Inverse Dynamics for Action Recognition

Al Mansur, Yasushi Makihara, and Yasushi Yagi

***Abstract*—Pose-based approaches for human action recognition are attractive owing to their accurate use of human motion information. Traditionally, such approaches used kinematic features for classification. However, in addition to having high dimensions and a small interclass variation, kinematic features do not consider the interaction of the environment on human motion. In this paper, we propose a method for action recognition using dynamic features, derived by applying inverse dynamics to a physics-based representation of the human body. The physics-based model is articulated and actuated with muscles and consists of joints with variable stiffness. Dynamic features under consideration include the torques from the knee and hip joints of both legs and, implicitly, gravity, ground reaction forces, and the pose of the remaining body parts. These features are more discriminative than kinematic features, resulting in a low-dimensional representation for human actions, which preserves much of the information of the original high-dimensional pose. This low-dimensional feature achieves good classification performance even with a relatively small training data set in a simple classification framework such as a hidden Markov model. The effectiveness of the proposed method is demonstrated through experiments on the Carnegie Mellon University motion capture data set and Osaka University Kinect action data set with various actions.**

***Index Terms*—Action recognition, dynamics feature, hidden Markov model (HMM), physics-based model.**

I. Introduction

THE recognition and interpretation of human actions and activities have sparked considerable interest in the computer vision research community, owing to the large number of potential applications in video retrieval, surveillance, human–computer interaction, robot learning and control, and imitation learning.

Depending on the way features are extracted, human action recognition approaches are roughly classified into two types, namely, appearance- and pose-based approaches.

In appearance-based approaches, features commonly used for action recognition include shape [1], [2], optical flow [3], [4], point trajectory [5], [6], and joint angle [7]. Most of these features encode the kinematics of human motion, and in general, they are high dimensional. Recognizing human actions from a video is a challenging task [8]. First, it is difficult to identify an action independently of the viewing direction. Second, extracting stable features is complicated owing to noise or the weakness of the feature extraction method itself.

On the other hand, pose-based approaches [9]–[16] usually employ the joint angles, point trajectories, or other motion information of the human body and its parts to model the dynamics of an action, thereby overcoming many of the limitations of appearance-based approaches. However, accuracy of these methods largely depends on the accuracy of the human pose tracking. Tracking all the body parts accurately is a nontrivial task. Compared with the upper body parts, tracking the lower body parts (e.g., legs) is much easier as several constraints can be imposed (such as ground contact and limited degrees-of-freedom) for common human actions. Recently, tracking using low-cost depth-measuring devices such as Kinect (TM) has become popular [17]–[19]. The accuracy of the tracking results obtained by these approaches is adequate for human action recognition, and therefore, pose-based approaches have become easier to deploy.

Traditional pose-based approaches use kinematic features only, making it very difficult to distinguish certain actions. In general, for many actions, kinematic features are high dimensional with a small interclass variation. Moreover, they do not consider the interaction of the environment on human motion. Recently, physics-based models have been successfully used for 3-D people tracking [20], [21]. These physics-based models provide parameterizations for effective modeling of plausible poses and motions. In addition, they are capable of capturing the influence of gravity, ground contact, and other physical interactions with the environment on pose and motion. In a related work [22], internal joint torques and external forces were recovered from the observed motion in a different problem setting.

Inspired by these works, we attempt using dynamic features obtained from a physics-based model of the human body for human action recognition. Use of dynamics allows us to represent an action by the torques and forces governing the human motion. We believe that features of this kind will result in a more discriminative feature.

The main contribution of this paper is that we represent actions using dynamic features, i.e., joint torques, which have several advantages over kinematic features. First, these features are more discriminative than kinematic ones. Second, they provide a low-dimensional representation of the actions, which is necessary in many cases to deal with the limited number of training data. We consider only the lower body torques for

Manuscript received March 1, 2012; revised August 17, 2012; accepted October 16, 2012. Date of publication December 10, 2012; date of current version July 15, 2013. This work was supported in part by the JSPS Grant-in-Aid for Scientific Research (S) under Grant 21220003 and in part by the "R&D Program for Implementation of Anti-Crime and Anti-Terrorism Technologies for a Safe and Secure Society," Strategic Funds for the Promotion of Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government. This paper was recommended by Associate Editor L. Shao.

The authors are with the Department of Intelligent Media, Institute of Scientific and Industrial Research, Osaka University, Ibaraki 567-0047, Japan (e-mail: mansur@am.sanken.osaka-u.ac.jp; makihara@am.sanken.osaka-u.ac.jp; yagi@am.sanken.osaka-u.ac.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCB.2012.2226879

action representation as these implicitly include configurations of the other limbs of the body and external forces acting on the body such as gravity and the ground reaction force (GRF). In most actions, the human body remains in an upright position with the legs supporting the body. As a result, the upper body parts have a greater influence on the dynamics of the lower body parts, whereas the lower body parts usually have little influence on the dynamics of the upper body segment. As such, we chose to use the lower body torques for action recognition. Due to the low-dimensional representation, we can use a simple learning and classification framework, e.g., the standard hidden Markov model (HMM) to achieve good classification performance with a small number of training data. Our experimental results using the Carnegie Mellon University (CMU) motion capture data set and the Osaka University Kinect action data set reveal that dynamic features provide good discrimination over different actions.

For joint torques computation, we used a generic inverse dynamics approach formulated in Lagrangian equations of motion. This is a common practice used in inverse dynamics of multibody systems. Ref. [23] uses the same approach for creating realistic human motion based on space–time optimization that minimizes the total muscle torques according to an actor's relative preference. We used a similar biomechanical model as used in [23]. However, another valid model with different parameters (e.g., [21]) should work in our method as long as the computed torques show the discriminativeness among different actions. In addition, we admit that HMM-based action classification is not new. However, we used this framework to show that the dynamic features are effective even when a simple classification scheme such as HMM is used.

This paper builds on a previous work [24], and we have expanded the database, techniques, and experimental results to show the effectiveness of the proposed method. In particular, we have included passive joint parameters (springs and dampers) in the articulated model to make the computed torques more realistic. In addition, we present evaluation results on a motion data set captured by a simple and inexpensive setup consisting of a Kinect (TM) and an RGB camera.

The organization of this paper is as follows. Following the literature review on human action recognition in Section II, Section III introduces the dynamic feature extraction process. Sections IV and V present the HMM-based classification framework and the experimental results, respectively, for two different data sets. Finally, Section VI concludes this paper and discusses future work.

II. Related Work

Human action and activity recognition has been an active area of research in the field of computer vision for more than two decades. As such, a large amount of literature exists. Comprehensive reviews of these works are presented in several survey papers [8], [25]. In this section, we limit our discussion to some of the papers most closely related to our research. Although many different approaches exist, action recognition algorithms can be mainly categorized into two categories, namely, appearance- and pose-based approaches.

A. Appearance-Based Approaches

Appearance-based approaches learn the appearance model of the human body or some of its body parts and try to match this to images in a test scene for action or gesture recognition. We can further categorize appearance-based approaches for human action recognition according to the action representation. Some of the most popular representations include learned geometrical models of human body parts, spatiotemporal templates, appearance or region features, shape information, interest-point-based representations, and motion/optical flow patterns. In the following paragraphs, we provide a brief summary of the related work.

In [26] and [27], using appearance-based features, actions are learnt by an HMM or other variants thereof. The appearance-based approach works well for gesture recognition since the overall appearance of the human hand does not change much in different people. However, these methods do not work well for whole body actions as they cannot handle the problem brought about by changes in clothing or appearance.

Shape-based representations utilize silhouettes of the human body [1] or the features derived from these silhouettes [2]. The basic idea behind shape-based representations is that an action consists of a sequence of poses that can be detected in a single frame. Usually, recognition is based on a single frame; however, to improve robustness, it may be extended to multiple frames. In [28], a silhouette-based representation was enhanced to characterize the outline of the human body in the space–time domain. This resulted in a spatiotemporal 3-D volume, constructed by stacking the silhouettes detected in each frame. Shape-based approaches work well on a number of actions. However, they also suffer from the problem of silhouette variations owing to clothing changes or imperfect segmentation.

Another class of works uses volumetric analysis of the video for action recognition [28]–[30]. Yilmaz and Shah [29] used spatiotemporal features to exploit both shape and motion features simultaneously, whereas Blank *et al.* [28] extended a method developed for analysis of 2-D shapes to deal with volumetric spatiotemporal shapes induced by human actions. The main advantage of the volume-based approach is that it is not necessary to build complex models of body configuration and kinematics. In addition, recognition can be directly performed using raw video.

Recently, interest-point-based representations have attracted considerable interest in action recognition research. They use spatiotemporal interest points and their trajectories for action and activity analysis [31]–[36]. The main strength of this representation is its robustness to occlusion since there is no need to track or detect the whole human body. In [37], performance of the space–time features was evaluated in a common experimental setup and concluded that for human action recognition, performance of regular sampling of space–time features is better than the other tested space–time interest point detectors. In another evaluation work [38], performance and computational efficiency of several part-based approaches of feature detection and feature description methods were evaluated in a human action recognition scenario. Using KTH action data set, it was found that the feature detection method in [33] combined with the improved LBP-TOP descriptor achieved the best recognition accuracy with low computational cost.

A number of researchers [3], [4], [33], [39]–[41] have used features based on motion and optical flow for action recognition. Bobick and Davis [39] proposed the motion energy image (MEI) to describe the cumulative spatial distribution of motion energy in a given sequence. Later, the idea of the MEI was extended to a motion history image (MHI) [42]. In optical flow-based approaches [3], [4], optical flow is used to derive a representation that is suitable for recognition. It was shown in [42] that MEI and MHI yield good discriminative performance for some specific simple action classes. However, it was reported in [40] that MEI and MHI show unsatisfactory performance for complex actions owing to overwriting the motion history.

B. Pose-Based Approaches

Pose-based approaches [9]–[14] usually rely on human body pose tracking to model the dynamics of body parts and exploit these models for action recognition. Sheikh *et al.* [9] expressed an action as a linear combination of spatiotemporal actions and proposed a framework for learning the variability of the execution of human actions that is unaffected by the changes. Using both motion capture and video data, Ikizler and Forsyth [10] proposed a generative model to query complex activities in a large collection of videos. Instead of inferring a 3-D pose in each frame of an action sequence, Lv and Nevatia [11] searched for a series of actions that best matched the input sequence and proposed an action representation scheme called Action Net, which inherently models the contextual constraints for action recognition. Embedding low-level shape and optical flow features into a high-level graphical model representation, Natarajan and Nevatia [12] presented an approach for simultaneous tracking and event recognition. Fanti *et al.* [13] proposed a hybrid probabilistic model for human motion recognition that combines global features (e.g., translation) with local variables (e.g., relative positions and appearance of body parts). Yilmaz and Shah [14] extended the standard epipolar geometry to the geometry of dynamic scenes for recognizing human actions in videos acquired by uncalibrated moving cameras. In a different problem setting, Kilner *et al.* [43] proposed a technique for automatic matching of human activities in outdoor sports broadcast environments. This technique is based on the analysis of recorded 3-D data of human activity and retrieving the most appropriate proxy action from a motion capture library. However, these approaches use kinematic information only to model human actions, which is not discriminative enough for many actions.

Only a few approaches employ sensors in feature extraction for action recognition. In these sensor-based approaches, human action is usually described by the output signal obtained from sensors either attached to the human body or installed in the environment (excluding cameras and depth sensors). In [44], an effective subject recognition approach is designed using GRF measurements of human gait. In [45], Yang *et al.* proposed an action recognition system using wearable motion sensor networks. In addition, there are several works that use acceleration data obtained from sensors [46] for action recognition. The main disadvantage of the sensor-based methods is that subjects either need to wear certain specialized devices or sensors or they have to perform the action in a specialized environment.

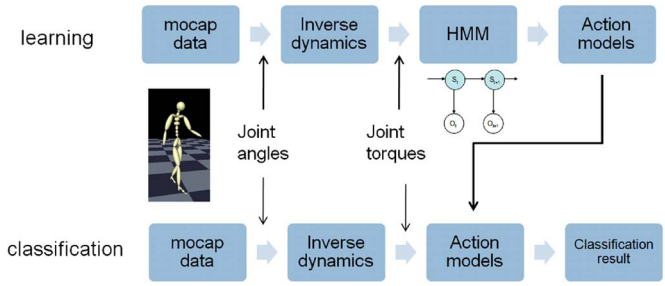


Fig. 1. Overview of our approach.

In summary, our proposed framework differs in three ways from previous representations. First, from the human motion, we compute the joint torques using a physics model. These features are more discriminative than kinematic features, and they can distinguish those actions that are very similar in terms of kinematic features. Second, information on ground interaction is encoded in the torques since the computation of the torques depends on external forces. Third, we use a low-dimensional representation of human actions, which allows us to achieve good classification performance with a relatively small training data set in a simple classification framework such as an HMM.

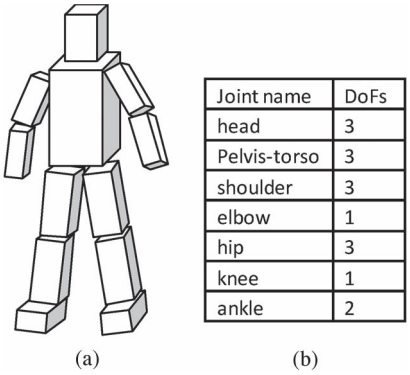
III. Our Approach

A. Overview

The overview of our approach is shown in Fig. 1. The approach is divided into two parts: 1) learning; and 2) classification. In the training part, motion capture data from several persons performing each action were used as input data. Motion capture data contain joint angle trajectories of various joints of the articulated human body. Using these joint angle trajectories, we compute torques at different joints of the body using inverse dynamics (see Section III-C). These joint torques are used in an HMM framework to learn the action models for each action. Then, in the classification part, we are given motion capture data from an unknown action, and our task is to predict the correct label. To do this, we again use inverse dynamics and compute joint torque corresponding to the given motion data. Now, we can use the previously learnt action models to predict the class label of the input motion.

B. Kinematic Model

Our 3-D articulated human body model (see Fig. 2) consists of 12 rigid body segments with a total of 26 degrees of freedom (DOF). We use three kinds of joints to link the segments to their parent segments: 1-DOF (hinge), 2-DOF (saddle), or 3-DOF (ball and socket) rotational joints. The position and orientation of the root segment are defined in world coordinates by a 6-DOF global joint. All segments are approximated by sticks of appropriate lengths. We use identical lengths of body segments, mass, and inertial parameters [47] for all subjects and actions. We estimate the internal joint torques by applying 3-D motion



Joint name	DoFs
head	3
Pelvis-torso	3
shoulder	3
elbow	1
hip	3
knee	1
ankle	2

(a) (b)

Fig. 2. (a) Kinematic model. (b) DOF of each part.

capture data to the model using the method described in the next section. Note that, since we are not dealing with a tracking problem in this paper, we assume that the pose tracking data are available to us. This is a realistic assumption, as current markerless tracking algorithms can provide accurate tracking results [17], [19], [48], [49].

C. Recovery of Torques Using Inverse Dynamics

The Lagrangian equations of motion, which include the effect of generalized forces associated with a DOF q_j, can be written as [23]

$$\sum_{i \in N(j)} \frac{d}{dt}\frac{\partial T_i}{\partial \dot{q}_j} - \frac{\partial T_i}{\partial q_j} = Q_j \tag{1}$$

where T_i is the kinetic energy of body segment i, $N(j)$ is the set of body segments in the subtree of a joint's DOF q_j, and Q_j is the total generalized force acting on q_j. The kinetic energy of body segment i can be found as

$$T_i = \frac{1}{2} tr\left(\dot{\boldsymbol{W}}_i \boldsymbol{M}_i \dot{\boldsymbol{W}}_i^T\right) \tag{2}$$

where $\boldsymbol{W}_i$ is the complete transformation from the root of the skeleton to segment i, and $\boldsymbol{M}_i$ is the mass tensor of body segment i. Now, the terms on the left-hand side of (1) can be written as

$$\frac{d}{dt}\frac{\partial T_i}{\partial \dot{q}_j} - \frac{\partial T_i}{\partial q_j} = tr\left(\frac{\partial \boldsymbol{W}_i}{\partial q_j}\boldsymbol{M}_i \ddot{\boldsymbol{W}}_i^T\right). \tag{3}$$

The total generalized force Q_j acting on DOF q_j is the sum of various component forces and is given as

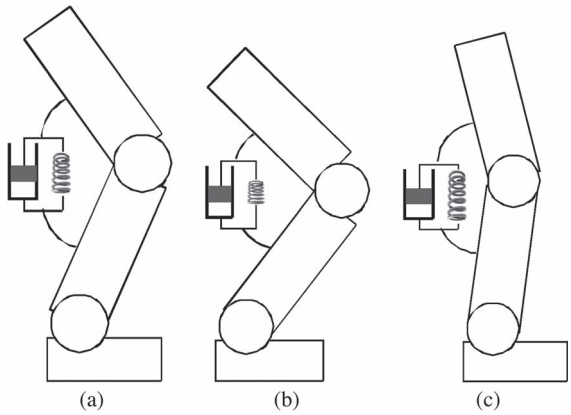
$$Q_j = Q_{m_j} + Q_{g_j} + Q_{p_j} + Q_{r_j} \tag{4}$$

where Q_{m_j}, Q_{g_j}, Q_{p_j}, and Q_{r_j} represent the generalized force due to muscles, gravity, passive element in the joints, and ground reaction, respectively. These component forces are computed as follows.

D. Gravity

As a result of gravity, a constant force $m_i \boldsymbol{g}$ acts on the center of the mass of each body part i. The generalized force at joint DOF q_j due to the effect of gravity is computed as [23]

$$Q_{g_j} = \sum_{i \in N(j)} \left(\frac{\partial \boldsymbol{W}_i}{\partial q_j}\boldsymbol{c}_i\right) \cdot (m_i \boldsymbol{g}) \tag{5}$$



(a) (b) (c)

Fig. 3. Spring constant for different states of a joint: (a) at rest, $Q_p = 0$; (b) flexion, $Q_p = -k_{s1}(q - \bar{q}) - k_d(\dot{q})$; and (c) extension, $Q_p = -k_{s2}(q - \bar{q}) - k_d(\dot{q})$.

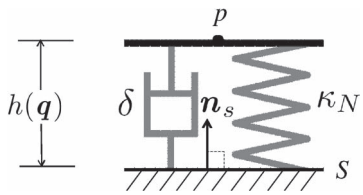


Fig. 4. Model for computing GRF.

where $\boldsymbol{c}_i$ is the center of body segment i (expressed in local coordinates), m_i is the mass of body segment i, and $\boldsymbol{g}$ is the acceleration due to gravity.

E. Joint Forces

Due to stretching of opposing muscles, tendons, and ligaments, passive joint forces are developed at different joints of the articulated human body model. Tendons are a kind of stretchy tissue that connects muscles to bones, whereas ligaments are a kind of fibrous tissue that joins adjacent bones across a joint to keep the joint in place. The combined effect of these muscles, tendons, and ligaments can be modeled as a spring and damper combination. The resulting force due to these passive elements can be described as [23]

$$Q_{p_j} = -k_{s_j}(q_j - \bar{q}_j) - k_{d_j}\dot{q}_j \tag{6}$$

where k_{s_j}, k_{d_j}, and $\bar{q}_j$ are the spring constant, the damping constant, and the resting angle of the joint, respectively. Each joint is characterized by two different spring constants k_{s1_j} and k_{s2_j}: one for flexion and the other for extension (see Fig. 3).

F. GRF

Since the estimates of the internal torques strongly depend on the external forces, we need to consider these forces to recover the joint torques. We consider foot contact only, i.e., GRF is created only when the foot is in contact with the ground. We model the foot ground contact with a damped linear spring modulated by two sigmoidal functions [22] (see Fig. 4). One sigmoid is used to prevent forces from being applied when point p

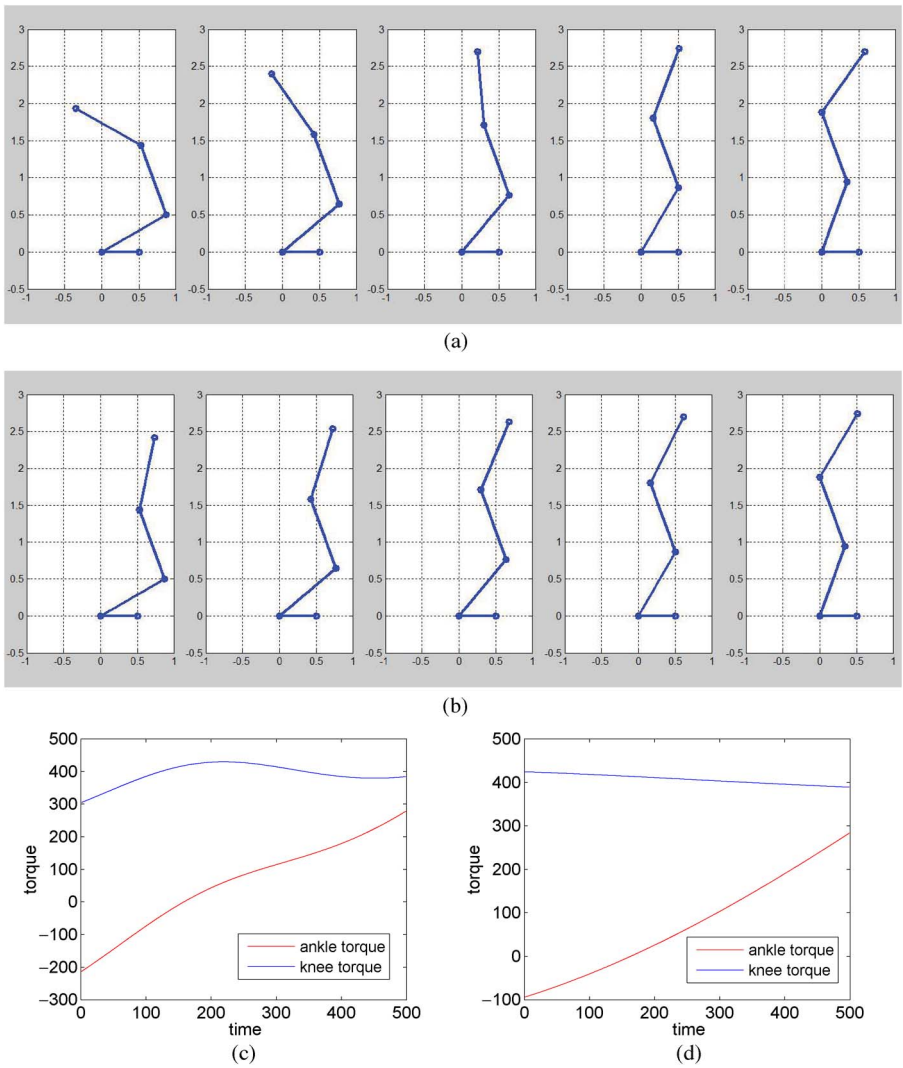


Fig. 5. Knee and ankle torques with the same lower body but different upper body motions. (a) Action 1. (b) Action 2. (c) Developed torques in action 1. (d) Developed torques in action 2.

(on the foot) is far from surface S, whereas the other sigmoid prevents forces that pull the foot when it moves away from the ground. The GRF acting on p is given by

$$Q_{r_j} = h(-60h(\boldsymbol{q}))h(5n_c)n_c\frac{\partial h(\boldsymbol{q})}{\partial q_j} \tag{7}$$

where $h(x) = (1/2)(1 + \tanh(x))$ is the sigmoidal function, and n_c is the normal spring force. $\partial h(\boldsymbol{q})/\partial q_j$ projects the normal force into the space of q_j. n_c is given by

$$n_c = -\kappa_N\left(h(\boldsymbol{q}) - \bar{h}\right) - \delta\dot{p}^T\boldsymbol{n}_s \tag{8}$$

where κ_N is the stiffness of the spring, $\bar{h}$ is the resting length of the spring, δ is the damping coefficient of the damper, and $\boldsymbol{n}_s$ is the unit normal of S at the point on S closest to p. The first term denotes the force due to the spring, whereas the second term denotes that due to the damper, which depends on the velocity of point p. Given a motion vector $\boldsymbol{X}$ with joint angle configurations $\boldsymbol{q}$ and ground contact forces, the joint torques (muscle forces) Q_{m_j} at time t can be computed from (1) and (4) as a function of the motion $\boldsymbol{X}$

$$Q_{m_j}(t, \boldsymbol{X}) = \sum_{i\in N(j)}\frac{d}{dt}\frac{\partial T_i}{\partial\dot{q}_j} - \frac{\partial T_i}{\partial q_j} - Q_{g_j} - Q_{p_j} - Q_{r_j}. \tag{9}$$

These joint torques are used as features in our proposed action recognition framework.

G. Low-Dimensional Representation

Torques computed on the lower body joints (e.g., knees and hips) depend on the pose or motion of the upper body limbs. Fig. 5 demonstrates this phenomenon. In Fig. 5(a) and (b), the human body is depicted stretching up from an almost sitting to an almost standing position. Although knee and hip motions are the same in both cases, the motion and initial posture of the upper body are different in each case. The upper body from the hip is represented by an inverted pendulum connected to the hip. The lumped mass is located as the center of mass of the upper body. In each case, the computed torques considerably

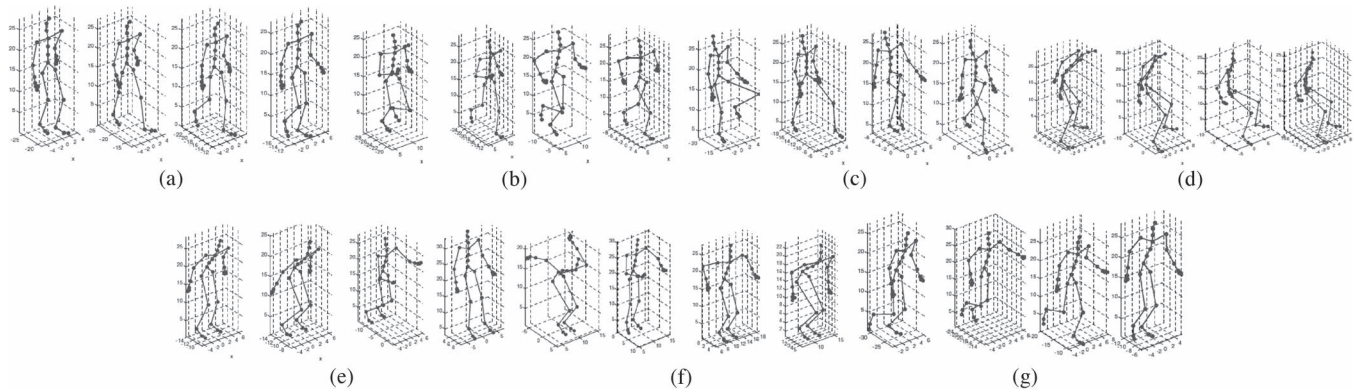


Fig. 6. Action samples from the CMU data set. (a) Walk. (b) Run. (c) March. (d) Sit. (e) Jump forward. (f) Jump in place. (g) Hopping.

differ [see Fig. 5(c) and (d)]. If these two motions represent two different actions, we can distinguish them with the help of lower body torques. Thus, we obtain a discriminative low-dimensional representation of the actions, which is the advantage of using dynamic features over kinematic features.

IV. Classification Framework

An HMM is a good probabilistic framework for modeling the dynamics of human action [26]. It can deal with sequential data and handle timescale changes in the data during recognition. In our HMM configuration, we have one hidden state variable S, and each state of this variable can emit a vector-valued observation $\boldsymbol{O}$ (torque) according to the continuous output probabilities, i.e., $B = \{b_i(\boldsymbol{O}_t)\} = \{P(\boldsymbol{O}_t|S_t = i)\}$, for $i = 1, \ldots, N$, where N is the number of states. Output density function $b_i(\boldsymbol{O}_t)$ is given by a mixture of multivariate Gaussian (vector observation) distributions

$$b_i(\boldsymbol{O}_t) = \sum_{m=1}^{M} c_{im}\mathcal{N}(\boldsymbol{O}_t; \boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im}) \tag{10}$$

where M is the number of mixture components in the Gaussian mixture; $\mathcal{N}(\cdots; \boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im})$ is a multivariate Gaussian distribution for the mth component whose mean and covariance matrix are $\boldsymbol{\mu}_{im}$ and $\boldsymbol{\Sigma}_{im}$, respectively; and c_{im} is the weight of the mth component, subject to $\sum_{m=1}^{M} c_{im} = 1$.

This model is based on two dependence assumptions: 1) hidden variable S_t at time t depends only on the hidden variable in the previous time step S_{t-1}; and 2) observation variable $\boldsymbol{O}_t$ at time t depends only on S_t. During the training step, the model learns the parameter set $\lambda = \{A, B, \pi\}$ consisting of prior probabilities $\pi = \{\pi_i\} = \{P(S_0 = i)\}$, transition probabilities $A = \{a_{ij}\} = \{P(S_t = j|S_{t-1} = i)\}$, and the observation probabilities $B = \{b_i(\boldsymbol{O}_t)\} = \{P(\boldsymbol{O}_t|S_t = i)\}$ for all states using the available observed data. Learning the parameters of these distributions means optimizing the model parameters (A, B, π) and, thus, maximizing the joint probability $P(\boldsymbol{O}, S)$. For each action class k, we learn a separate HMM model λ_k, $k = 1, 2, \ldots, C$. The expectation maximization algorithm is used for these estimations.

During recognition, for a classifier with C classes, we choose the model that best matches the observations from the C HMMs $\lambda_k = \{A_k, B_k, \pi_k\}$, $k = 1, \ldots, C$. In other words, given a test observation sequence $\boldsymbol{O}_{1:T}$, we select the class label as

$$c = \arg\max_k P(\lambda_k|\boldsymbol{O}_{1:T}) \tag{11}$$

where $P(\lambda_k|\boldsymbol{O}_{1:T})$ is the probability that the observed sequence is generated by λ_k, which is recursively computed using the Viterbi algorithm. Using low-dimensional torque and joint angle trajectories, we can successfully learn good models from a relatively small training data set with a simple HMM framework.

V. Experiments

A. Data Set

Experiments were performed on two different data sets containing 3-D motion capture sequences. The first was CMU mocap data set [50], whereas the second was Osaka University Kinect action data set created in our laboratory using a Kinect (TM).

CMU Data Set: Fig. 6 shows some typical sequences of seven actions from this data set. In total, we used 171 sequences of seven action classes, namely, walk, march, run, sit, jump forward, jump in place, and hop with 27, 21, 25, 21, 25, 32, and 20 instances, respectively. All seven classes have significant intraclass variations in terms of speed and style. In addition, for some actions, the interclass variation is very low. For example, the joint angle trajectories for *walk–march* and *jump in place–hop* pairs are quite similar for some sequences. *Walk*, *march*, and *run* classes have variations in terms of speed, stride length, bounce, and arm swing. In the sit class, the subjects move their legs randomly after sitting on a stool. This results in uncorrelated knee joint angles. In the case of *jump forward* and *jump in place* sequences, the knee and hip joint angles have good similarity, and sometimes, these actions are difficult to differentiate if translational motion of the body is not considered. In summary, all the action classes contain significant intraclass variations, and therefore, this is a very challenging data set. For evaluation, we used about 2/3 of the samples for training, with the remaining samples for testing.

Osaka University Kinect Action Data Set: This data set (available at [51]) was constructed in our laboratory using a Kinect (TM) and tracking on the depth sequence. Pose tracking

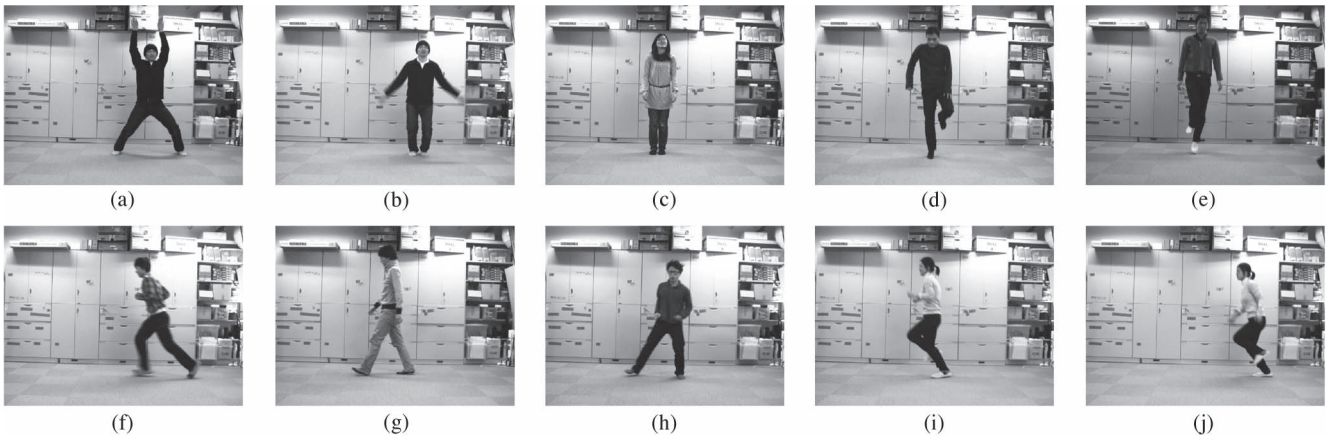


Fig. 7. Sample images from the Osaka University Kinect action data set. (a) Jack 1. (b) Jack 2. (c) Jump both legs. (d) Jump right leg. (e) Jump left leg. (f) Run. (g) Walk. (h) Side jump. (i) Skip left leg. (j) Skip right leg.

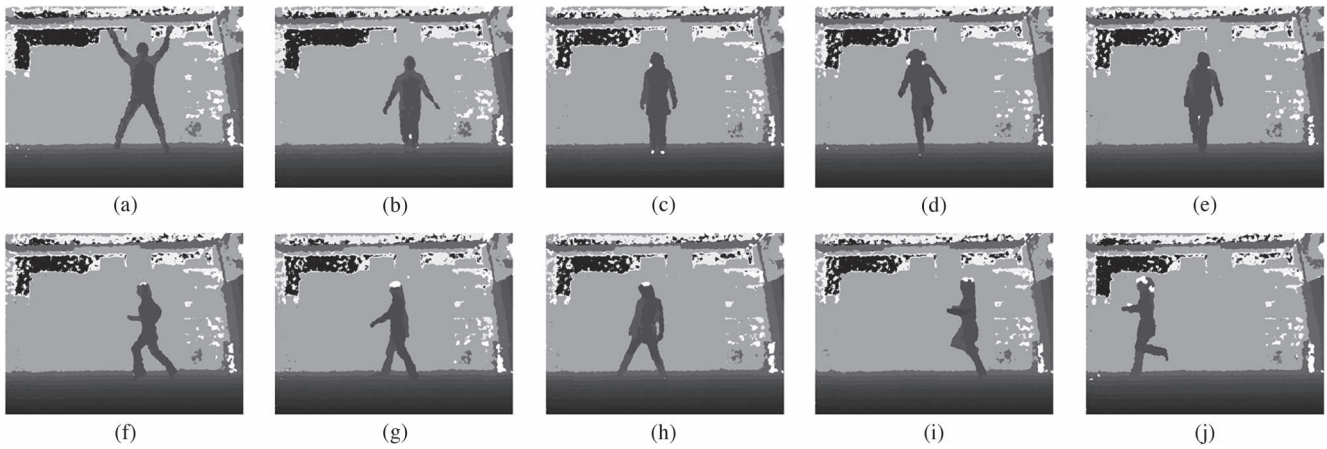


Fig. 8. Sample depth images from the Osaka University Kinect action data set. (a) Jack 1. (b) Jack 2. (c) Jump both legs. (d) Jump right leg. (e) Jump left leg. (f) Run. (g) Walk. (h) Side jump. (i) Skip left leg. (j) Skip right leg.

from the depth sequence was accomplished by a commercial motion capture software application [52]. This data set consists of 10 action classes with each action performed by eight subjects. These actions were recorded with a single Kinect (TM) and an RGB camera with frame size of 640 $\times$ 480 pixels. The frame rates for the Kinect (TM) and the RGB camera were 30 and 60 frames/s, respectively. The setup for motion capture was very simple and did not require any specialized environment or marker attachment. Examples of the RGB and depth sequences are shown in Figs. 7 and 8, respectively. Pose tracking obtained from the depth image sequences is used by the proposed method and the kinematics-based baseline method. On the other hand, the RGB image sequences are used by appearance-based baseline methods [53], [54]. The background and illumination conditions remain unchanged for all actions, thereby ensuring the fairness of the data set in application to appearance-based methods. Action types recorded are *jumping jack type 1*, *jumping jack type 2*, *jumping on both legs*, *jumping on right leg*, *jumping on left leg*, *running*, *walking*, *side jumps*, *skipping on left leg*, and *skipping on right leg*. These actions are very challenging owing to the small intraclass variations. We used a leave-one-subject-out cross-validation setting, with action sequences from seven subjects used for training and sequences of the remaining subjects used for testing.

B. Baseline Methods

We compared the results of the dynamic feature-based approach using three different baseline methods. The first method used joint angle features and the same HMM classifier. For this baseline method, we used two different feature sets for the experiments with the CMU data set. The first set comprised the six principal components obtained by reducing the original 26 dimensional joint angles by principal component analysis (PCA). More than 98% of the variance was accounted for by these six principal components. For the second set, we manually selected two knee joint angles and two hip joint angles. These particular joint angles have good consistency for a particular action class, whereas the other joint angles are noisy. Dimension reduction was necessary to avoid difficulty in learning a good model from a relatively small training data set. On the other hand, eight dimensional joint angle trajectories from both legs were used for the experiments with the Osaka University Kinect action data set.

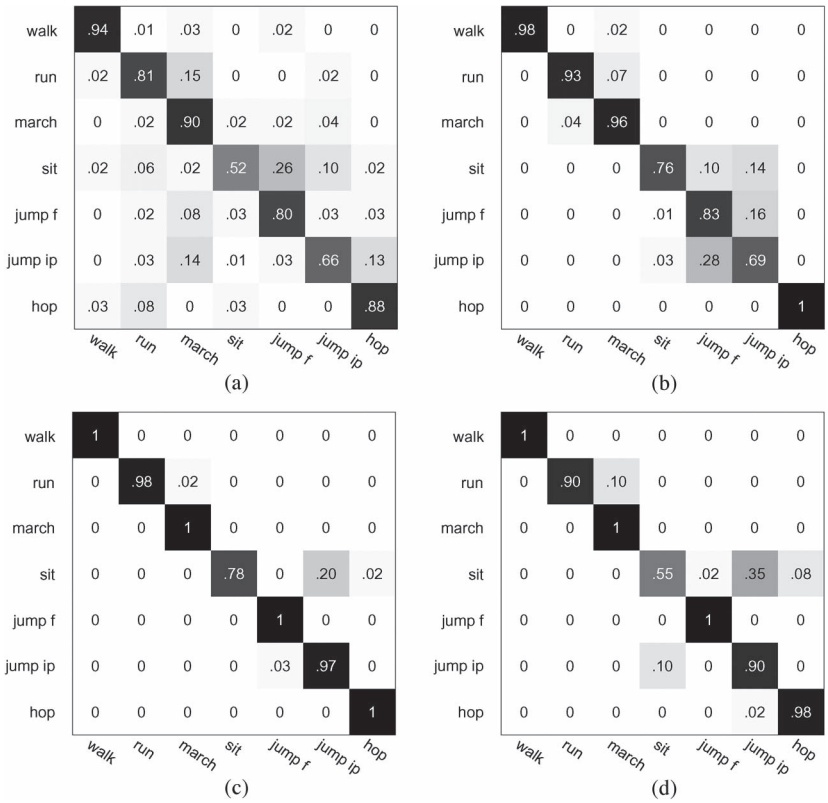


Fig. 9. Confusion matrices for the CMU data set. (a) Joint angle (six dimensions, using PCA). Average recognition rate is 78.4%. (b) Joint angle (four dimensions). Average recognition rate is 85.4%. (c) Joint torques (four dimensions). Average recognition rate is 95.7%. (d) Joint torques (four dimensions, computed from noisy joint angles). Average recognition rate is 90.4%.

We chose two other methods [53], [54] as additional baseline methods for comparison. Since code for these two methods is available online, we could test these methods on our own data set.

C. HMM Parameter Selection

To fine-tune the parameters of the HMM, we used tenfold cross validation. We tuned the number of states N_s and the number of mixture components K for the best performance. For the experiments with joint torques, we used $K = 3$ and $N_s = 3$. In the case of joint angles, we used $K = 4$ and $N_s = 2$ for the best classification results.

D. Results on the CMU Data Set

The classification results obtained using our approach and those using kinematic features are shown in Fig. 9. To produce the results, we computed the average over 10 runs with random permutations of training sets. Using dynamic features, we achieved mean accuracy of 95.7% on the entire data set. *Walk*, *march*, *jump forward*, and *hop* action classes were classified with 100% accuracy. There was some confusion in classifying *sit*, which was confused with *jump in place* (20%). The other confused actions are negligible ($\leq$ 3%). This is reasonable performance considering the similarity between these actions. Using the baseline method, we obtained mean classification accuracy of 78.4% in the first case (PCA reduced, six components) and 85.4% in the second case (four joint angles). Using dynamic features therefore outperforms both of these feature sets. These results also demonstrate that PCA-selected features are not sufficiently robust for discrimination. Confusion matrices for the kinematic features are shown in Fig. 9(a) and (b).

In the case of kinematic features, there was significant confusion during the classification of the *run*, *sit*, *jump forward*, and *jump in place* classes. However, the confusion was relatively small in the case of the dynamic features.

To evaluate the discriminative power of the dynamic and kinematic features, we compared some of these features for two action pairs: 1) *jump forward* and *jump in place* [see Fig. 10(a)]; and 2) *run* and *march* [see Fig. 10(b)]. It can be seen that the interclass distance is small in the case of joint angles, whereas the distance is larger in the case of the joint torques. We observed that the PCA-reduced kinematic features also showed poor discrimination.

We also carried out another set of experiments to test the robustness of the proposed method in the case of noisy motion capture data. To simulate noisy measurements, we added Gaussian noise to all joint angles. However, for the vertical distance of the foot from the ground, we used clean motion capture data because accurate determination of the ground contact event is required to compute the torques accurately. In our method, we need to differentiate the joint angles to obtain the angular velocities and accelerations. However, due to the abrupt change in the noisy joint angles, we obtained unrealistic values for velocities and accelerations. To overcome this problem, we sampled the noisy joint angles at regular intervals and reconstructed smoothed versions using polynomial

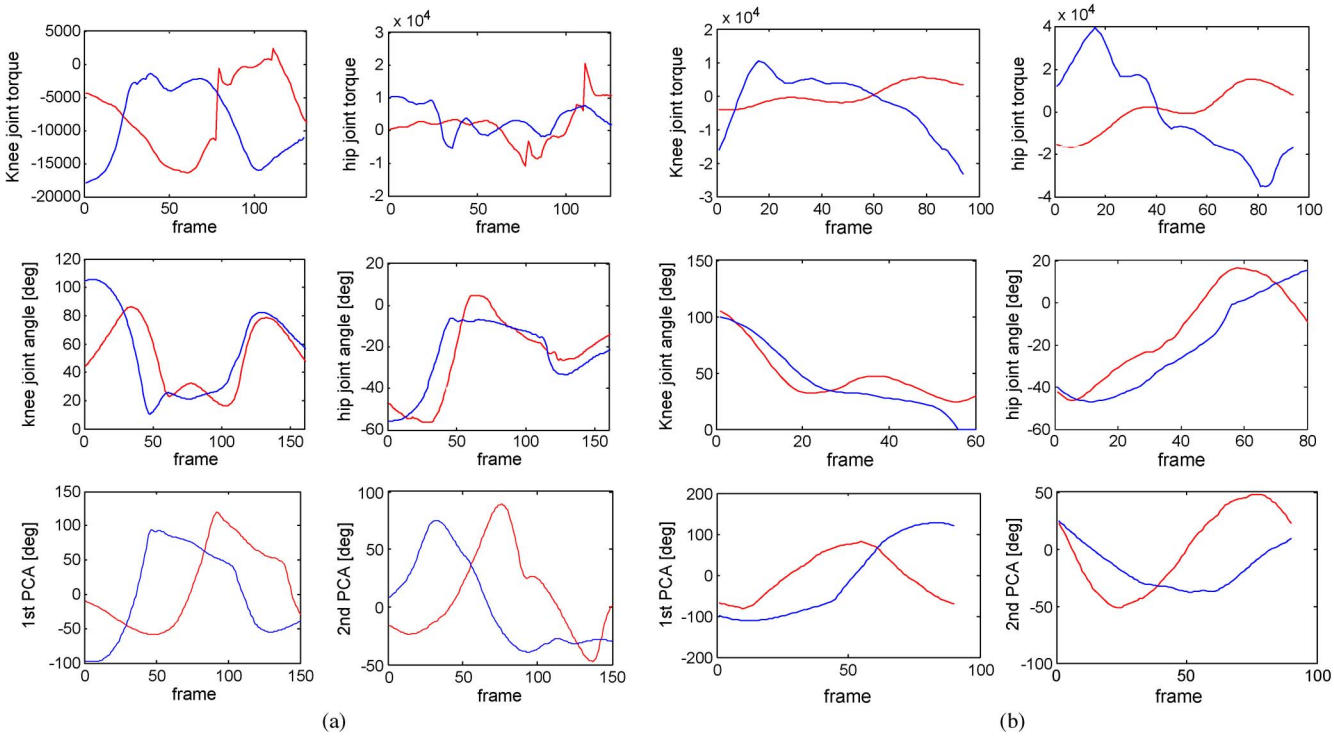


Fig. 10. Comparison of joint angle and torque features. (a) Comparison of joint angle and torque features for two action classes (i.e., jump forward and jump in place). (Top row) Knee and hip joint torques; (middle row) knee and hip joint angles; and (bottom row) first and second principal components of PCA-reduced joint angles. (Red) Jump forward; (blue) jump in place. The computed torques were not calibrated, and therefore, absolute units are not used. (b) Comparison of joint angle and torque features for two action classes (i.e., run and march). (Top row) Knee and hip joint torques; (middle row) knee and hip joint angles; and (bottom row) first and second principal components of PCA-reduced joint angles. (Red) Run; (blue) march. The computed torques were not calibrated, and therefore, absolute units are not used.

curve fitting. With these noisy joint angles, we computed the joint torques, with mean classification accuracy of 90.4%. The confusion matrix for this experiment is shown in Fig. 9(d). This result gives an idea of the performance of a method using the output of a tracker. In addition, it should be noted that even with the noisy data, dynamic features outperformed the kinematic features without noise.

E. Results on the Osaka University Kinect Action Data Set

For the evaluation on this data set, we used the joint angle trajectories with both the proposed method and the kinematics-based baseline method. The proposed method uses the computed torques from these trajectories, whereas the kinematics-based method uses the trajectories directly. The 8-D feature (joint angle or joint torque) obtained from the two hip joints (X and Z DOF) and two knee joints (X and Z DOF) was used. Upper joint trajectories were not used due to noise and errors in the tracking. For the other baseline methods [53], [54], which use appearance-based features, color images were used.

Confusion matrices for the proposed and baseline methods are shown in Fig. 11. Based on the results, it is clear that the appearance-based methods poorly perform on this challenging data set because many of the actions are very similar with respect to the appearance-based features. As expected, the joint angle feature performs much better, although there is still confusion between a few of the actions. For example, with regard to the kinematic features alone, the following action pairs produced similar joint angle trajectories: *skip left–jump left leg* and *side jump–jack 2*. Based on the results, there is obvious confusion between these action pairs. Using dynamic features, however, reduces such confusion, and the superiority of the dynamic features is clear from these results.

In our current implementation using MATLAB, the system takes a few seconds to compute the torques from an action comprising about 100 frames. However, processing could be made faster by using a more efficient code, for example, by implementing it in C/C++.

VI. Conclusion

In this paper, we have introduced dynamic features to realize the idea of employing a physics model for human action recognition. These dynamic features are computed from the available kinematics together with the known mass and inertia properties of the human body. Using these features, we expressed action classes in terms of hip and knee joint torques and used these torques for action recognition. The low-dimensional representation enabled us to achieve good classification accuracy with a small training data set. We carried out experiments on the CMU motion capture data set and the Osaka University Kinect action data set containing different human actions and demonstrated the superiority of dynamic features over kinematic features and other appearance-based methods.

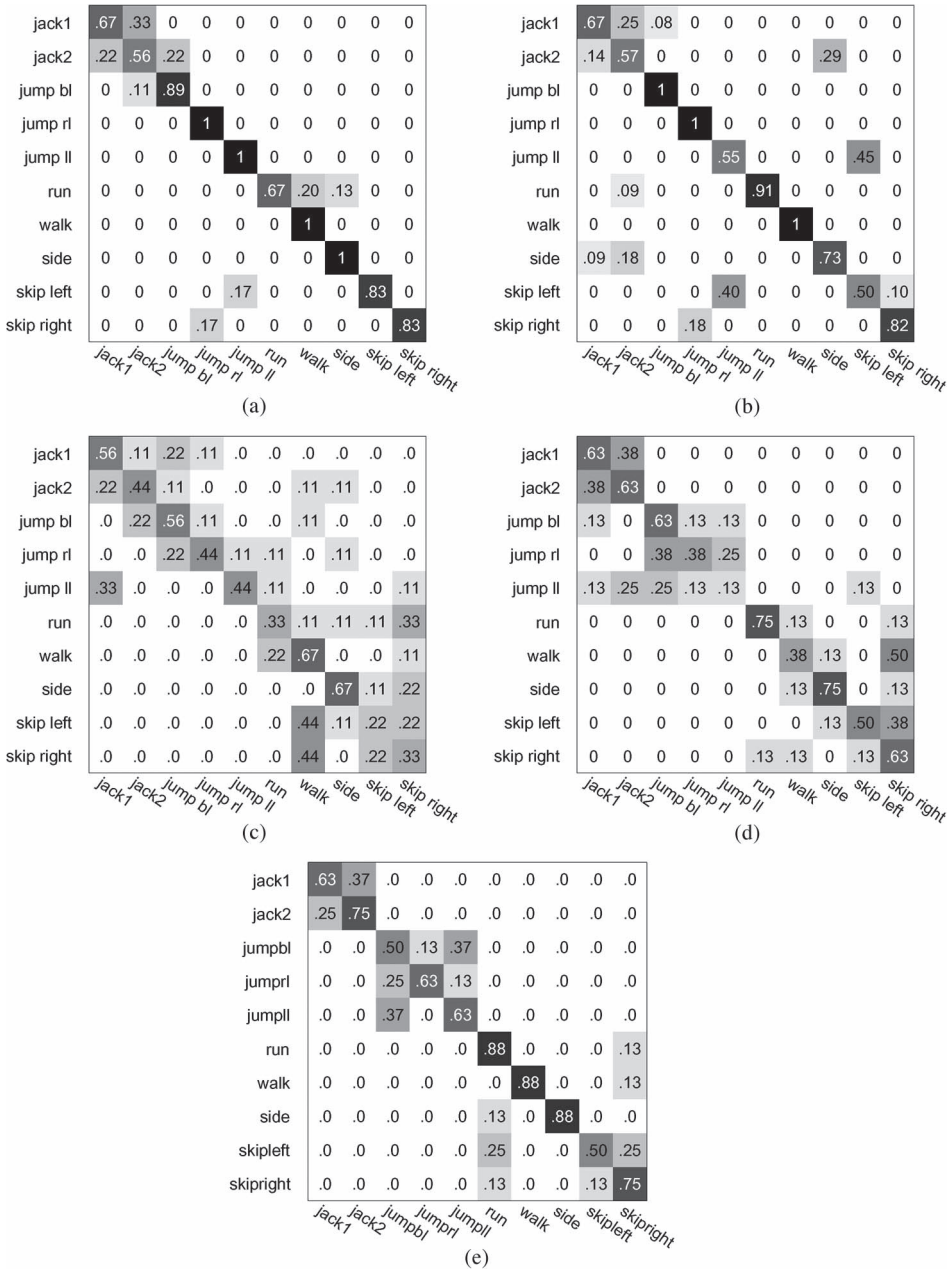


Fig. 11. Recognition performance on the Osaka University Kinect action data set. (a) Proposed method. Average recognition rate is 84.5%. (b) Kinematics feature. Average recognition rate is 77.5%. (c) Li *et al* [55]. Average recognition rate is 46.6%. (d) Bregonzio *et al* [53]. Average recognition rate is 54.1%. (e) Le *et al* [54]. Average recognition rate is 70.3%.

Future work may include investigating the use of dynamic features for other application domains such as human gait recognition.

REFERENCES

[1] L. Wang and D. Sutar, "Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model," in *Proc. IEEE CVPR*, 2007, pp. 1–8.

[2] D. Weinland and E. Boyer, "Action recognition using exemplar-based embedding," in *Proc. IEEE CVPR*, 2008, pp. 1–7.

[3] D. Tran and A. Sorokin, "Human activity recognition with metric learning," in *Proc. ECCV*, 2008, pp. 548–561.

[4] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 288–303, Feb. 2010.

[5] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *Proc. ICCV*, 2009, pp. 104–111.

[6] J. Sun, X. Wu, S. Yan, L. Cheng, T. Chua, and J. Li, "Hierarchical spatio-temporal context modelling for action recognition," in *Proc. IEEE CVPR*, 2009, pp. 2004–2011.

[7] A. Bissacco, A. Chiuso, and S. Soatto, "Classification and recognition of dynamical models: The role of phase, independent components, kernels and optimal transport," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 11, pp. 1958–1972, Nov. 2007.

[8] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, Jun. 2010.

[9] Y. Sheikh, M. Sheikh, and M. Shah, "Exploring the space of a human action," in *Proc. IEEE ICCV*, 2005, pp. 144–149.

[10] N. Ikizler and D. Forsyth, "Searching video for complex activities with finite state models," in *Proc. IEEE CVPR*, 2007, pp. 1–8.

[11] F. Lv and R. Nevatia, "Single view human action recognition using key pose matching and Viterbi path searching," in *Proc. IEEE CVPR*, 2007, pp. 1–8.

[12] P. Natarajan and R. Nevatia, "View and scale invariant action recognition using multiview shape-flow models," in *Proc. IEEE CVPR*, 2008, pp. 1–8.

[13] C. Fanti, L. Zelnik-Manor, and P. Perona, "Hybrid models for human motion recognition," in *Proc. IEEE ICCV*, 2005, pp. 1166–1173.
[14] A. Yilmaz and M. Shah, "Recognizing human actions in videos acquired by uncalibrated moving cameras," in *Proc. IEEE ICCV*, 2005, pp. 150–157.
[15] S. Ali, A. Basharat, and M. Shah, "Chaotic invariants for human action recognition," in *Proc. IEEE ICCV*, 2007, pp. 1–8.
[16] Y. Shen and H. Foroosh, "View-invariant action recognition from point triplets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1898–1905, Oct. 2009.
[17] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE CVPR*, 2011, pp. 1297–1304.
[18] I. Oikonomidis, N. Kyriazis, and A. Argyros, "Efficient model-based 3d tracking of hand articulations using Kinect," in *Proc. British Mach. Vis. Conf.*, 2011, pp. 101.1–101.11.
[19] M. Ye, X. Wang, R. Yang, L. Ren, and M. Pollefeys, "Accurate 3d body pose estimation from a single depth image," in *Proc. ICCV*, 2011, pp. 1–7.
[20] M. Vondrak, L. Sigal, and O. C. Jenkins, "Physical simulation for probabilistic motion tracking," in *Proc. IEEE CVPR*, 2008, pp. 1–8.
[21] M. Brubaker and D. J. Fleet, "The kneed walker for human pose tracking," in *Proc. IEEE CVPR*, 2008, pp. 1–8.
[22] M. Brubaker, L. Sigal, and D. Fleet, "Estimating contact dynamics," in *Proc. IEEE ICCV*, 2009, pp. 2389–2396.
[23] C. Liu, A. Hertzmann, and Z. Popovic, "Learning physics based motion style with nonlinear inverse optimization," *ACM Trans. Graph. (SIGGRAPH)*, vol. 24, no. 3, pp. 1071–1081, Jul. 2005.
[24] A. Mansur, Y. Makihara, and Y. Yagi, "Action recognition using dynamics features," in *Proc. IEEE ICRA*, 2011, pp. 4020–4025.
[25] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.
[26] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time sequential images using hidden Markov model," in *Proc. IEEE CVPR*, 1992, pp. 379–385.
[27] T. Starner and A. Pentland, "Visual recognition of American sign language using hidden Markov model," in *Proc. Int. Workshop Autom. Face Gesture Recognit.*, 1995, pp. 1–52.
[28] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space–time shapes," in *Proc. ICCV*, 2005, pp. 1395–1402.
[29] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *Proc. IEEE CVPR*, 2005, pp. 984–989.
[30] E. Shechtman and M. Irani, "Space–time behavior based correlation," in *Proc. IEEE CVPR*, 2005, pp. 405–412.
[31] A. Oikonomopoulous, I. Patras, and M. Pantic, "Spatiotemporal saliency for human action recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2005, pp. 1–4.
[32] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial–temporal words," in *Proc. BMVC*, 2006, pp. 1249–1258.
[33] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. IEEE Int. Workshop VS-PETS*, 2005, pp. 65–72.
[34] I. Laptev and T. Lindeberg, "Space time interest points," in *Proc. IEEE ICCV*, 2003, pp. 432–439.
[35] I. Laptev and P. Pérez, "Retrieving actions in movies," in *Proc. IEEE ICCV*, 2007, pp. 1–8.
[36] A. Kläser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *Proc. BMVC*, 2008, pp. 1–10.
[37] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. BMVC*, 2009, pp. 1–11.
[38] L. Shao and R. Mattivi, "Feature detector and descriptor evaluation in human action recognition," in *Proc. ACM Int. CIVR*, 2010, pp. 477–484.
[39] A. F. Bobick and J. Davis, "An appearance-based representation of action," in *Proc. CVPR*, 1996, pp. 307–312.
[40] A. F. Bobick, "Movement, activity, and action: The role of knowledge in the perception of motion," *Philos. Trans. Roy. Soc. Lond. B*, vol. 352, no. 1358, pp. 1257–1265, Aug. 1997.
[41] M. Yang, F. Lv, W. Xu, K. Yu, and Y. Gong, "Human action detection by boosting efficient motion features," in *Proc. IEEE ICCV Workshop*, 2009, pp. 522–529.
[42] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.
[43] J. Kilner, J.-Y. Guillemaut, and A. Hilton, "3D action matching with key-pose detection," in *Proc. IEEE ICCV Workshop Search 3D Video*, 2009, pp. 1–8.
[44] S. Moustakidis, J. Theocharis, and G. Giakas, "Subject recognition based on ground reaction force measurements of gait signals," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 6, pp. 1476–1485, Dec. 2008.
[45] A. Y. Yang, R. Jafari, S. S. Sastry, and R. Bajcsy, "Distributed recognition of human actions using wearable motion sensor networks," *J. Ambient Intell. Smart Environ.*, vol. 1, no. 2, pp. 103–115, Apr. 2009.
[46] N. Ravi, N. Dandekar, P. Mysore, and M. Littman, "Activity recognition from accelerometer data," in *Proc. 17th Conf. Innov. Appl. Artif. Intell.*, 2005, pp. 1541–1546.
[47] P. de Leva, "Adjustments to Zatsiorsky–Seluyanov's segment inertia parameters," *J. Biomech.*, vol. 29, no. 9, pp. 1223–1230, Sep. 1996.
[48] B. Rosenhahn and T. Brox, "Scaled motion dynamics for markerless motion capture," in *Proc. IEEE CVPR*, 2007, pp. 1–8.
[49] A. Balan, L. Sigal, M. J. Black, J. Davis, and H. Haussecker, "Detailed human shape and pose from images," in *Proc. IEEE CVPR*, 2007, pp. 1–8.
[50] Online. Available: http://mocap.cs.cmu.edu/
[51] Online. Available: http://www.am.sanken.osaka-u.ac.jp/~mansur/dataset.html
[52] Online. Available: http://www.ipisoft.com/
[53] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space–time interest points," in *Proc. IEEE CVPR*, 2009, pp. 1948–1955.
[54] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. IEEE CVPR*, 2011, pp. 3361–3368.
[55] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Proc. IEEE CVPR Workshop*, 2010, pp. 9–14.

Al Mansur received the M.Eng. degree in telecommunications from the Asian Institute of Technology, Pathumthani, Thailand, in 2002 and the Ph.D. degree in computer science from Saitama University, Saitama, Japan, in 2008.

He is currently a Researcher in the Institute of Scientific and Industrial Research, Osaka University, Ibaraki, Japan. His research interest includes computer vision, human motion analysis, gait, and action recognition.

Yasushi Makihara was born in Japan in 1978. He received the B.S., M.S., and Ph.D. degrees from Osaka University, Suita, Japan, in 2001, 2002, and 2005, respectively, all in engineering.

He is currently an Assistant Professor in the Institute of Scientific and Industrial Research, Osaka University, Ibaraki, Japan. His research interests are gait recognition, morphing, and temporal super resolution.

Dr. Makihara is a member of IPSJ, RSJ, and JSME.

Yasushi Yagi received the Ph.D. degree from Osaka University, Suita, Japan, in 1991.

He is the Director of the Institute of Scientific and Industrial Research, Osaka University, Ibaraki, Japan. In 1985, he joined the Product Development Laboratory, Mitsubishi Electric Corporation, where he worked on robotics and inspections. He became a Research Associate in 1990, a Lecturer in 1993, an Associate Professor in 1996, and a Professor in 2003 at Osaka University.

International conferences for which Prof. Yagi has served as Chair include ROBIO2006 (Program Cochair), ACCV2007 (Program Chair), PSVIT2009 (Financial Chair), ICRA2009 (Technical Visit Chair), ACCV2009 (General Chair), ACPR2011 (Program Cochair), and ACPR2013 (General Chair). He has also served as the Editor of IEEE ICRA Conference Editorial Board (2007–2011). He is a member of the Editorial Board of the International Journal of Computer Vision, the Editor-in-Chief of IPSJ Transactions on Computer Vision and Applications, and the Financial Chair of the Asian Federation of Computer Vision Societies.