IPSJ Transactions on Computer Vision and Applications Vol.7 121–137 (Oct. 2015)

[DOI: 10.2197/ipsjtcva.7.121]

Research Paper

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

Masaki Hayashi[1,a) Kyoko Oshima[2,b) Masamoto Tanabiki[3,c) Yoshimitsu Aoki[1,d)

Received: August 12, 2014, Accepted: July 13, 2015, Released: October 20, 2015

Abstract: We propose a per-frame upper body pose estimation method for sports players captured in low-resolution team sports videos. Using the head-center-aligned upper body region appearance in each frame from the head tracker, our framework estimates (1) 2D spine pose, composed of the head center and the pelvis center locations, and (2) the orientation of the upper body in each frame. Our framework is composed of three steps. In the first step, the head region of the subject player is tracked with a standard tracking-by-detection technique for upper body appearance alignment. In the second step, the relative pelvis center location from the head center is estimated by our newly proposed poselet-regressor in each frame to obtain spine angle priors. In the last step, the body orientation is estimated by the upper body orientation classifier selected by the spine angle range. Owing to the alignment of the body appearance and the usage of multiple body orientation classifiers conditioned by the spine angle prior, our method can robustly estimate the body orientation of a player with a large variation of visual appearances during a game, even during side-poses or self-occluded poses. We tested the performance of our method in both American football and soccer videos.

Keywords: human pose estimation, body orientation estimation, poselets, tracking-by-detection, feature selection

1. Introduction

In team sport video analysis, tracking players with computer vision-based methods is widely studied because the trajectories of players are the most basic and important kinds of information. There have been many applications that track players using a monocular view [1] or multiple views [2]. However, these player-tracking methods can only achieve location-based activity recognition of players, for example [3], [4]. Conversely, there are a few studies on group activity recognition for sports videos using per-player actions as features [5], [6]. While these methods can infer the semantic actions of each player (e.g., "running", "jumping"), they cannot recognize the fine-grained activity differences between activity classes because they just perform (discretized) classification between those semantic actions.

If sport-specific *upper body pose* patterns can be estimated from a vision-based recognition technique, there will be a new opportunity to realize the more detailed *pose-based* activity recognition of team sports players. The extraction of the upper body pose will achieve a deeper understanding of player actions by recognizing changes in the upper body pose, such as the gradual spine angle change during running from the starting position, defensive bending poses, and blocking poses. Moreover, gaze and direction-based attention prediction [7] and group activity recognition from orientations and relative locations of people [5], [6], [8] can also be achieved even for the team sports videos in the future. However, estimating the upper body orientation or body tilt from team sports videos has rarely been explored.

Realistic upper body pose appearances of team sports players have a wider variety of articulated pose patterns (**Fig. 2**) than pedestrian cases [9], [10], [11] with the following variations:

- Many types of spine angle (body tilt) (Fig. 2 (a)).
- Running while looking backward. (Fig. 2 (b))
- While moving back (Fig. 2 (c)).

These postural patterns, unseen in surveillance videos for pedestrian tracking, make it more difficult to realize the upper body pose estimation method for team sports videos. Specifically, sports players in team sports videos have more *body tilt variations* than pedestrians poses because they tend to bend their (upper) body while in defensive actions or some specific actions (e.g., passing action in soccer). Larger body tilt variations make the body orientation problem more difficult because the previous body orientation estimators during pedestrian tracking or detection [11], [12] depend on the alignment of the input window by the pedestrian detector for only standing walking poses.

To cope with those postural variations in team sports, we propose a framework for estimating the upper body orientation of a player with the *head-center-aligned* upper body region using the selected classifier conditioned by spine angle (**Fig. 1**). Our framework, which depends on the alignment of the upper body appearance, not only estimates the body orientation of the track-

[1] Keio University, Yokohama, Kanagawa 223–8522, Japan
[2] Panasonic Solution Technology Corporation, Minato, Tokyo 105–0013, Japan
[3] Panasonic Corporation, Yokohama, Kanagawa 224–8539, Japan
[a) mhayashi@aoki-medialab.org
[b) ohshima.kyoko@jp.panasonic.com
[c) tanabiki.masamoto@jp.panasonic.com
[d) aoki@elec.keio.ac.jp

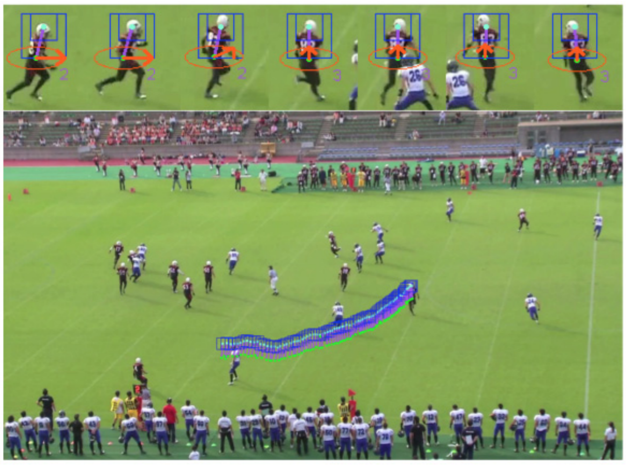
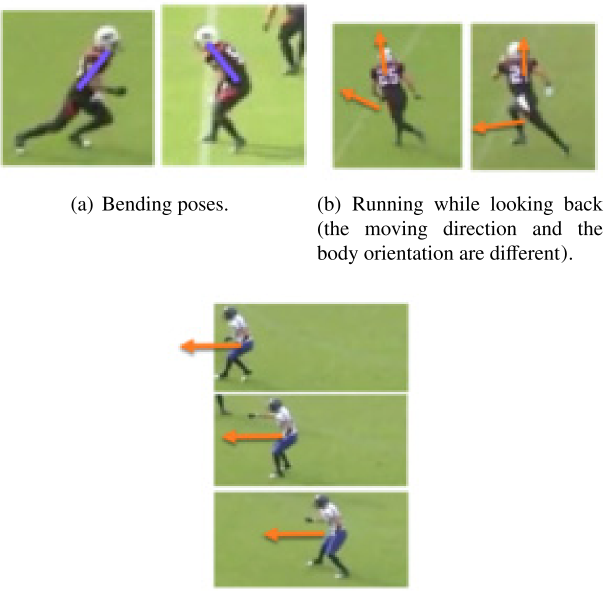


Fig. 1 Example result from our framework. Images in the upper row show the tracked player images in each frame of the input video with the estimated horizontal body orientation (orange arrow), and 2D spine pose (2D purple line). The lower image is a summary image from a test video, which shows the location of the head region and the pelvis center of the subject player in each frame. Our method can track and estimate the upper body pose even when lower body occlusion occurs because it only utilizes the upper body region appearance of the tracked player.

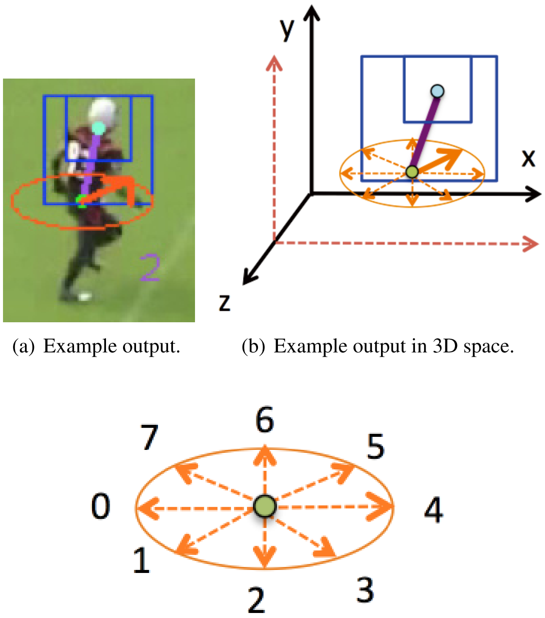


(a) Bending poses.

(b) Running while looking back (the moving direction and the body orientation are different).

(c) Moving back.

Fig. 2 Variation of human poses in team sports videos. These poses rarely appear in pedestrian surveillance videos and produce visual patterns for classifying body orientation.



(a) Example output.

(b) Example output in 3D space.

(c) Body orientation classes.

Fig. 3 System output in image and 3D space. (a) Example results shown in image. The smaller blue rectangle is the head region tracked by the head tracker. The larger blue rectangle is the upper body region for the poselet-regressor and the body orientation classifiers. (b) Visualized result in 3D space. The orange arrows show the eight horizontal orientation classes. The purple line is the spine, which includes the head center position (cyan) and the pelvis center position (green). The number showing at the right-bottom corner of the upper body means the selected spine-class s. (c) Class numbers for each body orientation class.

ing player (orange arrow in **Fig. 3** (a)) as in previous work on surveillance [11], [12], but also estimates the *2D spine pose* of the players (purple line in Fig. 3 (a)), which consists of the head center location and the pelvis center location *even during the bending poses*.

Our framework is composed of three steps. In the first step, the head position of the moving player is tracked by the head tracker of Ref. [13]. In the second step, we estimate the relative pelvis center location from the head center position using our proposed *poselet-regressor* with *head-center-aligned* Histogram-of-Oriented Gradients (HOG) [14] features within the upper body window. This results in estimating the 2D spine pose in each frame. Our poselet-regressor has continuous output space, while the original poselets [15] are trained as a pose exemplar detector. In the third step, we use the estimated 2D spine angle from the second step as a conditional prior for selecting a corresponding upper body orientation classifier, and then the upper body orientation is estimated by the head-center aligned (or pelvis-center aligned) upper body region HOG features within a corresponding spine angle range, which we call a spine class.

This framework is the extension of our previous body orientation and spine pose estimation work [16] with two major contributions: (1) relative spine pose estimation with the head tracker and the poselet-regressor with head-centered-aligned upper body appearance; and (2) classifier selection scheme using spine-angle prior and aligned appearance window, which was inspired by Refs. [17], [18], but with the difference that our conditional prior is the 2D spine angle class. Note that this paper only focuses on the body orientation estimation problem while the previous version [16] also proposed the head orientation estimator.

The first contribution of this paper is the proposed poselet-regressor (in step 2) that estimates the 2D spine pose *without* using pictorial-structures-based pose detectors such as Refs. [19], [20], [21] and the poselet detectors [15], [22], which has been a popular strategy for human pose estimation in computer vision. Our poselet-regressor predicts the relative pelvis position from the head center using the head-center-aligned upper body appearance determined from regression forests [23]. Compared with the multi-view pose estimation method using poselets [24], our framework tries to estimate the 2D spine pose (the line between the head center and the pelvis center) using the poselet-regressor and the head-center-aligned window from the head tracker. In other words, our poselet-regressor is a relative joint location predictor that depends on the local origin (head center) estimated by the head tracker in each frame of the video.

The second contribution of this paper is the switching scheme between multiple upper body orientation classifiers (in step 3) using the spine angle value as a conditional prior. This enables each body orientation classifier (random decision forests) to focus on selecting the important (HOG) features from the conditioned subset of the whole training dataset that includes similar and spine-pose-aligned upper body appearances with the same spine angle range. Previous body orientation estimation approaches for surveillance videos [10], [11] do not deal with bending poses but only pedestrians walking or standing upwright. Our conditioned prior scheme is inspired by the conditional regression forests [17], [18], but our conditional prior is the spine angle, which has never been explored before as a prior.

In addition to these major two contributions, another key contribution of this framework is that our method estimates the upper body pose of the player even when partially occluded, because it depends on the head tracker and the upper body orientation classification using *only* the selected features within the upper body region during training time with Regression Forests [23]. Since our previous work [16] depends on tracking the whole body, it can only track and estimate the pose of isolated players. Our new design, which tracks the head and uses the head-center-aligned upper body appearance, which is inspired by our another lower body pose estimation framework [25], opens up more chances to track and estimate the pose of players even in congested situations in team sports by only tracking the head region and using only the tracked upper body region for spine pose and body orientation estimation.

Since the variation of the arm pose is very large in unconstrained team sport player appearances, it is also important to depend aligned global appearance of the person. Because parts exemplars such as poselets [15], [22] and pictorial structures [20], [21] have to deal with the all of the arm parts for pose prediction, they are not always scalable to the unconstrained huge patterns of articulated poses. Instead of training parts detectors, we introduce the alignment of the whole body appearances from the tracker, which is the standard approach of body orientation classification methods during tracking-by-detection [11], [12], [16], and utilizes only discriminative HOG features within the upper body region selected by random decision forests training. Moreover, our poselet-regressor also uses the same aligned-global appearance approach to predict the relative pelvis center location from the (tracked) head center.

In summary, our alignment via head tracking and the feature selection with conditional spine pose prior are aimed at dealing with pose appearance patterns of sports players rather than pedestrians (for estimating their body orientation). By acquiring head and pelvis-center aligned upper body images, each body orientation classifier needs to deal with only the smaller appearance distribution within the corresponding spine angle class. To achieve this, even for an unconstrained setting, we propose the regression forests-based skeletal spine pose estimation.

The rest of the paper is organized as follows: Section 2 investigates related work. Section 3 introduces the overview of our framework. Section 4 introduces the spin pose estimation procedure using the head tracker and the poselet-regressor. Section 5 introduces the multiple body orientation classifiers conditioned by spine angle prior. Section 6 is the evaluation of our method with American football and soccer scenes. Section 7 is the conclusion of this paper.

2. Related Work

The estimation of head and body orientation from low resolution videos has been studied for the purposes of video surveillance [10], [11], [26], [27], [28]. The body orientation of the subject, which our framework predicts, has also been used for group activity recognition [29], [30], [31] as context features between interacting people.

There are two main approaches for body pose estimation from a fixed camera view: 1) human pose/orientation estimation from a *single image*; and 2) human body pose or orientation estimation from *videos* based on the position of the person/head *tracker*. Inferring the skeletal body pose or body orientation is quite useful for many applications, such as searching for semantic key poses from TV shows [32], recognizing pose and clothing attributes of a person [33], recognizing the interaction between two people [34], [35], and recognizing the interaction between an object and a person [36], [37]. We will review human pose estimation methods from a single image in Section 2.1, and human pose estimation methods from videos in Section 2.2.

Another categorization of related work is the human pose estimation for (1) the *skeletal* pose and (2) the head/body *orientation*. Our proposed poselet-regressor is the skeletal pose estimator of the (simplified) 2D spine line while the proposed spine-conditioned body orientation estimators are the body orientation estimators. The single image methods in Section 2.1 are related to the poselet-regressor, while the tracker-based orientation estimation methods in Section 2.2 are related to our spine-conditioned body orientation estimators.

While skeletal pose estimations are mostly performed using a part-detector strategy, our strategy is to also use the global aligned window appearance to estimate the skeletal spine pose. Our tracker-based joint location estimation of the poselet-regressor is inspired by our previous work [25], which also uses the body tracker for the window alignment and used random classification forests to estimate the lower body joint locations.

2.1 Human Pose Estimation from a Single Image

There are many papers that try to estimate the skeletal body pose from a *single image* [15], [19], [24], [32], [33], [34], [35], [36], [37]. These approaches mainly estimate the *frontal* 2D skeletal bone of the subject and the regions of each part of the body.

The most popular approach to skeletal human pose estimation uses pictorial structures [38], [39], where the body parts configuration of the person in the image is represented as a graphical tree model of body part region appearances. After the success of the Deformable Part Models (DPM) [40], the pictorial structure framework for people detection is extended with DPM to jointly detect the part locations with structured prediction, such as a flexible mixture-of-parts model (FMP) [20].

FMP [20] is robust for the frontal pose images because it employs a tree-structured graphical model of part detectors, which are trained as the mixtures of the part appearances from the training dataset using k-means. While FMP [20] gives very good results for frontal human poses, it cannot infer the partially-occluded side poses accurately because the tree-structure of the graphical model does not appear when the arms and legs partially occlude each other. In team sports videos, there are many side-view pose appearances in which it is difficult to estimate the parts locations using part-based models.

Part-model approaches find it difficult to estimate the pose: (1) when multiple parts are occluded; (2) when the image is low-resolution and the parts have similar appearances, making it hard to discriminate each part correctly; and (3) when the person is in non-frontal side poses, which are hard to model with tree part-models of pictorial structure.

There are also per-frame classification or regression methods for estimating each part location from a single depth image [18], [41]. Additionally, there are part-model approaches using multi-view images [42], [43]. In contrast, our approach does not use depth image or multi-view inputs, but only monocular RGB images.

Poselets: Detection of One Specific Pose

Reference [15] proposed human (partial) pose detectors, poselets, which can be also regarded as a human (partial) pose silhouette detector. The poselets have been also used for the middle-level parts for human pose estimation methods with part-based models [19]. Reference [21] proposed the poselet-conditioned pictorial structures approach, which uses the poselets as a mid-level representation of multiple body parts. While pictorial structures using poselets [19], [21] can cover the pictorial structures with only local part detectors [20], classification still depends on the pictorial structures and cannot cover many types of articulation (with only a few specific poselets). Particularly when the arms or legs are partially occluded or hidden behind other parts, we need more poselets (exemplars) to represent those person apperances. This makes it more difficult to represent a huge number of part configurations, even when occlusion or part-disappearance occurs.

The poselets can be also used for key-frame extraction such as in Ref. [44] for activity recognition via key-frame responses. However, poselets cannot detect detailed or in-between poses because poselets are discretized key-pose exemplar detectors (but they can detect key-poses as attributes or actions, see Refs. [24], [33]). Maji et al. [24] proposed multi-view poselets for the purpose of single image action recognition from the detected pose with action-specific poselets.

Pose Regression

Recently, appearance-based regression of some kinds of human poses has been studied [45], [46].

Classic approaches for estimating skeletal poses, such as Refs. [45], [46], [47], try to train regression models (with low dimensional latent variables) of typical human movements for individual activities (e.g., walking, jumping). References [47], [48] estimated the joint locations of a target walking person using a fixed side-camera view in each single frame of a video. In these papers, cameras are set to capture the person from side views on the road or street so that people can be captured in only side-view poses. This side-view camera setting is often used in gait recognition [49]. In contrast, our proposed method can estimate the body orientation and the spine pose from any camera view, because our poselet-regressor learns the various types of human upper body pose appearances from all camera views using only one model.

Conditional regression forests [17], [18], which divides the visual feature space into each-view spaces or some other subcategories, is the closest approach to our proposal. For facial images, when the view of the face is restricted, visual patterns of fiducial points become smaller because the facial parts cannot move so much. However, the whole-body appearance has a wide variety of patterns (body orientation or spine angle, in our case) even when the camera view is restricted. Also gaze direction estimation from the eye-region appearance is explored with conditional regression forests conditioned on the head pose [50]. For non-articulated objects, regression-based pose estimation [51] can be done. However, articulated-pose estimation from RGB images has not yet been explored, while some depth-based methods have proposed the regression of part locations [18], [41].

2.2 Human Orientation Estimation from Low-Resolution Videos

Head and body orientation estimation approaches during *people tracking* have been proposed, mostly for surveillance videos [10], [11], [13], [26], [52], [53]. Another popular scene setting is the frontal video of automobiles [27], [28]. These methods train scene-specific or clothing-specific head or body orientation classifiers, which typically classify the horizontal eight directions into eight classes, by combining those classifiers with the head or body trackers to jointly estimate the orientations and the location with filtering techniques. They mainly estimate the head and body orientations of *pedestrians* and cannot deal with poses where the subjects are bending their upper bodies. While we would like to review only body orientation estimation methods, since our proposed method estimates the body orientation in each frame, we will also review head orientation estimation methods. The latter are closely related and adopt very similar approaches to classifying the direction, and our method uses the same head tracker as used in this work [13], [26], [52], [53]. Those head orientation methods are also combined with body orientation estimation [9], [10], [54].

Benfold and Reid [52] proposed the first approach that adopts feature-selection for learning a per-frame head orientation classifier using randomized trees with a color feature. Benfold and Reid extended [52] in Ref. [13] by introducing a HOG features [14] with a color feature to create a robust head pose classifier and proposed a multi-target tracking scheme using a HOG-based head detector and an optical-flow tracker for surveillance videos. This per-frame classification with tracking-by-detection proposed in Ref. [13] has been a standard approach for recent head or body orientation estimation methods [10], [11], [16].

For team sports videos with low-resolution settings, Hayashi et al. [16], our previous version of this work, proposed a head and body orientation estimation and spine pose estimation of American football players in videos. This work performs head and body orientation classification independently with tracking-by-detection with a player tracker, as well as head detection within each frame. In sports videos, it is easier to train a robust head orientation classifier than training it for surveillance videos, because the visual appearance of the head region is similar between different players wearing similar uniforms, especially helmeted players in our experimental American football videos. However, since head appearance is versatile in other team sports (e.g., basketball or soccer) and in surveillance videos such as Ref. [13], the head orientation classifier needs higher dimensional or discriminative features than for sports players. Conversely, body features in sport videos have more pose variations than in pedestrians. In this sense, tackling various types of postural appearances of sports players is the main focus and contribution of this work.

Schulz et al. [27] proposed a joint head pose estimator and head localizer for pedestrians for the risk assessment of car drivers. Later, Schulz et al. [28] proposed a sequential Bayesian tracking extension of Ref. [27] with a particle filter. In Ref. [28], they use a head pose classifier result as the per-frame likelihood of a particle filter, and jointly predict and update the head location and head pose by tracking a pedestrian on video. Benfold et al. [26] used conditional random fields to train the head pose estimator of pedestrians in an unsupervised manner in a new video scene. These papers make use of the temporal transition constraints on head location and also the temporal continuity of the head orientations of pedestrians via filtering. While the head locations can exist only in the upper region of the detection window because pedestrians are always standing and walking upright [27], [28], the head locations of sports players have larger variation because they often bend their bodies and sometimes dive into opposing players.

Compared with the Town Dataset setting of Benfold et al. [13], [26], players in team sports videos have much more random transitions of the head/body orientation between frames and it is harder to assume the smoothness of the head/body orientations between consecutive multiple frames. For this reason, we will investigate *per-frame* classification of body orientation in this paper without using a temporal connection while the head tracker is performed with a Kalman filter, which assumes temporal smoothness of (only) head locations.

Cheng et al. [9] proposed a temporal framework for joint estimation of body orientation and location of the subject pedestrian using a particle filter with sparse codes of multi-level HOG features. Baltieri et al. [11] proposed body orientation estimation for pedestrians using the mixture representation of Extremely Randomized Trees classifiers. These methods have only been tested for standing pedestrians, while our method covers even non-standing bending poses owing to the flexibility of our poselet-regressor. In addition, this work and our previous work [16] estimate the upper body orientation using only upper body HOG features, while previous papers estimate the (whole) body orientation using the appearance of the whole body.

2.2.1 Joint Estimation of Head and Upper Body Orientation from Videos

Chen et al. [54] proposed joint tracking of head and body pose in surveillance videos. They used a particle filter to jointly estimate the head and body orientation combined with the movement direction. Later, in Ref. [10], they extended their work to the semi-supervised learning setting with their own kernel learning scheme by learning the relationship between the parameters governing head orientation, body orientation, and movement direction.

Different from Refs. [10] and [54], which leverage the assumption of the lower velocity of pedestrians and combine the velocity with the body orientation, our method tries to deal with the upper body appearances of sports players at higher speeds, which are already aligned by the head tracker without needing to make a connection between movement direction and body orientation. For this reason, good alignment by the head tracker is key in our proposed method, because we do not depend on the relationship or a prior from the movement direction and perform only per-frame body orientation classification.

Poselets: detector of one specific pose

Reference [15] proposed human (partial) pose detectors, poselets, which can be also regarded as a human (partial) pose silhouette detector. Poselets have been also used for the middle-level parts for human pose estimation methods with part-based models [19]. The poselet framework has an advantage in creating detectors for side-view poses, which are hard to deal with for part-based whole-person models [20].

Poselets can be also applied for key-frame extractor such as Ref. [44] for activity recognition via key-frame responses. However, poselets cannot detect detailed or in-between poses because they are discretized detectors. Poselets can only detect discretized rough poses (See Fig. 2 in Ref. [44] for poselets examples), while poselets can detect side-view poses that occur often in activity recognition videos.

3. Overview of Proposed Framework

In this section, we will show the overview of our framework for estimating the upper body pose of the player: the *spine pose* and the *body orientation*. **Figure 4** shows the flowchart of our proposed framework. Our framework is composed of the following three steps:

(1) Tracks the head region in each frame with the head tracker [13] to estimate the head center location in each frame.

(2) Estimates the relative pelvis position against the head center in each frame with the poselet-regressor, using the HOG features of the upper body region aligned to the head center as input. This step results in estimating the spine pose.

(3) Estimates the body orientation with a classifier selected by the spine angle value of the player. (Optionally: estimates the head orientation with a head orientation classifier in the same way as in our previous work [16].)

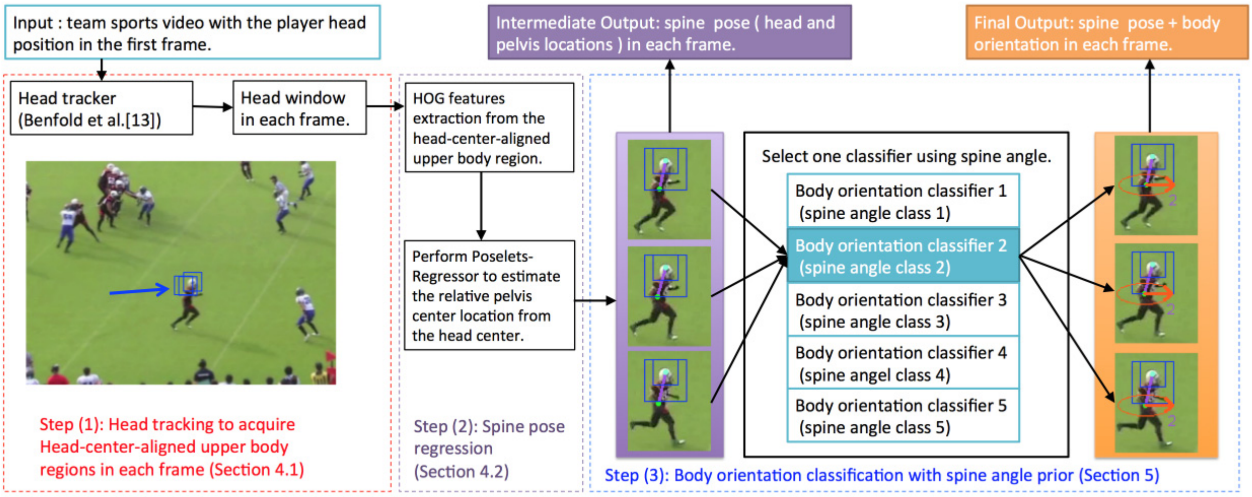

Input : team sports video with the player head position in the first frame.
Intermediate Output: spine pose (head and pelvis locations) in each frame.
Final Output: spine pose + body orientation in each frame.
Head tracker (Benfold et al.[13])
Head window in each frame.
HOG features extraction from the head-center-aligned upper body region.
Perform Poselets-Regressor to estimate the relative pelvis center location from the head center.
Select one classifier using spine angle.
Body orientation classifier 1 (spine angle class 1)
Body orientation classifier 2 (spine angle class 2)
Body orientation classifier 3 (spine angle class 3)
Body orientation classifier 4 (spine angel class 4)
Body orientation classifier 5 (spine angle class 5)
Step (1): Head tracking to acquire Head-center-aligned upper body regions in each frame (Section 4.1)
Step (2): Spine pose regression (Section 4.2)
Step (3): Body orientation classification with spine angle prior (Section 5)

Fig. 4 Proposed framework.

Steps (1) and (2) estimate the spine pose $\mathbf{s}_t = (\mathbf{h}_t, \mathbf{p}_t)$ where $\mathbf{h}_t = (x^h_t, y^h_t)$ is the head location and $\mathbf{p}_t = (x^p_t, y^p_t)$ is the pelvis location in each frame t. Step (3) estimates the body orientation $\mathbf{o}^b_t \in \{0, 1, \ldots, 7\}$ of the player in each frame (Fig. 3 (c)). The spine angle calculated from the spine pose is used to select one corresponding upper body orientation classifier f^b_s from multiple body orientation classifiers for each spine angle range (Fig. 8). In other words, the spine pose acts as a mediator between the steps (1) and (2) and step (3). We will define the procedures of steps (1) and (2) in Section 4, then define the procedure of step (3) in Section 5.

4. Head Tracking and Spine Pose Estimation with Poselet-regressor

To estimate the relative pelvis position $\mathbf{p}_t = (x^p_t, y^p_t)$ from the head center location $\mathbf{h}_t = (x^h_t, y^h_t)$ estimated by the head tracker at each frame t, we propose a body spine pose regressor, which we call the *poselet-regressor*. Using the head tracker and our poselet-regressor, our framework can estimate the 2D spine pose of the player $\mathbf{s}_t = (\mathbf{h}_t, \mathbf{p}_t)$ at each frame t of the video by regression (see **Fig. 5**).

With the global coordinate head center location $\mathbf{h}_t = (x^h_t, y^h_t)$ in a video frame (**Fig. 6** (a)), our poselet-regressor first calculates the multi-level HOG feature $\mathbf{x}^b_t$ within the upper body region around the head center location $\mathbf{h}_t$ (Fig. 6 (b)). Then it estimates the relative pelvis position $\mathbf{p}'_i = (x'^p_i, y'^p_i)$ from the local coordinate head center $\mathbf{h}'_t = (0, 0)$ (Fig. 6 (c)) using $\mathbf{x}^b_t$ for the random regression forests input vector. In other words, we use the selected visual features from the upper body region for this regression.

4.1 Head Tracking

To estimate the head center locations in each frame, we first perform head tracking for one subject player in a team sports video. We employed the head tracking approach proposed by Benfold et al. [13] to track the head region. This method tracks the head region of a person using tracking-by-detection with a Kalman filter. It uses a SVM (support vector machines) head rectangle detector with HOG features as the likelihood of Kalman filter and uses local feature tracking results to predict the next state. In our experiments, we trained scene-specific 24 × 24 head detectors for each scene. We regard the center of the tracked head regions in each frame t as the head location $\mathbf{h}_t$ used in the later steps.

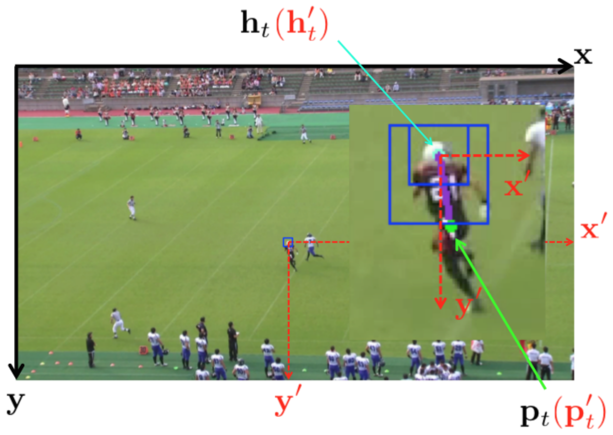
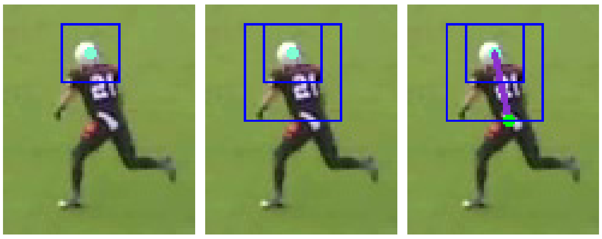

$\mathbf{h}_t\,(\mathbf{h}'_t)$
x
x'
x'
y'
y
y'
$\mathbf{p}_t\,(\mathbf{p}'_t)$

Fig. 5 Local coordinate system for the poselet-regressor and the global coordinate system for the head tracker. The black axes **x** and **y** are the global coordinates and the red axes **x′** and **y′** in the magnified player image are the local coordinates. Pelvis position estimation is performed in these local relative coordinates.



(a) Head region tracked by the head tracker.

(b) Upper body region(larger rectangle) used for poselet-regressor.

(c) 2D spine estimated by poselet-regressor to estimate pelvis center position.

Fig. 6 Estimating procedure of spine pose with head tracker and poselet-regressor.

4.2 Poselet-Regressor of Spine Pose

After the head region is estimated in each frame of the video in the first step, we employ the poselet-regressor to estimate the 2D spine pose, which consists of the 2D head position $\mathbf{h}'_t = (x'^h_t, y'^h_t)$ and the 2D pelvis center position $\mathbf{p}'_t = (x'^p_t, y'^p_t)$ in a local coordinate system whose origin $\mathbf{h}'_t = (0, 0)$ is the global head location $\mathbf{h}_t = (x^h_t, y^h_t)$ in each image (Fig. 5). The upper body region for the poselet-regressor is located 20 pixels to the left and 12 pixels above the head center location $\mathbf{h}_t$ (Fig. 3 (a)).

The proposed poselet-regressor does not try to detect segments of the subject person like the poselets detector [15], [22] or FMP detector [20]. Instead, our poselet-regressor estimates the continuous change of the relative joint position (pelvis center) from the center position (head center) using the selected discriminative features within the head-center-aligned upper body region HOG appearances. Adopting this design of *relative joint location estimation* by discarding the detection and segmentation ability of the original poselets [15], [22], our poselet-regressor can obtain the relative movement of one joint location from the center joint of the poselets window in the upper body visual feature space. Our poselet-regressor can also be regarded as the regressive version of the label-grid classifier [25], a visual grid classifier of the lower body joint location from the pelvis center with HOG-grid resolution. The center of alignment of the HOG features window in the label-grid classifier [25] is the pelvis center, while the alignment center of the poselet-regressor is the head center position in this paper.

Given this head-center-aligned upper body region at frame t, the poselet-regressor estimates the (regressed) 2D pelvis center location $\mathbf{p}'_t = (x'^p_t, y'^p_t)$ from the head center location $\mathbf{h}'_t = (0, 0)$. $\mathbf{p}'_t = (x'^p_t, y'^p_t)$ can be also regarded as an offset vector from the local origin $\mathbf{h}'_t$. We train the poselet-regressor $f^s(\mathbf{x}^b_t)$ as regression forests [23] to estimate $\mathbf{p}'_t$ with the selected features from the whole HOG feature vector $\mathbf{x}^b_t$ in the upper body region:

$$\hat{\mathbf{p}}'_t = f^s(\mathbf{x}^b_t) \tag{1}$$

Figure 7 shows some example results. Here, we assume that the head center locations in each frame were already tracked by the head tracker in Section 4.1. We then use the poselet-regressor to estimate $\mathbf{p}'_t$ in each frame.

We adopt the same feature vector as Ref. [11], a three-level pyramid of HOG features within the upper body region as a D dimensional feature vector $\mathbf{x}^b_t \in \mathbb{R}^D$ from the $W \times H$ window. The block size of the HOG is 2×2 at every level. In the same way as in Ref. [11], a dimensionality reduced image (by PCA) is used as an input for HOG features calculation.

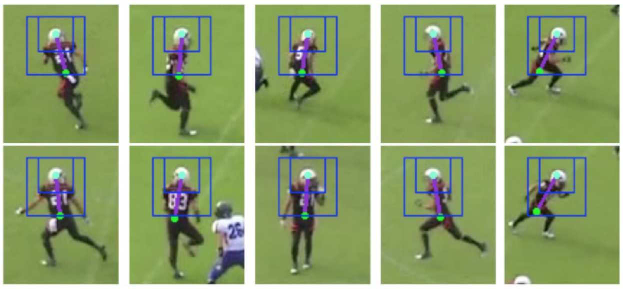


Fig. 7 Example results of the relative pelvis position estimation using poselet-regressor.

We name our relative pelvis location regressor as the poselet-regressor because it can be regarded as a regressive version of the poselets framework [15]. While the original poselets are *detectors*, our poselet-regressor trains a regressor of the relative pelvis offset using the upper body visual features, whose head positions are aligned. Note that one poselet detector is typically trained from people with many types of clothing and hair styles, while our poselet-regressor (in this paper [*1]) is trained from the upper body regions of different players and poses wearing only one specific American football or soccer uniform type (see our experimental setting in Section 6).

Training the Poselet-regressor

The poselet-regressor is trained from the dataset $\mathcal{D}_{pose} = \{(\mathbf{x}^b_i, \mathbf{p}'_i), i = 1 \ldots, n_b\}$, where n_b denotes the number of samples in the dataset $\mathcal{D}_{pose}$, $\mathbf{x}^b_i$ denotes the feature vector from the upper body region aligned with the head center, and $\mathbf{p}'_i$ denotes the pelvis center offset from the head center location in local coordinates. As already mentioned, we use regression forests [23] as the poselet-regressor to train the local pelvis center location $\mathbf{p}'_i$ in a continuous 2D image space using the dataset $\mathcal{D}_{pose}$. Each sample $(\mathbf{x}^b_i, \mathbf{p}'_i)$ in $\mathcal{D}_{pose}$ is collected and labeled from the videos from the same match and the players from the same team.

Optionally, the original training dataset $\mathcal{D}_{pose}$ is augmented to the $\mathcal{D}^{aug}_{pose}$ with some slides of the head center position to make the poselet-regressor (and the body orientation classifiers in Sec. 5) recognize a slanted upper body appearance. The slide vector $\mathbf{v_s} = (x_s, y_s)$ where x_s denotes the x-axis slide value of the head center position and y_s denotes the y-axis slide value of the head center position. By using slide vectors, head center positions are augmented while the images in the original $\mathcal{D}_{pose}$ remain the same. Since our body pose estimators depend on the head-center-aligned upper body appearance, the drift of the head tracker often provides a little slanted upper body appearance. Although HOG [14] has local deformation invariance through block histogram quantization, our data augmentation procedure will provide additional robustness to the not-well-aligned head tracking results (we test this setting on women's soccer scenes in Section 6.

Potential Advantage of the Poselet-regressor

Figure 7 shows some example results from our poselet-regressor in our experiment (Section 6). These results show us two advantages of the design of the poselet-regressor. The first advantage is that the poselet-regressor realizes spine pose estimation for all types and views of sports player poses, even when some part-occlusions happen (Fig. 7), because it achieves regression using only randomly-selected features within a holistic upper body region. As discussed in Section 2, part-detector based methods can estimate poses when the pictorial structure of *all* parts can be found. Conversely, our poselet-regressor uses only head-center-aligned upper body region appearance so that every upper-body visual features (including part-occluded poses and side poses) can be trained from the training images. This is a

[*1] While poselet-regressor can be learned from people with various types of clothing, we only use it for one specific clothing type for the team.

necessary trait for the human pose estimation methods for team sports *videos*, because players run right or left with side-view poses and their arms often disappear from images because of part-occlusions.

The second advantage of the poselet-regressor is a separate and sequential estimation of *joint* locations via relative joint location estimation (in local coordinates). While FMP [20] and other *part*-based detectors need to decide the locations of all local parts simultaneously, our poselet-regressor enables part-location estimation one by one with global appearance. This flexibility evades the necessity of part localization for bone estimation (in our case, spinal bone) and also achieves the spine pose estimation that most previous work does not primarily focus on (or which is often ignored). Moreover, by integrating it with head tracking (in this paper), a better alignment of the body appearance is produced, which makes the pose estimation problem simpler [*2], and can be done even while tracking the person. The head region has good *rigidity* for tracking and achieving good alignment for the pose estimators, while the previous body orientation estimation work depends on the pedestrian detection window as alignment.

5. Multiple Body Orientation Classifiers with Spine Angle Prior

The third and final step in our framework is to estimate the body orientation using the body orientation classifier with a corresponding spine angle range. We train each n_s upper body orientation estimator as eight-class random decision forests $\mathcal{F}^b = \{f_s^b, s = 1, \ldots, n_s\}$ with a training dataset from one team in a specific scene. Each body orientation classifier f_s^b is responsible for estimating the body orientation $\mathbf{o}_t^b \in \{0, 1, \ldots, 7\}$ (Fig. 3 (c)) within the corresponding spine angle range class using the input feature vector $\mathbf{x}_t^b$ at frame t:

$$\hat{\mathbf{o}}_t^b = f_s^b(\mathbf{x}_t^b) \qquad (2)$$

After tracking the head region and estimating the spine pose in the two previous steps (Section 4), we select a s-th class f_s^b from $\mathcal{F}^b$ to estimate the body orientation according to the spine angle θ_t of the player at frame t (**Fig. 8**).

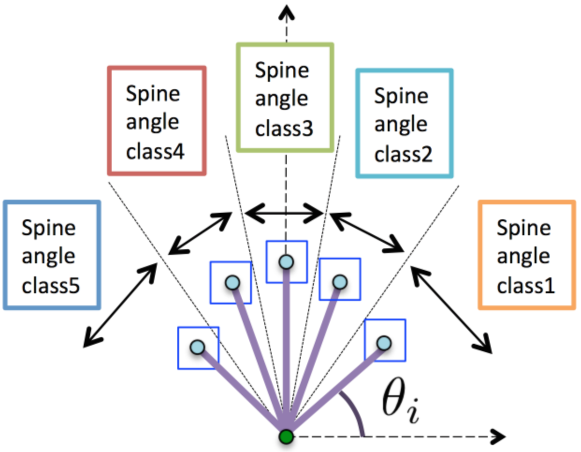


Fig. 8 Spine angle classes. The blue circle is the head center **h** and the green circle is the pelvis center **p** of the subject player. The spine angle range of the training dataset is divided into five spine angle classes.

We use random decision forests [55] to train each upper body direction classifier using the subdatasets divided by the spine angle value (see **Fig. 9**). We use the same feature window size of the poselet-regressor calculated from $W \times D$ upper body region (Section 5.1) for the two-level HOG features of the multiple body orientation classifiers.

5.1 Learning Multiple Upper Body Orientation Classifiers by Dividing the Dataset According to the Spine Angle

To estimate the body orientation of the player, we use one classifier f_s^b selected from n_s classifiers $\mathcal{F}^b = \{f_s^b, s = 1, \ldots, n_s\}$, where f_s^b is a body orientation classifier for each spine angle class s (Fig. 8).

To train each f_s^b, we first prepare the dataset $\mathcal{D} = \{(\mathbf{x}_i^b, \mathbf{o}_i^b, \mathbf{s}_i), i = 1 \ldots, n_b\}$ as in Section 4.2, where $\mathbf{x}_i^b$ is the upper body region feature vector, $\mathbf{o}_i^b \in \{0, 1, \ldots, 7\}$ is the body orientation label, and $\mathbf{s}_i = (\mathbf{h}_i, \mathbf{p}_i)$ is the spine pose label of the i-th sample in the dataset, respectively. After the learning procedure, each f_s^b uses different features selected from the same upper region feature $\mathbf{x}_i^b \in \mathbb{R}^D$. We also define the spine angle θ_i on the image plane as the angle against the x-axis direction. This spine angle θ_i can be calculated from the spine pose estimated with the 2D spine pose $\mathbf{s}_i$ (see Fig. 8).

Next, we divide $\mathcal{D}$ into n_s−subdatasets $\{\mathcal{D}_s, s = 1, \ldots, n_s\}$ according to the angle value θ_i of each i-th instance in $\mathcal{D}$ (Fig. 8). Spine angle θ_i space is separated into n_s regions, which we call *spine angle classes*, according to the spine angle value θ_i calculated from $\mathbf{h}_i, \mathbf{p}_i$. With each $\mathcal{D}_s$, we learn f_s^b as random decision forests. This dataset preparation procedure is shown in Fig. 9.

We use $n_s = 5$ by default as showed in Fig. 8. After the preliminary tests with our experimental datasets, we decided to divide the spine angle θ_i space into the following five classes:

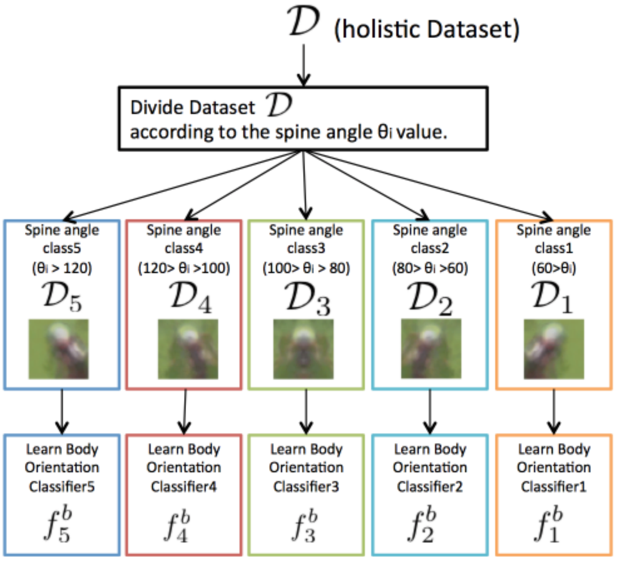

$\mathcal{D}$ (holistic Dataset)
Divide Dataset $\mathcal{D}$ according to the spine angle θi value.
Spine angle class5 (θi > 120) $\mathcal{D}_5$
Spine angle class4 (120> θi >100) $\mathcal{D}_4$
Spine angle class3 (100> θi > 80) $\mathcal{D}_3$
Spine angle class2 (80> θi >60) $\mathcal{D}_2$
Spine angle class1 (60>θi) $\mathcal{D}_1$
Learn Body Orientation Classifier5 f_5^b
Learn Body Orientation Classifier4 f_4^b
Learn Body Orientation Classifier3 f_3^b
Learn Body Orientation Classifier2 f_2^b
Learn Body Orientation Classifier1 f_1^b

Fig. 9 Learning multiple body orientation classifiers (random decision forests) by grouping datasets into the subsets having the same spine angle range. This conditional classifiers learning makes the random decision forests easier to select discriminative features for body orientation classification from only the spine pose aligned HOG features in each D_s.

[*2] Note that we insist on simplicity for (only) the pose estimation problem during tracking from low-resolution surveillance videos.

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

At test time, after the value of s was selected using Eq. (3), the corresponding classifier f^b_s was used to estimate the body orientation.

This spine-driven dataset grouping both at training time and test time makes it easier for each random decision forest to select more discriminative split functions at each node rather than using the whole dataset. The reason is that images in one grouped subdataset $\mathcal{D}_s$ are more similar and thus it is easier to automatically select the informative (different) feature dimensions for random forest learning (see the average images in each subdataset $\mathcal{D}_s$ for each spine angle class in Fig. 9).

Our idea of using spine angle value as a conditional prior for selecting body orientation classifier was inspired by the conditional regression forests in Ref. [17], but there is a difference. While Ref. [17] models the conditional prior of the head orientation by selecting the T trees from one holistic random decision forest for all the head orientations (conditions), our method trains multiple independent random decision forests for each spine angle range separately and do not use any trees from the different conditions. The reason for adopting this different approach is that the upper body appearance does not have continuous (manifold-like) feature space along the spine angle because arm and head appearances are sometimes inconsistent with the spine angle changes.

6. Experiments

We evaluated our framework with videos from an American football game and a women's soccer game. We performed experiments for each pose estimation step, namely (1) head and pelvis center estimation with head tracker and the pelvis center (Section 6.2), and (2) the body orientation estimation (Section 6.3). All videos were captured at a high place in the stadium with fixed cameras at 29 fps. In the American football videos, horizontally moving players are mainly shown. However, in the women's soccer videos, players often move diagonally (to the goal or the opponents). Hence, we can expect different body orientation statistics for each scene.

In each scene, we prepared our experimental data by dividing videos into a test dataset and a training dataset. The datasets were prepared with images and manually annotated labels of the spine pose and the body orientation. The American football videos were captured from one match of Panasonic IMPULSE [*3], which is a Japanese professional American football team. The women's soccer videos were captured from the match "*Waseda-Keio Game*" of the Keio University womens soccer team [*4].

We evaluate each of the three steps of our method in the following three subsections (Sections 6.1, 6.2, 6.3). We will focus only on the poses of the black-uniformed players in the American football videos and focus only on the poses of the brown-uniformed players in the women's soccer videos (see the figures for those uniform colors).

We prepared a test video dataset $\mathcal{D}_{test}$ with 12 test videos for the American football game and prepared $\mathcal{D}_{test}$ with 10 videos for the women's soccer game, with 22 test video sequences in total. Each test video sequence in both games is composed of 80 frames, and we track one player in each video to evaluate our framework. To evaluate the advantage of using only the upper body region appearances for human pose estimation, some of the tests include frames with some lower body occlusions between players. In addition, to demonstrate the advantages of our body orientation classifiers conditioned by spine angle class, some of the tests include bending or twisting poses (where the upper body orientation and the lower body orientation (movement direction) are different). Later in Section 6.4, we discuss the effectiveness of our method for occlusion cases and bending poses by showing the visualized result images.

For each scene, we independently trained a head detector (which we will not evaluate), the poselet-regressor of the pelvis center and the body orientation classifiers from the training dataset $\mathcal{D}_{train}$ with manually labeled poses, which only includes team players from one specific team. In addition, to overcome the head-center drift of the head tracker in the soccer scene, we augmented the $\mathcal{D}_{train}$ of the soccer scene to $\mathcal{D}^{aug}_{train}$ to train pose estimators that also understand shifted examples. We made $\mathcal{D}^{aug}_{train}$ with slide vectors $(-8, 0), (8, 0), (0, 8), (0, -8)$.

To train the body orientation classifiers with the symmetric images and labels of the originally labeled samples, we resampled the symmetric samples using the following procedure. First, we made the dataset $\mathcal{D}_{train}$ in specific scenes (e.g., American football or women's soccer, in our experiments) by manually labeling the images from the training videos. Second, we made a flipped copy of $\mathcal{D}_{train}$ as $\mathcal{D}'_{train}$, which is composed of the flipped images and flipped labels from $\mathcal{D}_{train}$. Finally, we acquired the whole training dataset $\mathcal{D} = \{\mathcal{D}_{train}, \mathcal{D}'_{train}\}$ with symmetrically resampled images and labels. In the American football scenes, about 40 percent of the images were used in the test videos (tests 2,3,4,10 and 12) [*5]. In the soccer scene, we completely separated the test dataset and the training dataset. Note that again our goal was training scene-specific (or sport-specific) body orientation estimators that can deal with bending poses. Hence, we tried a supervised-learning approach as a first step toward unconstrained sport pose estimators.

We trained and tested our method for one specific team (with the same clothing but with different body shapes). $\mathcal{D}_{train}$ in the American football scenes includes 16334 (after reflection, original dataset includes 8167 samples) feature vectors $\mathbf{x}^b_i$ calculated from images and their labels $(\mathbf{o}^b_i, \mathbf{s}_i)$. $\mathcal{D}_{train}$ in the women's soccer scene includes 1053 examples, which is small for random decision forests training. For this reason, we augmented the original

[*3] http://panasonic.co.jp/es/go-go-impulse/

[*4] http://keio-soccer.net/

[*5] Even though some test sequences were overlapped with the training dataset, most of the test images in each frame were different from the ones in the training dataset because tracking provides a shifted upper body window appearance based on the drift of the head tracker

with reflections and four slide vectors. After data augmentation, we obtained $2822 \times 5 slides = 5265$ examples in total). We used a 48×48 upper body region for the American football scenes and a 64×64 upper body region for the women's soccer scenes, from which we calculated the feature vectors for both the poselet-regressor and the body orientation classifiers. For the body orientation classifiers of the women's soccer scene, we change the center of the upper body region to the pelvis center estimated by the poselet-regressor while the center of the upper body region for the classifiers of American Football scene is the head center estimated by the head tracker (larger window shows the regions for body orientation classifiers in each figure in this paper). And we also make the upper body region size for the soccer scene body orientation classifiers to 64×64 in order to include arm regions of all training samples.

To train five body orientation classifiers in each spine angle range, we divided the training dataset $\mathcal{D}_{train}$ (or $\mathcal{D}^{aug}_{train}$ in the soccer scene) into five subdatasets $\{\mathcal{D}_s, s = 1, \ldots, 5\}$ and trained each spine angle class classifier f^b_s with D_s independently as in Section 5.1. In the American football scenes, each $\mathcal{D}_s$ had 934, 4005, 6456, 4005, and 934 examples (16334 in total) respectively. In the women's soccer scenes, each $\mathcal{D}_s$ had 336, 437, 1216, 437, and 336 images (5265 in total), respectively. For training the American football scene poselet-regressor, $\mathcal{D}_{train}$ with 16334 examples was used. For training the women soccer scene poselet-regressor, $\mathcal{D}^{aug}_{train}$ with 5265 was used.

We performed the evaluations for players on the lower side of the field in each scene so that the image scale of the players became almost the same. For the same reason, we also collected training examples from the players who played on the lower side of the field. As a result, we could only consider a small range of scales of the players in the experiments (and also the camera views and distance).

6.1 Head Tracking

To apply the head tracking algorithm from Ref. [13], we trained head detectors for each scene as linear SVMs using HOG features [14] with 4×4 cells within a 24×24 pixels window so that the head region was included in the $50--70$ percent of the window size. The first column in **Table 1** is the result of the error in the head tracker for our test dataset. The error in the spine angle class will be evaluated in the next subsection.

6.2 Spine Pose and Spine Angle Class Precision

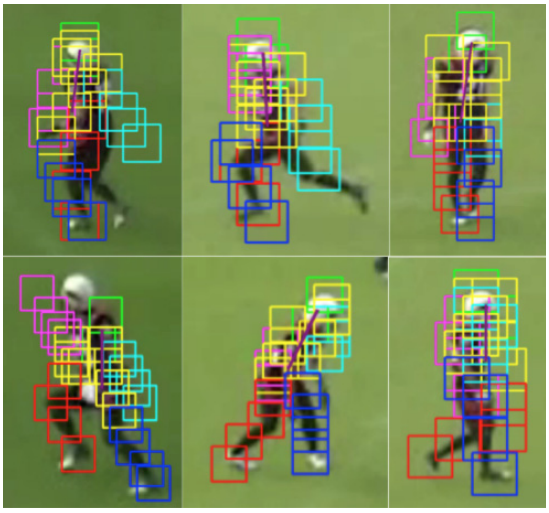
We evaluated the precision of the poselet-regressor using the test dataset from two perspectives: (1) precision of the poselet-regressor itself; and (2) precision of the assignment of the spine angle class. As a baseline of (1), we also evaluated the performance of the Flexible-Mixtures-of-Parts (FMP) [20] as the head center and the pelvis center estimator. Since FMP is a person detector, we used the 200×200 image centered at the head position from the head tracker to (re-)detect the FMP for this evaluation. We used software and the default model of FMP provided by the authors of [20]. We resized the 200×200 image to 400×400 size so that the FMP model could detect the person with the trained size.

Table 1 Average estimation error (Euclidean distance in pixels) of the head center and the pelvis center in each test dataset.

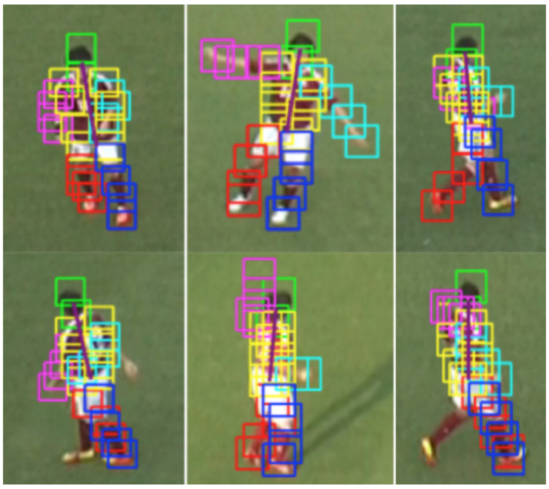
	American football	women's soccer
head center (ours)	3.99 (12 tests)	7.68 (11 tests)
head center([20])	10.44 (10 tests)	7.40 (9 tests)
pelvis center (ours)	4.03 (12 tests)	6.25 (11 tests)
pelvis center([20])	9.75 (10 tests)	7.69 (9 tests)

The first row in Table 1 shows the average error of the head center locations estimated by the head tracker with our estimators and FMP [20]. The second row in Table 1 is the result of the pelvis center location estimated by the poselet-regressor from the tracked head locations in each frame and the results of FMP [20]. Note that in both scenes, some tests are omitted for calculating the results of FMP (test 2,3,4 in American football and test 5 in women's soccer). The reason is that the subject player was not detected by the FMP because of inter-player occlusion. **Figure 10** shows some example results of FMP detection on our test sequences.

While our approach for estimating the pelvis center location is two-step estimation (steps (1) and (2)), the average error of the



(a) American football scene.

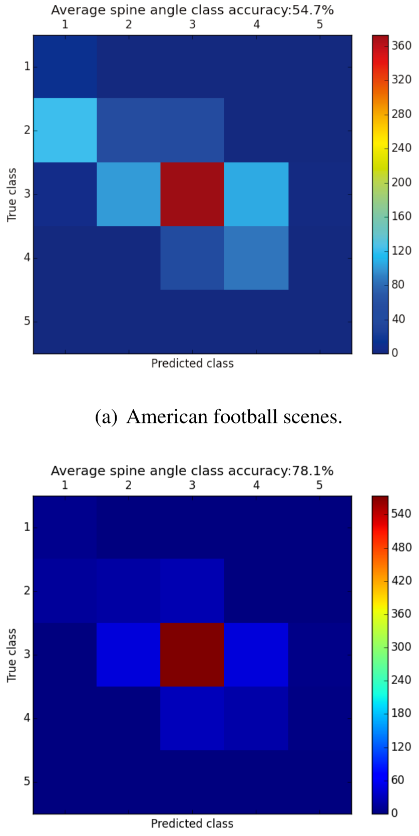


(b) Women's soccer scene.

Fig. 10 Example results of skeletal pose estimation with FMP [20]. The purple line is the spine pose between the head center and the pelvis center.

pelvis locations remains almost half the size of the cell size 8×8 of HOG features for the body orientation classifiers input vector. This makes HOG features pooling effective for estimating the same output body orientation class as long as the translation of the upper body window is small. In the American football scenes, our method predicted accurate spine pose even when the spine angle is acute or even when the pose is side-view. At the same time, FMP [20] also shows accurate results for spine pose estimation (Fig. 10). While the results for the two joints are good (for our body orientation classification in step 3), all parts of the FMP do not fit well for the subject person. While torso parts (yellow rectangles) and head parts (green rectangles) are well fitted to the players, arm parts and leg parts are not well fitted because of the hard occlusions or disappearance of those parts. Hence, FMP is not valid for our purposes, even though the head center and the pelvis center seem well fitted.

Next, we evaluated the precisions of the spine-angle classification of the whole test dataset $\mathcal{D}_{test}$. **Figure 11** shows the confusion matrix of the results of the spine angle classification performed with the whole test dataset $\mathcal{D}_{test}$ in each scene type. American football videos include mainly standing poses (spine class 2,3,4) and few bending poses (spine class 1,5). There are some samples whose true class 3 is wrongly classified as one of the neighboring classes 2 and 4. The results for the women's soccer videos shows the accuracy of the spine angle classification: 78.1 percent, which is higher than the results for the American football scene (54.7 percent).



(a) American football scenes.

(b) Soccer scenes.

Fig. 11 Confusion matrices of the spine angle class estimation.

While our spine angle range strategy reduces the effects of the spine pose estimation error by discretizing the spine angle, higher accuracy of the assignment of the spine angle class s also strengthens the accuracy of the spine angle class prior for selecting the appropriate body orientation classifier f_s^b. In this sense, we would prefer the more precise poselet-regressor, while this might be difficult for our problem setting with very low-resolution videos.

6.3 Body Orientation Precision

We evaluated the precision of the body orientation classifiers (Section 5) using the test dataset. We compared the result of proposed body orientation classifiers to the result of our body orientation classifier in our previous work [16], which uses only one body orientation classifier for the whole training dataset. To test a fair comparison in terms of alignment, we trained body orientation classifiers for the women's soccer scenes with a pelvis-aligned window, where the pelvis is located at (32, 48) from the top left corner of the 64×64 window (see Fig. 15).

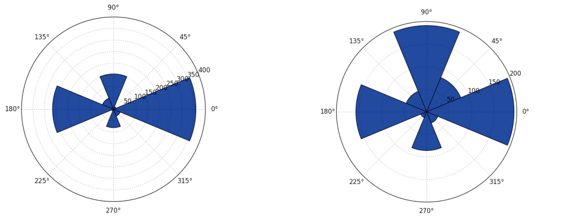
To compare with the body orientation classifier of [16], we also trained one body orientation classifier with random decision forests using the overall training dataset $\mathcal{D}_{train}$ for American football scenes, $\mathcal{D}_{train}^{aug}$ for the women's soccer scenes. As described in Section 4.2, we used the input feature vector for random forests with the 3-level pyramid HOG with a PCA-reduced one-channel image for our proposed method. For Ref. [16], we used a one-level pyramid HOG as in Ref. [16].

For the American football scenes, we used the spine angle class boundaries [60, 80, 100, 120] between two neighboring classes of Eq. (3). For the soccer scenes, we used different boundaries [70, 80, 100, 110] because the spine angles of the soccer players are not so acute as those of the American football players.

We compared the results of the proposed body orientation classifiers with the spine angle prior and the body orientation classifier of [16] from two perspectives with the same tracked results of spine poses: a multi-class classification perspective and a body orientation angle estimation perspective.

Multi-Class Classification

We first evaluated our body orientation classifiers as a multi-class classifier. **Figure 12** shows the body orientation class distribution, where 0 degrees is equivalent to class 4 and 180 degrees is equivalent to class 0 (see Fig. 3 (c) for the class index assignment). Since most of the subject players (who are mainly running back and wide receivers) run horizontally in the American



(a) Body orientation distribution in 12 American football tests.

(b) Body orientation distribution in 10 women's soccer tests.

Fig. 12 Body orientation class distribution (histogram) in each type of scene.

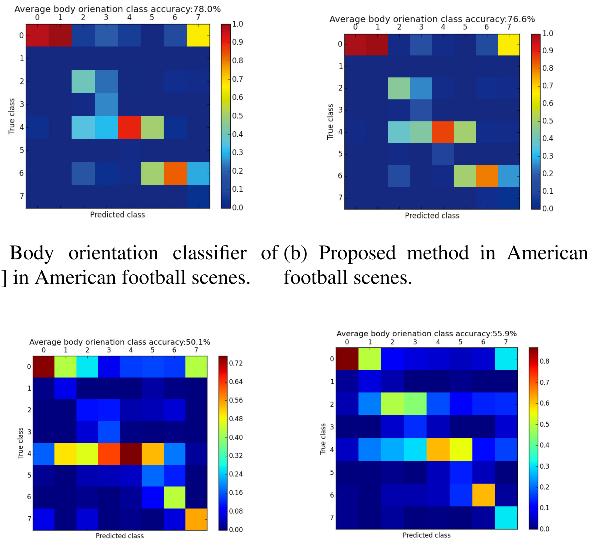
football scenes (Fig. 12 (a)), most of the body orientations are in horizontal directions (class 0 (left) or class 4 (right)) and there are only few diagonal orientations. In the women's soccer scenes (Fig. 12 (b)), the target brown clothing team is attacking to the right direction in the field. Most of the frames are in the right direction, with some diagonal orientations.

Figure 13 shows the confusion matrices of the classification results using the body pose classifier in Ref. [16] (Fig. 13 (a) and (c)), and the body orientation classifiers proposed in this paper (Fig. 13 (b) and (d)) in each scene. The confusion matrices for the proposed method show slightly more precise results than those of Ref. [16], while the class prediction accuracy is almost the same. However, Fig. 13 (c) shows a little more misclassification (i.e., predicted class 1 vs. true class 5 is salient in Fig. 13 (c)) while Fig. 13 (d) shows the more misclassification to the neighborhood classes.

Orientation Angle Error

As also evaluated in the other papers for head or body orientation estimation reviewed in Section 2, we evaluated the average angle error of the estimated body orientation angle from the same test results.

Table 2 shows the average estimation error of the body orientation angle in degrees for the test dataset $\mathcal{D}_{test}$ by our proposed method and our previous method [16] in each type of scene. Although the average errors of the proposed method shows that it performs the better than Ref. [16] in Table 2 in both scene types, this does not give us a good understanding of the overall results because the precision and accuracy are almost the same between the two methods. Hence, we will visualize many example results in specific tests and situations in the next Section 6.4 with further detailed discussion to provide evidence for each specific challenge.



(a) Body orientation classifier of [16] in American football scenes.

(b) Proposed method in American football scenes.

(c) Body orientation classifier of [16] in women's soccer scenes.

(d) Proposed method in women's soccer scenes.

Fig. 13 Confusion matrices of body orientation estimation results.

Table 2 Average estimation error (in degree) of the body orientation in each scene dataset. The baseline is our previous work [16].

	Proposed	Ref. [16]
American football scenes	20.90	23.57
Women's soccer scenes	39.99	47.02

6.4 Discussions

Running Poses

Since *running* is the most frequent action in team sports videos, the robustness of estimating running poses is the most important for evaluating the human pose estimation method for team sports. **Figure 14** shows some results from the proposed method and Ref. [16] in seven consecutive frames in the test dataset while the player is running (tests 1,2,3,4 in the American football scenes). Figure 14 (a) (test 1) is a typical example of a standard straight running case, which occurs very often in team sports videos. Figure 14 (b) (test 3) shows the results when the movement direction of the player is different from the body orientation (the player is moving to the left while the body orientation is in the upward direction). This capability of our framework is very important for team sports videos, where players often look at different directions from their movement direction. Figure 14 (c) (test 4) shows the results of twisting behavior (body rotation against the camera pose) during running. Even though Ref. [16] shows good results, the proposed method gives perfect results during the transition of body orientation (see 4th frame from the left on both rows in Fig. 14 (d)). **Figure 15** shows some key frames in the women's soccer tests. Figure 15 (a) and (c) (test1 and test7) shows running sequence examples. Compared with the American football scenes, the soccer scenes have more *diagonally running* players and diagonal body orientations: classes 1, 3, 5, and 7 in Fig. 3 (c).

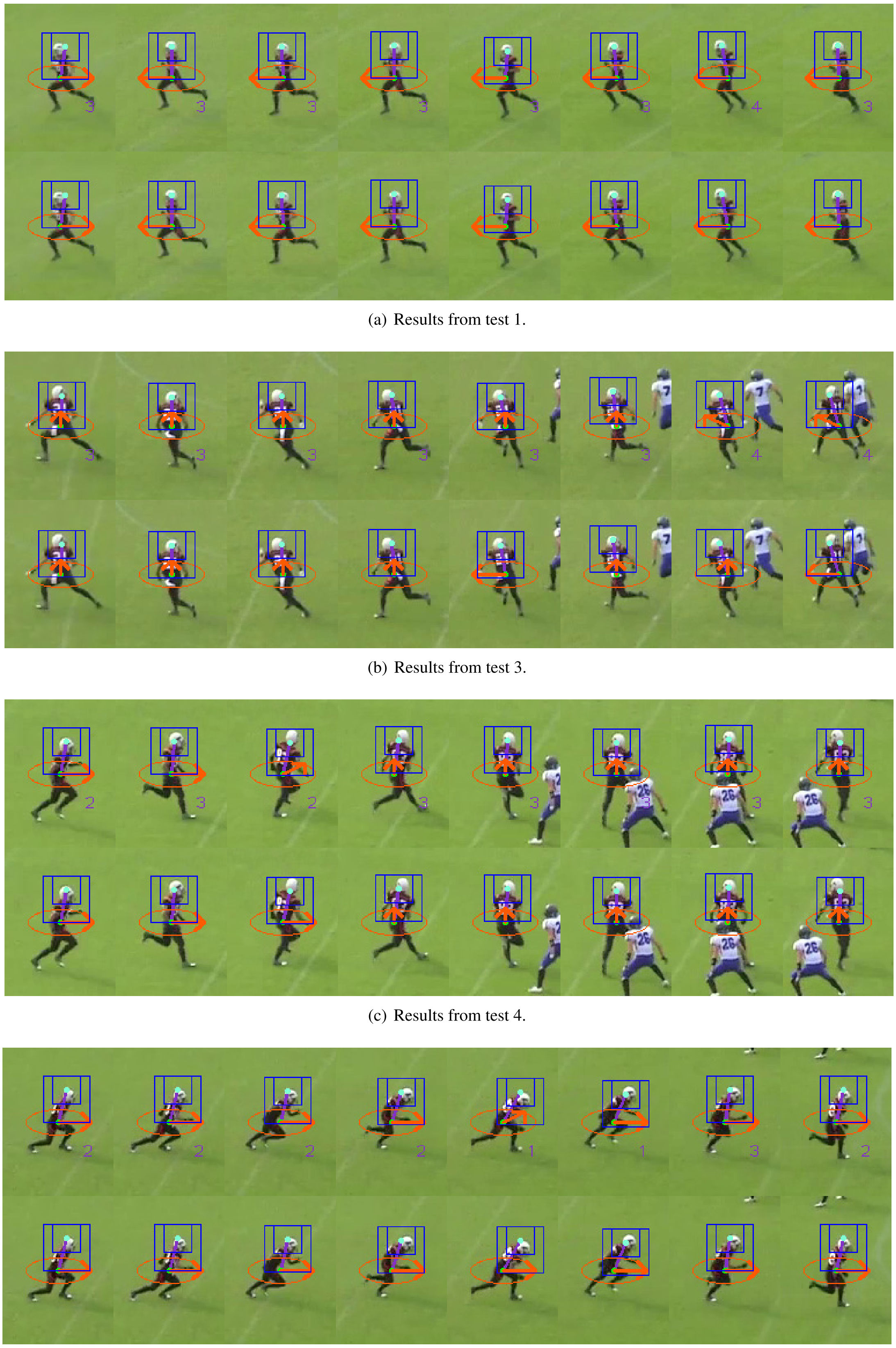
Occlusions and Using only the Upper Body Region

Figure 16 shows some cases with hard occlusions within the upper body region. Since our method only uses the upper body region (larger blue rectangle) for body pose estimation, the upper body poses are estimated correctly (Fig. 16 (b)). However, estimation of the upper body orientation tends to fail if most of the background becomes an unknown image pattern for the poselet-regressor and body orientation classifiers (Fig. 16 (a) and (c)). Even though the background in our videos consists of simple green flat texture and the random decision forests can select the features mostly of the foreground (player) HOG cells after the feature selection, our upper body pose estimation results become unstable when hard occlusions occur because we have not yet built the foreground-only selection. This is a very important problem in the application of our framework to other sports videos where the background consists of more complex textures.

Figure 15 (d) (test10 for women's soccer scenes) is also an example of the advantage of the usage of only the upper body region's appearance. Our method can estimate the upper body pose of soccer players even as they interact with the ball with their legs, because the ball does not affect the estimation at all as long as it does not enter the upper body region.

Bending Poses

Figure 17 shows the results of bending (side-view) poses. If the spine pose is estimated correctly, the upper body orientation is also classified correctly (Fig. 17 (a) and (c)). If the drifts of the head tracker become large, the poselet-regressor tends to provide an incorrect pelvis center location and the upper body orientation



(a) Results from test 1.

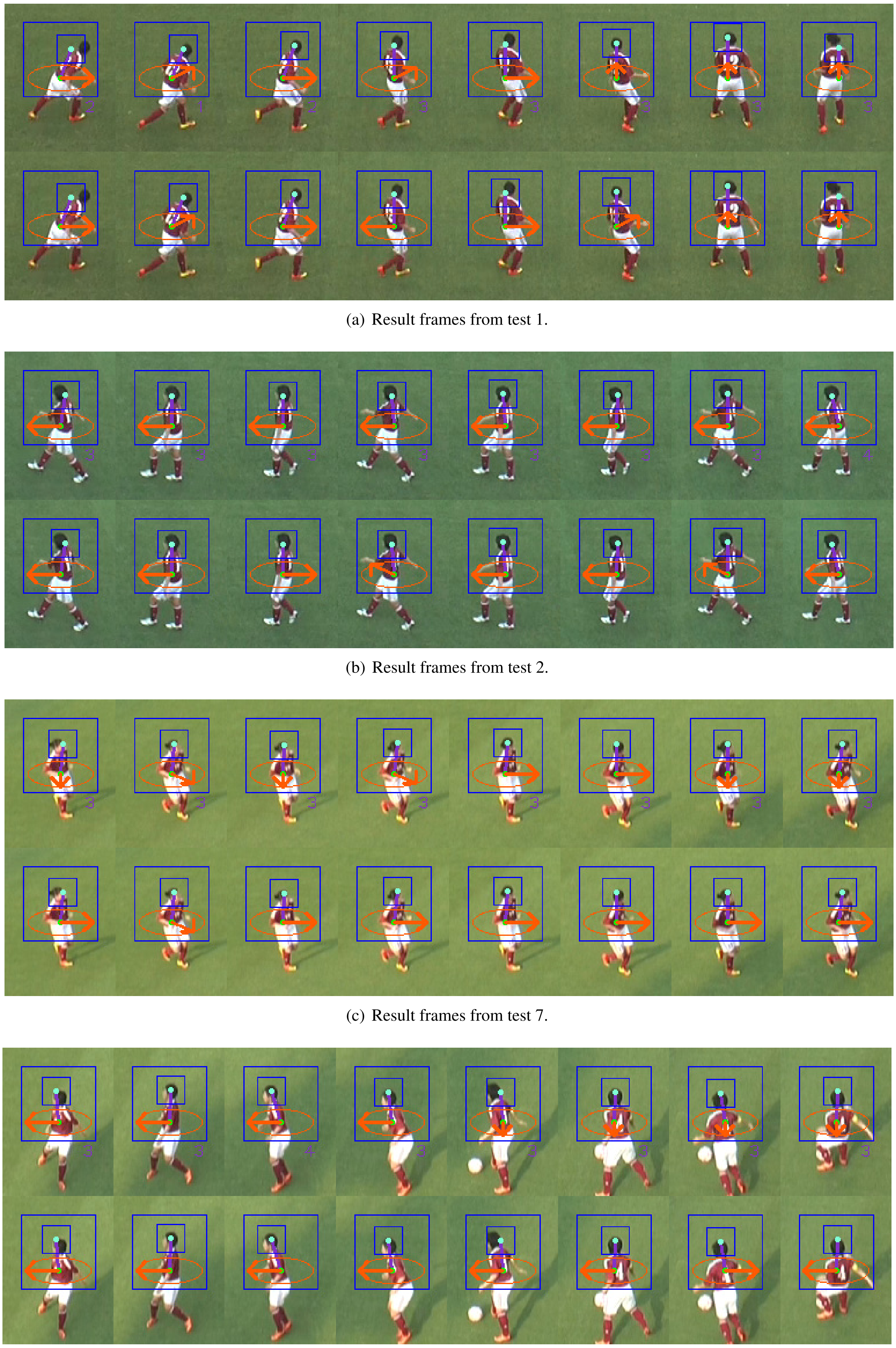
(b) Results from test 3.

(c) Results from test 4.

(d) Results from test 6.

Fig. 14 Results from the American football scenes. The first row shows the results of the proposed method and the second row shows the method of Ref. [16].

classifiers tend to fail because of the wrongly estimated spine angle class (Fig. 17 (b) and (d)). These spine pose estimations while bending from side-view monocular videos (not only for pedestrians) are novel outputs in the computer vision field, while having some errors in our experiments.



(a) Result frames from test 1.

(b) Result frames from test 2.

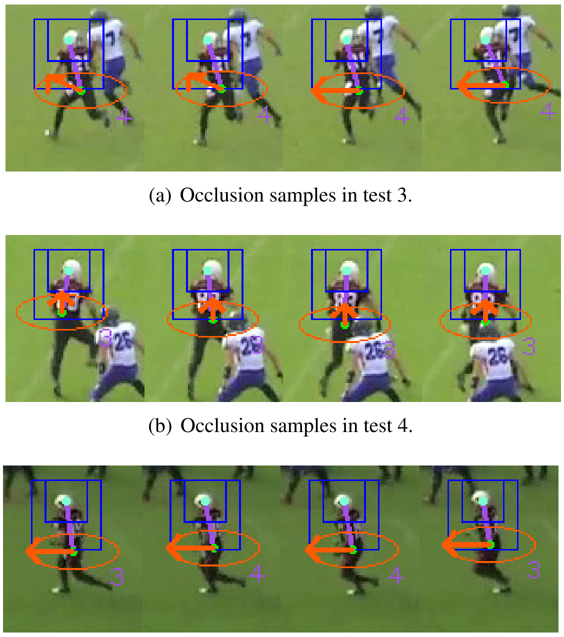
(c) Result frames from test 7.

(d) Result frames from test 10.

Fig. 15 Results from women's soccer scenes. The first row shows the results of the proposed method and the second row shows the method of Ref. [16].

The Effect of the Alignment of the Upper Body Region and the Selected Features.

Our body orientation method depends on the alignment of the tracker and the selected features in each spine angle. **Figure 18** shows some examples of the effect of the alignment of the head tracker and pelvis center estimation. In Fig. 18 (a) and (b), the

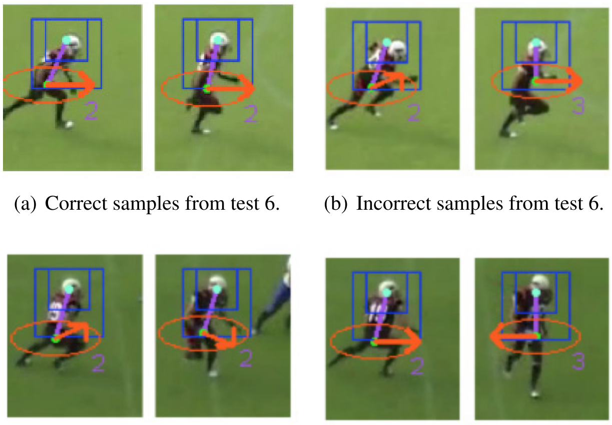


(a) Occlusion samples in test 3.

(b) Occlusion samples in test 4.

(c) Occlusion samples in test 10.

Fig. 16 Results during occlusion between players.



(a) Correct samples from test 6.

(b) Incorrect samples from test 6.

(c) Correct samples from test 12.

(d) Incorrect samples from test 12.

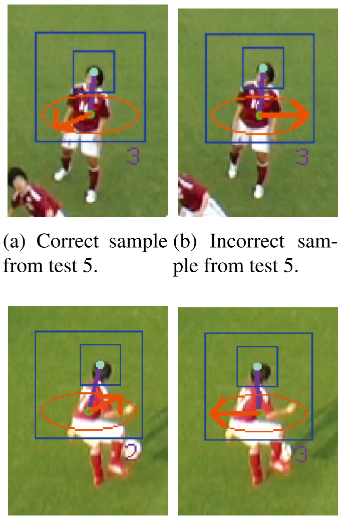
Fig. 17 Sample results of bending poses from American football tests.

head center location is not good. This misalignment causes the misclassification of the body orientation in Fig. 18 (b), while the body orientation is correct in Fig. 18 (a) because the misalignment of the head is small (for 8 × 8 pooling of HOG features for the poselet-regressor and the body orientation classifiers).

In Fig. 18 (c) and (d), the head tracker is good enough. However, in Fig. 18 (d), the pelvis estimation is not accurate and our algorithm selects the wrong spine angle class from the misaligned upper body region. There is a tradeoff between the selected features of each spine angles class vs. alignment of the upper body region and the precision of the pelvis center location.

Models for One Specific Team

Our evaluation for American football videos shows the robustness of our method mainly for running poses, which account for the vast majority of player poses not only in American football but also in all other team sports. Compared with FMP [20] or other frontal upper body estimators that must know the various types of clothing and hair styles, human pose estimators for team sports (such as our method) only need to know the appearance of the two team's uniforms for a specific match. In this sense, although our method uses fully supervised models for one specific team (or uniform), our experiments shows that our classifiers are robust enough to estimate the upper body poses of the target team players.



(a) Correct sample from test 5.

(b) Incorrect sample from test 5.

(c) Correct sample from test 8.

(d) Incorrect sample from test 8.

Fig. 18 Effect of the alignment of the upper body region for body orientation estimation.

7. Conclusion

We proposed an upper body pose estimation framework for team sports videos, which estimates the upper body orientation and the spine pose of one player from the tracked and aligned upper body appearances and feature selection with random decision forests. Our method employs a scene-specific head tracker, a spine pose regressor (poselet-regressor), and body orientation classifiers conditioned by the spine angle. Both our poselet-regressor and the conditioned body orientation classifiers are trained from the player images of the same team, and can be used for the players wearing the same uniform (or performing the same sports actions in the other scenes). Our alignment-based body orientation classification, guided by the 2D spine pose, can predict not only the body orientation but also the 2D spine pose even when hard-occlusions or part disappearance occurs, because it uses a few selected features within the aligned upper body window. This alignment-based pose estimation framework, is suitable for side view running poses as [25] and suitable for upper body bending poses which both frequently appear in team sports videos.

Moreover, our previous method [16] proposed a rough conversion of a 2D spine pose to a 3D pose by combining the 2D spine pose with horizontal body orientation recognition (Fig. 3 (b)). This means that the upper body poses estimated by the proposed method can be also used for generating (approximate) 3D upper body pose information as Ref. [16] does.

Future work includes upper body orientation estimation using team contexts such as movement directions of all players on the same team or their common attending direction. In addition, 3D geometry of the scene and players should be considered to restrict the variation of the human appearance on images. Also, we would like to utilize the spine pose information estimated from our poselet-regressor as a mid-level feature for action recognition, such as the understanding of defensive behavior from bending poses or the team activity analysis as proposed in Refs. [29], [30], [31] for surveillance.

Acknowledgments Authors would like to thank Japan American Football Association for providing the American football scene videos and also thank Keio University soccer team for providing the women soccer scene videos. Authors would like to thank the editor and reviewers' valuable comments.

References

[1] Carr, P., Sheikh, Y. and Matthews, I.: Monocular object detection using 3d geometric primitives, *Computer Vision–ECCV 2012*, pp.864–878, Springer (2012).

[2] Fleuret, F., Berclaz, J., Lengagne, R. and Fua, P.: Multicamera people tracking with a probabilistic occupancy map, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.30, No.2, pp.267–282 (2008).

[3] Atmosukarto, I., Ghanem, B., Ahuja, S., Muthuswamy, K. and Ahuja, N.: Automatic recognition of offensive team formation in american football plays, *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), pp.991–998, IEEE (2013).

[4] Bialkowski, A., Lucey, P., Carr, P., Denman, S., Matthews, I. and Sridharan, S.: Recognising team activities from noisy data, *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), pp.984–990, IEEE (2013).

[5] Lan, T., Sigal, L. and Mori, G.: Social roles in hierarchical models for human activity recognition, *2012 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.1354–1361, IEEE (2012).

[6] Wang, Z., Shi, Q., Shen, C. and van den Hengel, A.: Bilinear Programming for Human Activity Recognition with Unknown MRF Graphs, *2013 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.1690–1697, IEEE (2013).

[7] Kim, K., Grundmann, M., Shamir, A., Matthews, I., Hodgins, J. and Essa, I.: Motion fields to predict play evolution in dynamic sport scenes, *2010 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.840–847, IEEE (2010).

[8] Choi, W., Shahid, K. and Savarese, S.: What are they doing?: Collective activity classification using spatio-temporal relationship among people, *2009 IEEE 12th International Conference on Computer Vision Workshops* (*ICCV Workshops*), pp.1282–1289, IEEE (2009).

[9] Chen, C., Heili, A. and Odobez, J.-M.: Combined estimation of location and body pose in surveillance video, *Advanced Video and Signal Based Surveillance* (2011).

[10] Chen, C. and Odobez, J.: We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video, *2012 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.1544–1551, IEEE (2012).

[11] Baltieri, D., Vezzani, R. and Cucchiara, R.: People orientation recognition by mixtures of wrapped distributions on random trees, *Computer Vision–ECCV 2012*, pp.270–283, Springer (2012).

[12] Andriluka, M., Roth, S. and Schiele, B.: Monocular 3D Pose Estimation and Tracking by Detection, *IEEE Conference on Computer Vision and Pattern Recognition, 2010, CVPR 2010*, IEEE (2010).

[13] Benfold, B. and Reid, I.: Guiding Visual Surveillance by Tracking Human Attention., *BMVC*, pp.1–11 (2009).

[14] Dalal, N. and Triggs, B.: Histograms of Oriented Gradients for Human Detection, *Computer Vision and Pattern Recognition*, Vol.1, pp.886–893 (2005).

[15] Bourdev, L. and Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations, *2009 IEEE 12th International Conference on Computer Vision*, pp.1365–1372, IEEE (2009).

[16] Hayashi, M., Yamamoto, T., Ohshima, K., Tanabiki, M. and Aoki, Y.: Head and Upper Body Pose Estimation in Team Sport Videos, *2013 2nd IAPR Asian Conference on Pattern Recognition* (*ACPR*), pp.754–759, IEEE (2013).

[17] Dantone, M., Gall, J., Fanelli, G. and Van Gool, L.: Real-time facial feature detection using conditional regression forests, *2012 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.2578–2585, IEEE (2012).

[18] Sun, M., Kohli, P. and Shotton, J.: Conditional regression forests for human pose estimation, *2012 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.3394–3401, IEEE (2012).

[19] Wang, Y., Tran, D. and Liao, Z.: Learning hierarchical poselets for human parsing, *2011 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.1705–1712, IEEE (2011).

[20] Yang, Y. and Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts, *2011 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.1385–1392, IEEE (2011).

[21] Pishchulin, L., Andriluka, M., Gehler, P. and Schiele, B.: Poselet conditioned pictorial structures, *2013 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.588–595, IEEE (2013).

[22] Bourdev, L., Maji, S., Brox, T. and Malik, J.: Detecting people using mutually consistent poselet activations, *Computer Vision–ECCV 2010*, pp.168–181, Springer (2010).

[23] 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).

[24] Maji, S., Bourdev, L. and Malik, J.: Action recognition from a distributed representation of pose and appearance, *2011 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.3177–3184, IEEE (2011).

[25] Hayashi, M., Oshima, K. and M.T.Y.A.: 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).

[26] Benfold, B. and Reid, I.: Unsupervised learning of a scene-specific coarse gaze estimator, *International Conference on Computer Vision*, pp.2344–2351 (2011).

[27] Schulz, A., Damer, N., Fischer, M. and Stiefelhagen, R.: Combined Head Localization and Head Pose Estimation for Video–Based Advanced Driver Assistance Systems, *Pattern Recognition*, pp.51–60, Springer (2011).

[28] Schulz, A. and Stiefelhagen, R.: Video-based pedestrian head pose estimation for risk assessment, *2012 15th International IEEE Conference on Intelligent Transportation Systems* (*ITSC*), pp.1771–1776, IEEE (2012).

[29] Lan, T., Wang, Y., Yang, W., Robinovitch, S.N. and Mori, G.: Discriminative latent models for recognizing contextual group activities, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.34, No.8, pp.1549–1562 (2012).

[30] Chamveha, I., Sugano, Y., Sato, Y. and Sugimoto, A.: Social Group Discovery from Surveillance Videos: A Data-Driven Approach with Attention-Based Cues, BMVC (2013).

[31] Choi, W. and Savarese, S.: A unified framework for multi-target tracking and collective activity recognition, *Computer Vision–ECCV 2012*, pp.215–230, Springer (2012).

[32] Eichner, M., Marin-Jimenez, M., Zisserman, A. and Ferrari, V.: 2d articulated human pose estimation and retrieval in (almost) unconstrained still images, *International journal of computer vision*, Vol.99, No.2, pp.190–214 (2012).

[33] Bourdev, L., Maji, S. and Malik, J.: Describing People: Poselet-Based Attribute Classification, *International Conference on Computer Vision* (*ICCV*), (online), available from ⟨http://www.eecs.berkeley.edu/ lbourdev/poselets⟩ (2011).

[34] Patron-Perez, A., Marszalek, M., Reid, I. and Zisserman, A.: Structured learning of human interactions in TV shows, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.34, No.12, pp.2441–2453 (2012).

[35] Ramanathan, V., Yao, B. and Fei-Fei, L.: Social role discovery in human events, *2013 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.2475–2482, IEEE (2013).

[36] Yao, B. and Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities, *2010 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.17–24, IEEE (2010).

[37] Yao, B., Ma, J. and Fei-Fei, L.: Discovering object functionality, *Submitted to the IEEE International Conference on Computer Vision* (*ICCV*) (2013).

[38] 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).

[39] Andriluka, M., Roth, S. and Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation, *IEEE Conference on Computer Vision and Pattern Recognition, 2009, CVPR 2009*, pp.1014–1021, IEEE (2009).

[40] Felzenszwalb, P.F., Girshick, R.B., McAllester, D.A. and Ramanan, D.: Object Detection with Discriminatively Trained Part-Based Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.32, pp.1627–1645 (2010).

[41] Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M. and Moore, R.: Real-time human pose recogni-

tion in parts from single depth images, *Comm. ACM*, Vol.56, No.1, pp.116–124 (2013).

[42] Kazemi, V., Burenius, M., Azizpour, H. and Sullivan, J.: Multiview body part recognition with random forests, *British Machine Vision Conference* (2013).

[43] Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N. and Ilic, S.: 3d pictorial structures for multiple human pose estimation, *CVPR, IEEE* (2014).

[44] Raptis, M. and Sigal, L.: Poselet key-framing: A model for human activity recognition, *2013 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.2650–2657, IEEE, (2013).

[45] Ek, C.H., Torr, P.H. and Lawrence, N.D.: Gaussian process latent variable models for human pose estimation, *Machine learning for multimodal interaction*, pp.132–143, Springer (2008).

[46] Jaeggli, T., Koller-Meier, E. and Van Gool, L.: Learning generative models for multi-activity body pose estimation, *International Journal of Computer Vision*, Vol.83, No.2, pp.121–134 (2009).

[47] Urtasun, R., Fleet, D.J. and Fua, P.: 3D people tracking with Gaussian process dynamical models, *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol.1, pp.238–245, IEEE (2006).

[48] Andriluka, M., Roth, S. and Schiele, B.: People-tracking-by-detection and people-detection-by-tracking, *IEEE Conference on Computer Vision and Pattern Recognition, 2008, CVPR 2008*, pp.1–8, IEEE (2008).

[49] Han, J. and Bhanu, B.: Individual recognition using gait energy image, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.28, No.2, pp.316–322 (2006).

[50] Sugano, Y., Matsushita, Y. and Sato, Y.: Learning-by-synthesis for appearance-based 3d gaze estimation, *2014 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.1821–1828, IEEE (2014).

[51] Dollár, P., Welinder, P. and Perona, P.: Cascaded pose regression, *2010 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.1078–1085, IEEE (2010).

[52] Benfold, B. and Reid, I.: Colour Invariant Head Pose Classification in Low Resolution Video., *BMVC*, pp.1–10 (2008).

[53] Chamveha, I., Sugano, Y., Sugimura, D., Siriteerakul, T., Okabe, T., Sato, Y. and Sugimoto, A.: Appearance-based head pose estimation with scene-specific adaptation, *Proc. IEEE International Workshop on Visual Surveillance* (*VS2011*), pp.1713–1720 (Nov. 2011).

[54] Chen, C., Heili, A. and Odobez, J.-M.: A joint estimation of head and body orientation cues in surveillance video, *2011 IEEE International Conference on Computer Vision Workshops* (*ICCV Workshops*), pp.860–867, IEEE (2011).

[55] Breiman, L.: Random forests, *Machine learning*, Vol.45, No.1, pp.5–32 (2001).

Masaki Hayashi received his M.S. degree in Computer Vision and Image Processing from Keio University in 2006. Since 2012, he has been a Ph.D. candidate at Keio University. His research interests are video-based human pose estimation, activity recognition, and attention recognition, especially for team sports videos.

Kyoko Oshima received her B.A. of Liberal Arts from International Christian University in 1992. She is a staff engineer in Panasonic Corporation and works on research and development of computer vision systems.

Masamoto Tanabiki received his M.S. degree in electrical engineering from Waseda University in 1998. His research interests are video-based human tracking and pose estimation, especially for security, business intelligence, and sports.

Yoshimitsu Aoki is an Associate Professor, Department of Electronics & Electrical Engineering, Keio University. He received his Ph.D. in Engineering from Waseda University in 2001. From 2002 to 2008, he was an associate professor in Shibaura Institute of Technology. Since 2008, he has been an associate professor at Department of Electronics & Electrical Engineering in Keio University. He performs research in the areas of Computer Vision, Pattern Recognition, and Media Sensing/Understanding.

(Communicated by *Slobodan Ilic*)