appearance of the person used to estimate the target, from the other joint positions (e.g., the head center). After tracking the head position of a subject athlete in a monocular input video, we apply two independent modules per frame. The first module estimates the upper-body pose (upper-body orientation and spine pose), and the second module estimates the lower-body pose (the positions of the four lower-body joints). The proposed relative joint position estimation scheme based on window position alignment and the global features of the appearance of a person provides a simple but robust human pose estimation process, which is similar to the window sharing features of face detection and face recognition. Using the origin-aligned global appearance of a person also leads to the typical failure of previous pose estimation methods with part detectors when the parts are largely occluded. We demonstrate the effectiveness and robustness of the proposed method using soccer and American football videos.

1. Introduction

Data analysis for tactical use in professional team sports has become crucial to winning games because sensor-based or vision-based data acquisition has come into wide use. In order to semi-automatically acquire tracking data in real matches, vision-based multitarget tracking products, such as TRACAB [1] for soccer and SportVU [2] for basketball, are used by professional sports teams. Coaches and staff members can use the acquired trajectories of athletes to analyze the performance or tactics of both the opposing team and their own team. If the poses and trajectories of the athletes can be provided by the system, we can achieve more detailed vision-based data analyses, such as action recognition, skeleton-based pose-type classification, and head or body directional attention analyses.

However, at present, state-of-the-art human pose estimators are not suited to typical team sports videos. Human pose estimation in computer vision has primarily been performed using pose detection via the pictorial structures framework [3], which assumes that whole parts of a person appear in images. While popular methods based on part filters, such as Flexible-Mixtures-of-Parts [4] or Convolutional Neural Networks [5], are able to estimate only *frontal* and *all-parts-shown* poses, such as those in the LEEDs dataset [6] and the FLIC dataset [7]. The previous methods are rarely used with side-view poses (in which the body is facing left or right from the camera), which are not included in [6], [7], or other public human pose estimation datasets. In this sense, previous methods cannot be applied to all frames of the team sports videos in the stadium because most athletes are captured in side-view poses while running toward the goal.

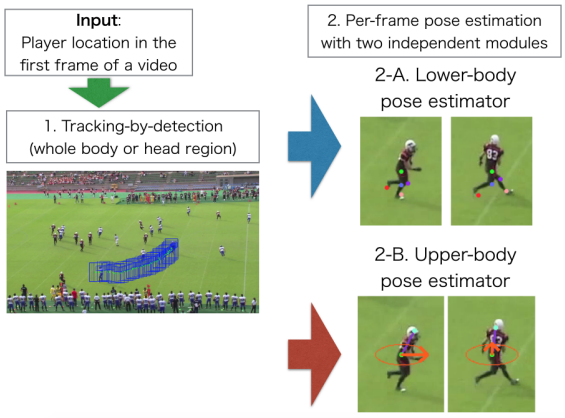
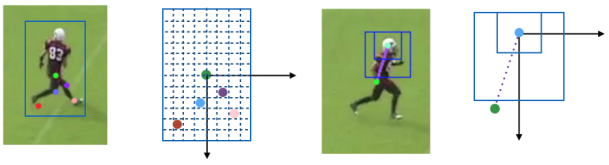


Fig. 1: Overview of the framework. (1) The left-hand side of the figure shows the tracking-by-detection part of the framework, and (2-A and 2-B) the right-hand side of the figure shows the two pose estimators: the lower-body pose estimator [8] and the upper-body pose estimator [9]. Both pose estimators use the center-aligned global appearance of a person in each frame to estimate the corresponding output poses.

1.1 Proposed method

In order to exceed the limited capability of these previous

[1] Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan
[a)] mhayashi@aoki-medialab.org
[b)] aoki@elec.keio.ac.jp



(a) The label-grid classifier (classification) estimates (discrete) the positions of joints (knee and foot) relative to the pelvis center (green circle).

(b) The poselets-regressor (regression) estimates the position of the pelvis center (green circle) relative to the head center (blue circle).

Fig. 2: Proposed method. Per-frame estimation of the positions of joints relative to an aligned center using the aligned global visual features obtained from person tracking or using detection windows (blue rectangles).

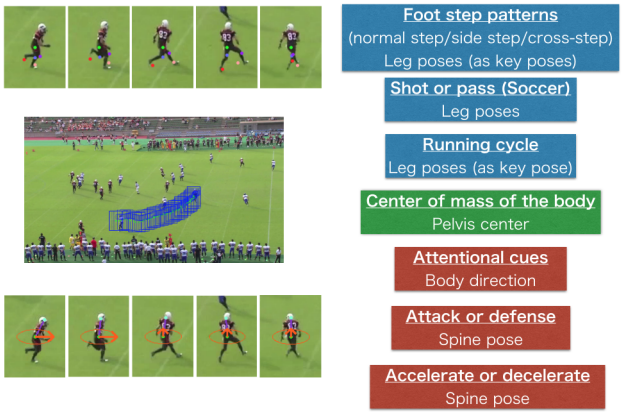


Fig. 3: Meaning of each type of pose in the context of team sports. The left-hand column shows the tracking summary and the five key frames of this sequence, along with the estimated body pose information, which is shown in red. In the right-hand column, the blue rectangles show the pose type (underlined) and the corresponding pose of the lower body. The red rectangles show the types of upper-body poses. The green rectangle shows the most common pose type (pelvis center).

methods, we separately proposed two half-body pose estimators for team sports videos, namely, the lower-body joint position estimator [8] and the upper-body orientation estimator conditioned by the 2D spine line angle (body tilt) [9].

Figure 1 shows the process flow common to these two estimators [8] and [9], in which the person is first tracked in a tracking-by-detection manner and pose estimators are then applied to the aligned appearance of a person in each frame, which are the same features typically used in pedestrian detection. Later, in the first author's dissertation [10], we integrated these two modules. In this technical report, we summarize these three frameworks in Sections 2, 3, and 4.

The proposed method is a simple human joint position estimation method using only the global appearance features (Figure 2). Note that we train the lower-body joint position estimators using the left joint in the image coordinates, as typically used in monocular human joint position estimation, such as in [4] and [5]. For instance, in the proposed method, the left knee is the knee joint located on the left-hand side in the image coordinates and does not necessarily refer to the joint of the left leg.

The lower-body joint estimation [8] used the label-grid classifier (Figure 2(a)). In every frame of the input video, the label-grid classifier estimates the joint position of the lower body using global histogram of oriented gradients (HOG) features [11] within the tracked person window. We trained the person detector using training images for which the window centers were aligned to the pelvis center. The person detector uses the random forests classifier [12] to classify the grid position of the target joint.

The upper-body estimation [9] introduced the poselets-regressor (Figure 2(b)), which can be regarded as the regression version of the label-grid classifier. Instead of using the person tracker, we used the head tracker to estimate the head center of the subject athlete. After tracking the head center, the poselets-regressor estimates the pelvis center relative to the (tracked) head center using upper-body region HOG features for which the local origin is aligned to the position of the head center. In this process, the 2D spine line pose, which consists of the head center and the pelvis center, is obtained. The upper-body orientation is then estimated using the body orientation estimator conditioned by the 2D spine line angle. By adopting the upper-body region alignment determined by the head tracking, the upper-body pose of the athlete of interest can be estimated even when the lower body is occluded by other athletes.

The integrated estimation method in [10] first tracks the head region using the head tracker of Benfold and Reid [13]. The position of the pelvis center relative to the head center is then estimated using the first poselets-regressor. Finally, the positions of the four lower-body joints (left knee, right knee, left foot, and right foot) relative to the pelvis center are then estimated using the first poselets-regressor and the corresponding poselets-regressor. The person window is obtained by tracking-by-detection of the pelvis-aligned person detector in the test stage.

1.2 Definition of Human Pose in Team Sports and Our Contribution

The poses of athletes estimated by the proposed method may be interpreted to have several meanings (Figure 3). These poses have not been estimated in team sports videos because these unconstrained pose patterns are not fully covered in previous human pose estimation studies because typical human pose estimation datasets [6] [7] only includes limited types of poses, as we already discussed. Moreover, by using the output poses from the proposed pose estimation methods, *tactical understanding using poses as input features* can be studied. In other words, our goal is to provide monocular human motion capture for team sports videos in order to develop pose-based analysis of athletes and tactics.

The contributions of the proposed simple pose estimation framework are summarized as follows:

- The proposed pose estimation based on person window alignment provides a very simple and computationally effective pose estimation pipeline by eliminating the need for the structural training procedure required in part-based pose detectors, such as flexible mixtures-of-parts (FMP) models and convolutional neural networks.

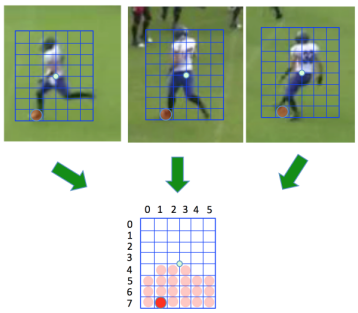


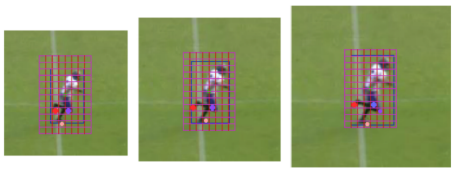
Fig. 4: Label-grid classifier. The red circle on the grid is the classified joint position $\boldsymbol{l}^j \in \mathbb{N}^2$ (red circle) on each athlete window from the trained label-grid class candidates (pink circles). In this example, the label-grid classifier for the j-th joint is on a (6×8) grid structure, and the estimated label-grid is at $\boldsymbol{l}_t^j = (1, 7)$ in all three images. The number of classes for the left foot label-grid classifier is 21 (= sum of pink circles and red circles).

- The alignment strategy also provides a visual feature-space similar to the classical face recognition pipeline [14], which uses the global window appearance given by the face detector. The behavior is easier to understand because we are very familiar with face recognition algorithms.
- Using the proposed method, various partially occluded skeleton poses, when the body direction is either right or left, can be estimated, whereas previous parts-based methods have difficulty dealing with self-occluded poses. This is important because athletes are captured primarily from the side using fixed cameras and the captured images are frequently self-occluded.
- An estimated 2D spine pose would provide novel cues for clarifying sport-specific athlete behavior in the future, although the computer vision community has been focusing only on pedestrian detection and tracking for surveillance applications.
- The proposed head tracker-based upper-body-pose estimation allows stable person pose estimation, even when the lower body of the athletes is occluded.

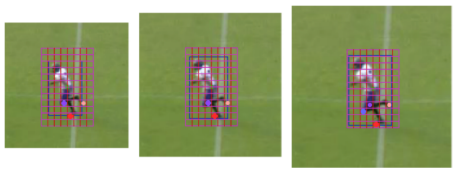
2. Lower-Body Pose Estimation Using the Label-Grid Classifier

The proposed lower-body pose estimation method [8] is the human joint position estimation that is integrated by a popular tracking-by-detection approach to provide the appearance window of a whole person whose center is aligned to the pelvis center. In [8], we proposed a label-grid classifier that estimates the discretized grid position $l^j \in \mathbb{N}^2$ of the j-th lower-body joint from the HOG features within the person window obtained by the tracking-by-detection approach (see Figure 4).

We trained the athlete window detector with training samples for which the window centers are aligned with the pelvis center position. Tracking-by-detection using the trained detector and a Kalman filter provides pelvis-aligned person windows for every frame of the video. Then, we use the label-grid classifier in each frame t to estimate the relative joint position l^j from the pelvis center obtained by tracking-by-detection.



(a) Example images from scaled dataset $\mathcal{D}_{sca}$ with different athlete scales $s = \{0.7, 0.8, 0.9\}$ (note that all athlete windows for the label-grid (purple grid) have the same fixed window size)



(b) Example images from mirrored dataset $\mathcal{D}_{mir}$ created from images and labels $\mathcal{D}_{sca}$

Fig. 5: Data augmentation. Using h_i^{pla} (the height of the blue window), images are scaled to scale s so that the center position $\boldsymbol{p}_i^{pel}$ remains at the center of the label-grid, even in the resized images.

2.1 Data Augmentation for Learning Multiple Appearance Scales

In [8], we independently trained four lower-body joint label-grid classifiers (left knee, right knee, left foot, and right foot) with an American football training dataset collected by ourselves (Figure 2 (a)). Using the trained label-grid classifiers, we performed tracking and pose estimation of one target athlete in each test video. The training multi-scale dataset $\mathcal{D}_{sca}$ was created by applying data augmentation to the original dataset so as to be applicable to multiple human scales $s = \{0.7, 0.75, \ldots, 1.0\}$ (Figure 5). In addition to augmenting resized images of multiple scales, flipping images is also performed in order to create mirrored dataset $\mathcal{D}_{mir}$. In order to perform this data augmentation procedure, each image is labeled with human height h in order to scale images to the target scale s. The training dataset for the j-th joint is $\mathcal{D} = \{(\mathbf{x}, s, h, l^j)_i\}$, where $\mathbf{x} \in \mathbb{R}^N$ is the N dimensional HOG feature vector of the i-th sample, s is the augmented scale of the person, h is the person height, and $l \in \mathbb{N}^2$ is the grid position in the person window.

2.2 Visual Features and Classifier

We used the random forests classifier as a label-grid classifier and used HOG features [11] as the input features for the label-grid classifier. The random forests classifier selects the hierarchical HOG feature space calculated from only the pelvis-center-aligned person windows in all training samples. We normally use a 64×96 person window and a label-grid size of 8×8, which is equal to the HOG cell size. For example, if the dataset $\mathcal{D}$ includes 32 label-grid positions in all training samples, we train 32 class random forests, and each l^j indicates one of the 32 label-grid classes.

2.3 Experiments

In [8], we performed experiments involving the lower-body joint position estimation framework using American football

(a) Test (1) (b) Test (2) (c) Test (3) (d) Test (4) (e) Test (5)

Fig. 6: Example results from frontal pose experiments. The panels on the left-hand side in each subfigure [(a) through (e)] show the results for the label-grid classifiers, and the panels on the right-hand side show the results for the FMP model, where only the four detected joints are shown. (A lack of visualization of joints indicates that the FMP model could not detect anyone in the frame.)

videos captured in a stadium. The test dataset included ten 40-frame videos of a target athlete. The first five tests were frontal poses, and the other five tests were side-view videos in which athletes were primarily running or walking to the left or right with no body tilt. Hence, the primary focus of the experiment was to determine the usefulness of side-view running poses, which occur very frequently in team sports videos.

The average estimation error in pixels for each joint was approximately 10 to 20 pixels in 10 tests. Since the label-grid unit size is 8×8, the errors are within twice the label-grid size. Please refer to [8] for further details on the experimental results of each test.

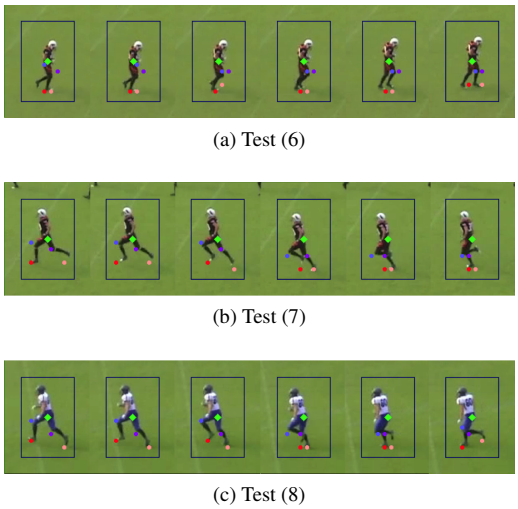
We also compared FMP models [4]. Figure 6 shows the results of FMP models for frontal pose tests. Although FMP models sometimes misdetected the person or fitted the pictorial structures incorrectly when some leg regions were self-occluded, the proposed methods correctly estimated all four joints owing to the globally aligned appearance feature usage. Moreover, in side-view pose tests (Figure 7), the proposed methods can estimate joint positions, whereas the FMP models could not deal with side-view appearances in the presence of severe partial occlusion.

The proposed framework does have limitations. In the experiments of our previous study on lower-body pose estimation [8], no body tilt was assumed because a person detector was used to align the person window and the joint position resolution could not be smaller than the label-grid size. We later addressed these limitations in the integrated method by proposing the use of a head tracker for continuous position estimation using a poselets-regressor (Section 4).

3. Upper-Body Pose Estimator Using a Poselets-Regressor

In our previous upper-body pose estimation study [9], we proposed the following two pose estimation modules:

- We proposed the poselets-regressor, which is a regression version of the label-grid classifier, and used it to estimate the relative pelvis center position from the head center position. Applying the poselets-regressor with the head center tracker provides a 2D spine line approximation of the tracked athlete.
- We proposed conditional regression forests [15] for body orientation estimation, which is conditioned by the 2D spine angle range. Conditioning the body appearance feature space with the 2D spine angle range allows training of the spine-angle-specific body orientation classifier, which only knows the appearance pattern within the spine angle range. This is possible as a result of the newly proposed poselets-regressor of the spine line pose.

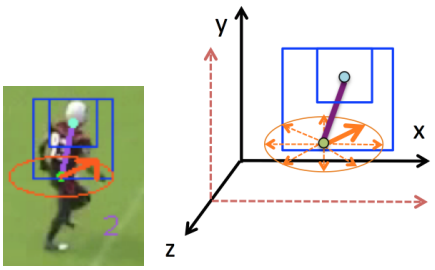


(a) Test (6)

(b) Test (7)

(c) Test (8)

Fig. 7: Example results of side pose experiments.



(a) Example of upper-body pose estimation results in image coordinates (b) Visualized pose estimation results in 3D space

Fig. 8: System output of the proposed upper-body pose estimator laid over the original image and in 3D spatial coordinates. The purple line is the spine line formed by the head center and the pelvis center. The orange arrows show the eight quantized horizontal body directions. The number indicates the s-th spine angle class. Note that we only estimate the 2D spine pose projected onto the image plane in (a) but assume the original 3D pose in (b).

In summary, the proposed upper-body pose estimation method [9] consists of the following three steps:

(1) Tracking the head center of a target athlete using the head tracking method of [13] (Section 3.1).
(2) 2D spine pose estimation using the poselets-regressor (Section 3.2).
(3) Body orientation estimation using conditional regression forests using the spine angle estimated in the previous step (Section 3.3).

The proposed method estimates the spine pose and the body orientation of the head-tracked target athlete (see Figure 8). We

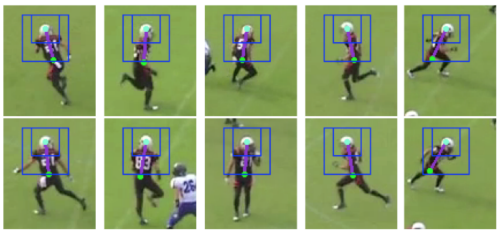


Fig. 9: Example results for estimating the relative position of the pelvis using a poselets-regressor.

summarize the above steps in the following subsections.

3.1 Head Tracking

We use the multitarget head tracking method of Benfold and Reid [13] to track the head position $\mathbf{h}_t \in \mathbb{N}^2$ of the target athlete in each frame t of a test video. The approach of [13] uses HOG features and a support vector machine (SVM) classifier for the likelihood of the Kalman filter and uses optical flow keypoint tracking for the motion prediction of the Kalman filter. Whereas the original approach [13] uses a multitarget state space for the Kalman filter, we use the state space of only one target head because our objective is to track only one target athlete to estimate his or her pose. Another difference is that we train sport-specific HOG-SVM head detectors (e.g., the head detector of American football athletes and the head detector of soccer athletes), whereas [13] trained and used a more generic head detector.

3.2 2D Spine Pose Estimation Using a Poselets-Regressor

Given the head center position $\mathbf{h}_t$ at frame t, we estimate the pelvis center position $\mathbf{p}_t \in \mathbb{N}^2$ at frame t using the poselets-regressor, which was newly proposed in [9]. Figure 9 shows some estimation results obtained using our spine poselets-regressor, where the purple line indicates the 2D spine line formed by $\mathbf{h}_t$ at frame t and $\mathbf{p}_t$. The smaller blue rectangle is the head region given by the head tracker, and the larger blue rectangle is the upper-body region used to calculate the HOG features for the poselets-regressor to estimate the relative pelvis center from $\mathbf{h}_t$ at frame t.

The proposed poselets-regressor is the regression version of the label-grid classifier [8] by simply replacing classification forests with regression forests. The poselets-regressor estimates the relative joint position $\mathbf{j}^t \in \mathbb{N}^2$ of a target joint t from another joint position $\mathbf{j}^o \in \mathbb{N}^2$ using the person HOG features within the HOG window for which the local origin is aligned with another joint $\mathbf{j}^o$. In our previous upper-body pose estimation study [9], we trained a poselets-regressor that estimates the relative position of the pelvis center $\mathbf{p}_t$ based on the position of the head center $\mathbf{h}_t$ given by the head tracker using global upper-body HOG features as an input vector for the regression forests (see Figure 8 for the upper-body HOG region). Data augmentation is also performed by a label-grid classifier to train the models that know multiple scales of athlete poses.

We referred to this label-grid classifier as the poselets-regressor in [9] because this classifier can be regarded as regressing the original poselets. The poselets-regressor is a detector of one specific pose in which the joints are aligned. The poselets-regressor uses the known local origin of the joint position $\mathbf{j}^o$ and can regress the other target joint position $\mathbf{j}^t$.

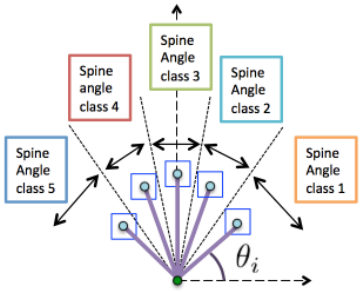


Fig. 10: Spine angle classes. The blue and green circles indicate the head center $\mathbf{h}_t$ and pelvis center $\mathbf{p}_t$, respectively, of the subject athlete. The spine angle range of the training dataset is divided into five spine angle classes.

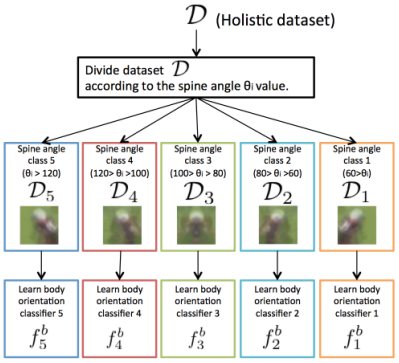


Fig. 11: Learning multiple body orientation classifiers by grouping datasets into subsets having the same spine angle range. The images are average images for each D_s.

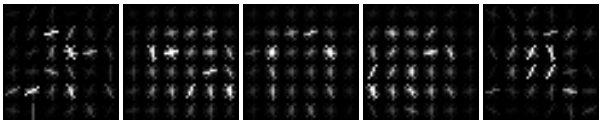
3.3 Estimating Body Orientation Using Conditional Random Forests Classifiers

Inspired by conditional pose estimation approaches [15], [16], we proposed the conditional classification forest for estimating the discretized horizontal body orientation $\mathbf{o}_t^b \in \{0, 1, \ldots, 7\}$ using the 2D spine angle range as the conditional prior.

Given the 2D spine pose in Section 3.2, we first calculate the 2D spine angle θ_t at frame t and discretize θ_t into the spine angle class s using the following condition:

$$s = \begin{cases} 1 & (60 > \theta_t) \\ 2 & (80 \geq \theta_t > 60) \\ 3 & (100 \geq \theta_t > 80) \\ 4 & (120 \geq \theta_t > 100) \\ 5 & (\theta_t > 120) \end{cases}$$

Figure 10 also shows the visual illustration of spine angle class s. Using the spine angle class s, we estimate the body orientation $\mathbf{o}_t^b$ using a corresponding classifier f_s^b (classification forests) trained with only the samples within the spine angle range in Equation 3.3 (see Figure 11 for the training procedure). For the



(a) $s = 5$ (b) $s = 4$ (c) $s = 3$ (d) $s = 2$ (e) $s = 1$

Fig. 12: Visualization of the importance of HOG features of each body orientation classifier for spine angle class s trained from the American football dataset. Whiter orientations indicate greater importance in trained random forests. See Figure 11 for the average image of the training samples for each spine angle class.

input features of each f_s^b, we also used the same features as the spine poselets-regressor described in Section 3.2, which are the HOG features within the head-center-aligned upper-body region.

3.4 Comparison with Related Methods and the Advantage of the Proposed Method

The advantage of the proposed conditional upper-body orientation estimation strategy is that random forests can only focus on a sample in which the spine angle is roughly aligned with the corresponding spine class s to extract features by which to classify the body orientation class with only a compact random forest tree structure. Since upper-body angles are roughly aligned in each spine angle class and edges around a body boundary appear in the same position, selecting important HOG features or disregarding non-discriminative features with a smaller feature distribution is easier using random forests. Figure 12 shows the HOG features selected through random forests training for each spine angle class. Without spine angle priors, random forests must cluster the feature space by itself and cannot guarantee good feature selection from the various pose appearances in the training dataset.

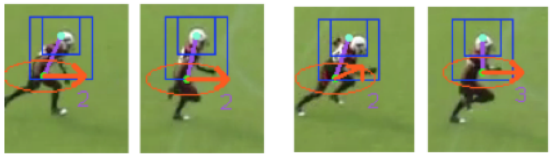
Previous body orientation estimation approaches applied to surveillance videos, such as [17], [18], typically used pedestrian detector output rectangles for the feature region of body orientation estimators. On the other hand, the proposed method uses head-center-aligned upper-body regions for body orientation estimation. For standing pedestrians only, good region alignment is easy to obtain using a pedestrian detector. However, since athletes tend to have various spine angles (body tilt angles), person detectors cannot always provide aligned person appearances. This is the main reason why we used head tracking results as the alignment center for both the body orientation estimators and the poselets-regressor of the relative pelvis position. Compared with the other alignment keypoint of the person appearance, the head center of an athlete, which often becomes stable for team sports videos (in which head appearances tend to be similar), is easier to estimate using only a head tracker.

3.5 Experiments

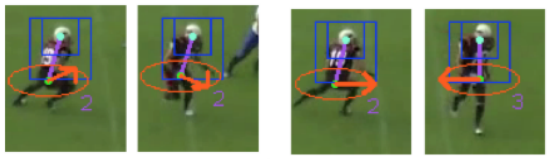
We performed experiments on both spine pose estimation and body orientation estimation. We used American football and women's soccer videos in the experiments and trained a sport-specific spine poselets-regressor and a conditional regression forest for estimating body orientation. Test sequences consist of 12 American football videos and 10 soccer videos. Each video contains 80 frames of the target athlete. The athletes in the videos sometimes bend their upper body and change their upper-body direction.

Table 1: Average estimation error (in degrees) of the body orientation in each scene dataset.

Dataset	Proposed method	[19]
American football scenes	20.90	23.57
Women's soccer scenes	39.99	47.02



(a) Correct samples from test 6 (b) Incorrect samples from test 6



(c) Correct samples from test 12 (d) Incorrect samples from test 12

Fig. 13: Sample results of bending poses obtained through American football tests. Head center misalignment tends to result in incorrect body orientation.

3.5.1 Spine Pose Evaluation

For both FMP models and the proposed method (head tracking and the poselets-regressor of the pelvis center position), the average head center position error and the pelvis center errors were small and were approximately the same. However, FMP models tend to misfit some of body parts while the spine pose itself is correct, and we cannot determine whether the spine pose is correct when this misfitting occurs . See Figure 14 for the FMP detection results in our test videos.

3.5.2 Body Orientation Evaluation

We applied both the proposed method and a commonly used previous procedure using only one body orientation classifier for the whole-spine angle classes in our previous method [19], which is the same type of approach used in typical classical body orientation estimation [17]. Note that we trained a sport-specific classifier. That is, we trained body orientation classifiers with only American football data and tested the classifiers with only the American football test sequences.

Table 1 shows the average body orientation errors (in degrees) for angles converted from the eight body-direction classes. Figure 15 shows the body orientation estimation results as confusion matrices. While the average body orientation errors of the proposed method are slightly better those of the previous method [19], the confusion matrices of the propose method have smaller errors because the diagonal grids are thicker, as shown in Figures 15(b) and 15(d). This indicates that conditional classifier separation results in better fitting and focuses on smaller variations of patterns in each spine angle class, as discussed in Section 3.4. Note that these results are obtained from head-center-aligned HOG features, whereas previous body orientation papers use per-

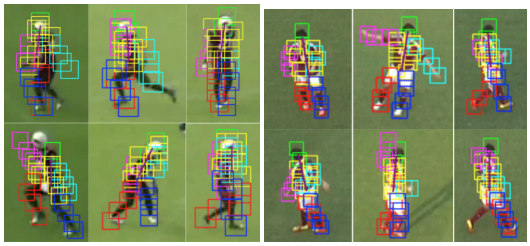
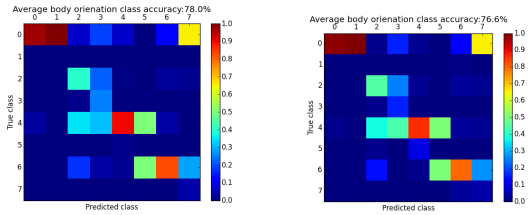
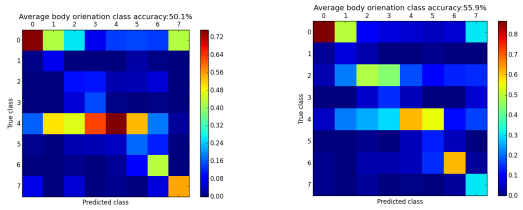


Fig. 14: Example results of skeletal pose estimation using the FMP model [4]. The purple line indicates the spine line formed by the head center and the pelvis center.



(a) Body orientation classifier of [19] for American football scenes

(b) Proposed method [9] applied to American football scenes



(c) Body orientation classifier of [19] for soccer scenes

(d) Proposed method [9] applied to soccer scenes

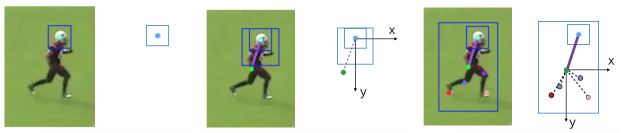
Fig. 15: Confusion matrices of body orientation estimation results.

son detectors designed for detecting pedestrians and ignore the person appearance when body tilt occurs. Hence, the results also indicate that body orientation estimation can be performed using head-center-aligned person (upper-body) HOG features.

Although promising results were obtained through the experiments, some frames had errors that are thought to originate from the alignment-based algorithm. Figure 13 shows the typical errors that occur when the alignment of the upper-body region is insufficient. While HOG features are pooled and quantized as the resolution of cell grid size, the body orientation estimator tends to fail when the error of the tracked head position is larger than the cell size. Thus, the proposed framework may be too dependent on the head-center alignment as compared with part-based pose estimators, which can use many anchor points via multiple part detections.

4. Whole-Body Pose Estimator with Two Stages of Poselets-regressor

In the first author's doctoral thesis [10], we integrated our lower-body pose estimator [8] and our upper-body pose estimator [9] into a single framework by simply using the pelvis center as the connecting point between two estimators (see Figure 16). Namely, we first apply a head tracker and a spine pose estimator in the manner described in [9] to estimate the pelvis center position in each frame. Then, we use four lower-body joint poselets-regressors by using the pelvis-center-aligned HOG features as input features, as we did in the label-grid classifier [8]. In other words, we generalize the poselets-regressor to the relative landmark position regression between any pair of landmarks by applying lower-body landmarks regression.



(a) Step 1: The head region (blue rectangle) in each frame is tracked by the head tracker.

(b) Step 2: The 2D spine pose is estimated by the poselets-regressor. The larger rectangle indicates the HOG features region.

(c) Step 3: The lower-body joints are estimated by poselets-regressors for each joint. The larger rectangle indicates the HOG features region.

Fig. 16: Pose estimation procedure of the integrated method.

4.1 Proposed Method

The proposed method consists of the following steps:

(1) Tracking the head center of a target athlete using the head tracking method of [13] (Section 3.1).

(2) 2D spine pose estimation using a poselets-regressor (Section 3.2) to estimate the pelvis center position.

(3) Estimation of four lower-body landmarks using four poselets-regressors.

The first two steps are as described in our previous upper-body pose estimation paper [9], and the third step uses the same alignment landmark and features as our previous lower-body pose estimation paper [8]. We use two stages of poselets-regressors by regarding the pelvis center as the connection point between two stages.

4.2 Experiments

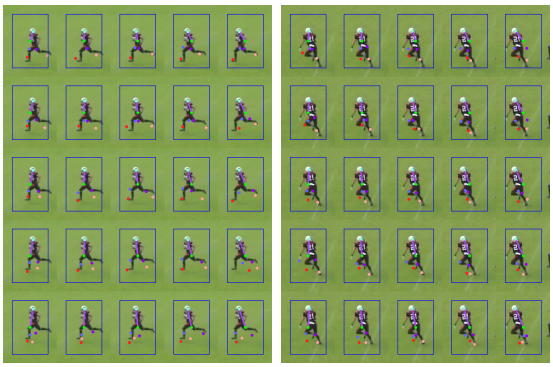
The proposed integrated framework is similar to our previous approaches. In order to examine our previous frameworks [8] and [9] in greater detail, in our thesis, we performed experiments on lower-body joint position estimation accuracy using the following settings, which were not considered in our previous studies [8] and [9]:

- Different cell sizes of HOG features (8×8 or 4×4).
- Comparison of whole-body region HOG features (as in [8]) and lower-body region HOG features.
- Comparison joint labeling approaches between the approach (1) based on the left or right joint on image coordinate (which is usually used in human pose estimation studies) and (2) that based on the joint of left leg or joint of right leg.

Table 2 shows the estimation errors for combinations of settings. Based on the results shown in the table, changing both the body region size and the HOG cell size does not appear to affect the estimation accuracy for all four joints in American football appearance patterns. Basing the labeling policy on the left leg or right leg rather than the left joint or right joint on image coordinate deteriorates the accuracy of the estimation.

Table 2: Average estimation error for each joint in American football tests (1) through (10) for four settings. All errors are in pixels. The columns list the results with the combination of prepared conditions. The results are obtained using the poselets-regressor to estimate the position of the specified joint.

Input region	Whole body	Whole body	Lower body	Whole body
Label policy	Image (left/right)	Image (left/right)	Image (left/right)	Leg (left/right)
Cell size	8 ×8	4 ×4	8 ×8	8 ×8
Left knee	6.73	8.35	7.85	11.57
Right knee	9.07	10.18	9.98	18.04
Left foot	6.92	7.50	7.96	8.88
Right foot	5.56	6.72	6.69	10.44
Head	0.60			
Pelvis	7.06			



(a) Example results for Test (1) (b) Example results for Test (3)

Fig. 17: Results of window-shifted tests. The columns indicate the movement along the x-axis $(-8, -4, 0, 4, 8)$, and the rows indicate movement along the y-axis $(-8, -4, 0, 4, 8)$. The central figure shows the movement of point (0,0) with the ground-truth position of the pelvis center.

4.3 Shifting the Input Window for the Lower-Body Joint Poselets-Regressors

In another experiment, we deliberately shift the pelvis center from the ground-truth position in order to provide the shifted HOG features for the lower-body joint poselets-regressor. We determined how the estimation errors change when the appearance window is deviated along the x-axis with $(-8, -4, 0, 4, 8)$ moves and along the y-axis with $(-8, -4, 0, 4, 8)$ moves.

As a result, we confirmed that the joint estimation error increases as the pelvis center error increases. Figure 17 shows example test results. As indicated by the result, the more the pelvis center moves, the greater the joint estimation error becomes. Additional details are provided in Chapter 5 of the thesis [10].

5. Conclusion

We proposed a human pose estimation approach, which we refer to as the poselets-regressor, based on the window alignment and relative joint position estimation method. To the best of our knowledge, the proposed approach for team sports videos is the first method that can deal with *all types of poses*, including poses that have been previously ignored in human pose estimation studies, such as side-view poses, poses with body tilt, and poses with severe partial occlusions. Although the proposed method is too dependent on the alignment of the origin landmark for the poselets-regressor, it can approximately estimate the correct joint positions as long as the alignment by the tracker or the first-stage poselets-regressor is adequate.

We would like to increase the robustness of the proposed method by combining typical part-based human pose detection approaches. Moreover, we would like to develop pose-based athlete behavior recognition using the spine pose, body orientation, and leg poses.

Acknowledgments The present research was conducted as joint research with Panasonic Corporation and AVC Networks Company. We would like to thank our fellow researchers and staff at these companies, as well as our colleagues at Keio University. American football videos were provided by Panasonic IMPULSE, and women's soccer videos were provided by the Keio University soccer team.

References

[1] TRACAB: `http://chyronhego.com/sports-data/tracab`.
[2] SportVU Player Tracking: `http://www.stats.com/sportvu/sportvu-basketball-media/`.
[3] Felzenszwalb, P. F. and Huttenlocher, D. P.: Pictorial structures for object recognition, *International Journal of Computer Vision*, Vol. 61, No. 1, pp. 55–79 (2005).
[4] Yang, Y. and Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts, *CVPR* (2011).
[5] Toshev, A. and Szegedy, C.: Deeppose: Human pose estimation via deep neural networks, *CVPR* (2014).
[6] Johnson, S. and Everingham, M.: Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation., *BMVC*, Vol. 2, No. 4, p. 5 (2010).
[7] Sapp, B. and Taskar, B.: MODEC: Multimodal Decomposable Models for Human Pose Estimation, *CVPR* (2013).
[8] Hayashi, M., Oshima, K., Tanabiki, M. and Aoki, Y.: Lower Body Pose Estimation in Team Sports Videos Using Label-Grid Classifier Integrated with Tracking-by-Detection, *IPSJ Transactions on Computer Vision and Applications*, Vol. 7, No. 1, pp. 18–30 (2015).
[9] Hayashi, M., Oshima, K., Tanabiki, M. and Aoki, Y.: Upper Body Pose Estimation for Team Sports Videos Using a Poselet-Regressor of Spine Pose and Body Orientation Classifiers Conditioned by the Spine Angle Prior, *IPSJ Transactions on Computer Vision and Applications*, Vol. 7, No. 1, pp. 121–137 (2015).
[10] Hayashi, M.: Human Body Pose Estimation Framework for Team Sports Videos Integrated with Tracking-by-Detection, PhD Thesis, Keio University (2015).
[11] Dalal, N. and Triggs, B.: Histograms of Oriented Gradients for Human Detection, *CVPR* (2005).
[12] Criminisi, A., Shotton, J. and Konukoglu, E.: Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning, *Foundations and Trends® in Computer Graphics and Vision*, Vol. 7, No. 2–3, pp. 81–227 (2012).
[13] Benfold, B. and Reid, I.: Guiding Visual Surveillance by Tracking Human Attention., *BMVC* (2009).
[14] Jain, A. K. and Li, S. Z.: *Handbook of face recognition*, Springer (2005).
[15] Dantone, M., Gall, J., Fanelli, G. and Van Gool, L.: Real-time facial feature detection using conditional regression forests, *CVPR* (2012).
[16] Sun, M., Kohli, P. and Shotton, J.: Conditional regression forests for human pose estimation, *CVPR* (2012).
[17] Andriluka, M., Roth, S. and Schiele, B.: Monocular 3d pose estimation and tracking by detection, *CVPR* (2010).
[18] Chen, C. and Odobez, J.: We are not contortionists: coupled adaptive learning for head and body orientation estimation in surveillance video, *CVPR* (2012).
[19] Hayashi, M., Yamamoto, T., Ohshima, K., Tanabiki, M. and Aoki, Y.: Head and Upper Body Pose Estimation in Team Sport Videos, *International Joint Workshop on Advanced Sensing/Visual Attention and Interaction (ASVAI2013), on ACPR 2013* (2013).