Human Action Recognition Based on 3D Human Modeling and Cyclic HMMs

Shian-Ru Ke, Hoang Le Uyen Thuc, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi

Human action recognition is used in areas such as surveillance, entertainment, and healthcare. This paper proposes a system to recognize both single and continuous human actions from monocular video sequences, based on 3D human modeling and cyclic hidden Markov models (CHMMs). First, for each frame in a monocular video sequence, the 3D coordinates of joints belonging to a human object, through actions of multiple cycles, are extracted using 3D human modeling techniques. The 3D coordinates are then converted into a set of geometrical relational features (GRFs) for dimensionality reduction and discrimination increase. For further dimensionality reduction, k-means clustering is applied to the GRFs to generate clustered feature vectors. These vectors are used to train CHMMs separately for different types of actions, based on the Baum–Welch re-estimation algorithm. For recognition of continuous actions that are concatenated from several distinct types of actions, a designed graphical model is used to systematically concatenate different separately trained CHMMs. The experimental results show the effective performance of our proposed system in both single and continuous action recognition problems.

Keywords: Human action recognition, 3D modeling, hidden Markov model, geometrical relational features.

Manuscript received July 2, 2013; revised Nov. 25, 2013; accepted Dec. 3, 2013.

Shian-Ru Ke (corresponding author, srke@uw.edu) and Jenq-Neng Hwang (hwang@uw.edu) are with the Department of Electrical Engineering, University of Washington, Seattle, WA, USA.

Hoang Le Uyen Thuc (hluthuc@dut.udn.vn) is with the Department of Electronic and Telecommunication Engineering, Danang University of Technology, Danang, Vietnam.

Jang-Hee Yoo (jhy@etri.re.kr) is with the SW·Content Research Laboratory, ETRI, Daejeon, Rep. of Korea.

Kyoung-Ho Choi (khchoi@mokpo.ac.kr) is with the Department of Information and Electronics, Mokpo National University, Muan, Rep. of Korea.

I. Introduction

Human action recognition is a growing topic in video analysis and understanding — one of the most popular areas in the community of computer vision — thanks to its applications in surveillance, entertainment, and healthcare. In surveillance, human action recognition can be used in conjunction with video camera footage to help with the recognition and analysis of human actions. In entertainment, human–computer interaction can be helped to appear more natural via human action recognition, which in turn can help increase the entertainment experience. In healthcare, human action recognition can help detect abnormal gaits or assist in a patient's rehabilitation through an analysis of their actions.

However, it is challenging to recognize various human actions due to the high number of degrees of freedom associated with the average human body — namely, variations in human poses; variations in the colors of a person's clothing; changes in lighting and illumination; variations in viewpoints; and frequent self-occlusion. Moreover, the use of monocular video sequences further increases the difficulty for human action recognition.

Generally, the two main stages in human action recognition are: the feature extraction and representation stage and the classification stage.

In the feature extraction and representation stage, the features or characteristics of video frames, such as silhouette, shape, color, and motion, are extracted and represented in a systematic and efficient way. In a video sequence, the features that capture the relationship between space and time create what is known as the space–time volume (STV). The space–time correlation is one of the most popular features in the video analysis community. Blank and others [1] propose a method that

© 2014 ETRI

http://dx.doi.org/10.4218/etrij.14.0113.0647

consists of stacking segmented silhouettes (frame by frame) to form a 3D spatial–temporal shape. In a similar way, Ke and others [2] build STVs, for shape-based matching, from image features that are based on the consecutive silhouettes of objects along a time axis, including spatial–temporal region extraction and region matching. Kim and others [3] propose a spatiotemporal approach to detect the salient region in images and videos. Laptev and others [4], [5] propose a method to extract space–time interest points (STIP) by maximizing a normalized spatiotemporal Laplacian operator over spatial and temporal scales. In addition to space and time information, the frequency domain information of an image is also considered. Kumari and Mitra [6] use discrete Fourier transforms of small uniformly partitioned image blocks as the selected features for activity recognition. The STV, STIP, and DFT are global features extracted by globally considering the whole image.

However, the global features are sensitive to noise, occlusion, and variations of viewpoints. Therefore, some local descriptors are used to capture the characteristics of an image patch, such as scale-invariant feature transformation (SIFT) [7]–[8], Histogram of Oriented Gradient (HOG) [9], nonparametric weighted feature extraction (NWFE) features [10], and Lucas–Kanade–Tomasi (LKT) features [11]–[12]. For example, Scovanner and others [13] introduce a 3D SIFT descriptor, which can reliably capture the spatiotemporal nature of video sequences as well as 3D imagery such as MRI data. The SIFT descriptor is popular due to its invariance to image rotation and scale and is robust to affine distortion, noise corruption, and illumination changes. But SIFT has issues of high dimensionality and insufficient discrimination. Lu and Little [14] propose a template-based algorithm to track and recognize athletes' actions based on a PCA-HOG descriptor. Moreover, Kataoka and Aoki [15] propose an extension of HOG — namely, Co-occurrence Histograms of Oriented Gradients (CoHOG) for pedestrian detection. However, HOG features are extracted at a fixed scale; therefore, the size of a human body in an image has a great influence on performance. Moreover, by considering the distance information and the width feature of a silhouette, Lin and others [10] design a new feature for human activity recognition — namely, NWFE. Unfortunately, NWFE does not take advantage of color appearance information. Furthermore, Lucas–Kanade [7] and Tomasi [8] propose an LKT feature tracker based on the sum of squared intensity differences. Lu and others [16] use an LKT feature tracker to track human joints and recognize non-rigid human actions. Since the LKT feature tracker assumes that neighboring pixels in a small window have the same flow vector, it poses a limitation on large motion between frames.

Due to limitations on 2D global and local features, Weinland and others [17] use 3D exemplars; that is, 3D occupancy grids as features. Without 3D reconstruction, the learned 3D exemplars are used to produce 2D silhouette frames for matching. However, their method [17] is limited to view variations. Junejo and others [18] resolve the changes of view angles by using a self-similarity matrix (SSM), obtained by computing self-similarities (distance between low level features) of action sequences over time. But SSM is only useful for body occlusion.

To resolve body occlusion and take into account body configuration, 3D human modeling is thus considered. Subsequently, estimated 3D coordinates are converted into a feature space. For 3D human modeling, Rogez and others [19] use a series of view-based shape-skeleton models for video surveillance systems by projecting the input image frames onto the training plane. However, this needs an extensive training dataset due to the various viewpoints involved. Lee and Nevatia [20]–[21] use a multi-level structure model for 3D pose estimation from monocular video sequences by addressing automatic initialization, data association, self- and inter-occlusions. 3D human poses are inferred based on a method of Data-Driven Markov Chain Monte Carlo [22]–[23]; however, the computation cost is extremely high. Considering accuracy of pose estimation and time complexity simultaneously, Ke and others [24]–[25] propose a method to track 2D body parts by integrating shape, color, and temporal information to effectively estimate 3D human poses.

Generally, the 3D coordinates of human joints can be further converted into low-dimensional or more discriminative features, such as polar coordinate representation [26], Boolean features [27], or geometrical relational features (GRFs) [28], [29], for the purpose of effective recognition.

In the classification stage, the selected or converted features are sent to proper classification algorithms for detection or recognition. One of the well-known classification algorithms is the Dynamic Time Warping (DTW) algorithm [30], which is a similarity measurement between two sequences by a dynamic programming approach. Sempena and others [31] use DTW to recognize human activities such as waving, punching, and clapping. DTW is simple and fast, but it might need extensive templates for various situations, which unfortunately results in high computational costs.

Besides DTW, probability-based methods by discriminative and generative models for classification have also been proposed. Discriminative models learn a posterior probability distribution $P(Y|X)$ of a specific class label Y given the observed variable X. A support vector machine (SVM) [32]–[33], one of the more popular discriminative models, is designed to find the optimal dichotomic hyperplane that can maximize the margin of two classes. Schuldt and others [34] apply SVMs to recognize human activities by extracting local

space–time features in a video. The main drawback of an SVM is the high computation burden for the constrained optimization programming used in the learning phase. On the other hand, generative models learn the joint probability distribution $P(X,Y)$, which can be used to generate samples from the distribution. The hidden Markov model (HMM) [35] is one of the most popular generative models. It follows a doubly stochastic process with an underlying hidden first-order Markov stochastic process and an observed stochastic process that can produce a sequence of observed symbols. Yamato and others [36] train, based on low-level image mesh features, HMMs [37] to recognize actions of different tennis strokes by the Baum–Welch re-estimation algorithm. Considering the multiple cycles of human actions, Thuc and others [38] proposed a cyclic HMM (CHMM) to effectively adapt to most quasiperiodic human action recognition tasks. But each separately trained CHMM can only recognize a single action, rather than continuous actions with different types of concatenated actions during the testing phase.

In this paper, we propose a system to recognize single actions and continuous human actions concatenated from different types of actions. To deal with changes of illumination, changes of clothes, changes of viewpoints, and occlusions, 3D human modeling is considered. The estimated 3D coordinates of human joints are further converted into GRF vectors, which are then clustered into one-dimensional feature vectors based on a k-means clustering algorithm [39]. The separately trained CHMMs of different types of actions are trained by these one-dimensional feature vectors. Designed graphical models for switching CHMMs are used to recognize continuous actions through the Graphical Model Tool Kit (GMTK) [40].

The overview of the proposed system is shown in Fig. 1. The inputs of the system are monocular video sequences. In the first phase, a human object is segmented from selected video frames. Then the human object's 3D poses are estimated by using the 3D coordinates of their body joints. This is all done based on a 3D human pose estimation technique [25]. In the second phase, the estimated 3D coordinates of the body joints are converted into one-dimensional feature vectors, based on GRF conversion [29] and k-means clustering [39]. In the third phase, one-dimensional feature vectors are used to train CHMMs, one model for one type of action, and then designed graphical models are used to recognize continuous human actions concatenated from different types of human actions by switching between CHMMs. Finally, the recognized human actions are created.

This paper is organized as follows. Section II introduces the scheme of 3D human pose estimation. Section III describes the feature conversion and classification algorithm. Section IV explains the experiments and evaluation based on the proposed system. Finally, a conclusion is given in section V.

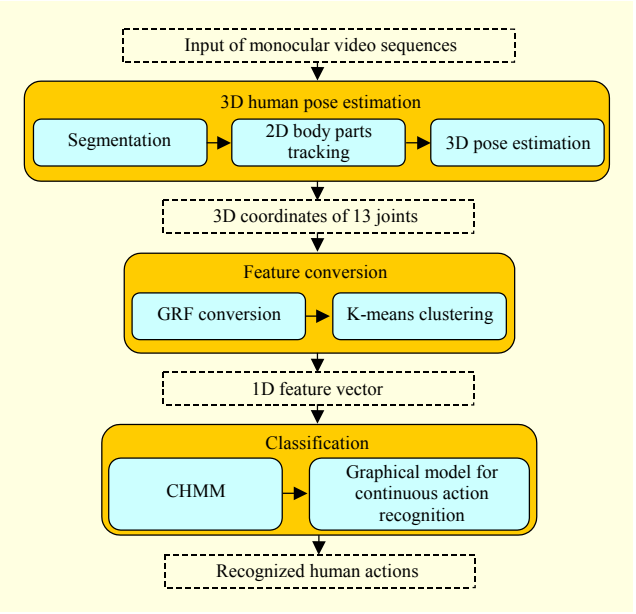


Fig. 1. Overview of proposed system.

II. 3D Human Pose Estimation

The objective of this phase is to extract the 3D coordinates of 13 human joints from monocular video sequences [25]. Three stages are in [25]: individual segmentation and 2D features extraction, 2D body parts tracking, and 3D pose estimation.

1. Segmentation and 2D Features Extraction

First of all, a human object is segmented and corresponding 2D features are extracted, as shown in Fig. 2. The segmentation with shadow removal, for a human object, is based on the distortions in both brightness and chromaticity [41] of each pixel in each frame. From this, a human object's silhouette for each frame can be generated. Further, the skin pixels can be detected in RG space [42]. Based on the segmented silhouette, an edge image is obtained by applying the Canny edge algorithm [43]. Finally, the motion image is the absolute

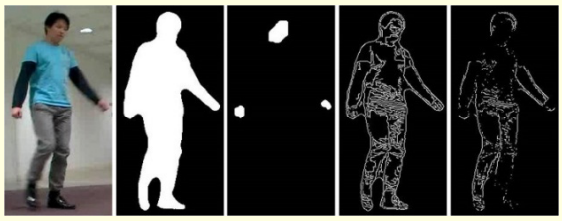


Fig. 2. 2D features (from left to right): input frame, silhouette, skin blobs, edge, and motion image.

ETRI Journal, Volume 36, Number 4, August 2014
http://dx.doi.org/10.4218/etrij.14.0113.0647

difference between the previous frame and the current frame.

2. 2D Body Parts Tracking

In this step, five blobs are used to represent the following 2D body parts: head, left hand, right hand, left foot, and right foot. These five blobs are tracked frame-by-frame, based on shape, color, and temporal information. Three techniques — 2D skeletonization scheme [25], mean-shift tracking algorithm [44]–[45], and Kalman filter prediction [46] — are separately applied to take advantage of the shape, color, and temporal information. The trajectories of the five blobs are shown in Fig. 3.

3. 3D Pose Estimation

The information of 2D features and locations of 2D body-part blobs are used to reconstruct 3D human poses, through the methodology of analysis-by-synthesis, by fitting 2D features to 3D poses. The downhill simplex algorithm [47] is applied to minimize the cost function so as to find the optimal 3D human pose for each frame. The cost function is a measure of the difference between the 2D features and the 3D model projections; it is composed of the following four scores: silhouette score, edge score, motion score, and feature point score. The estimated 2D skeleton, estimated locations of 3D joints, and corresponding estimated 3D human pose are shown in Fig. 4.

Fig. 3. Trajectories of the five blobs (head, left hand, right hand, left foot, and right foot).

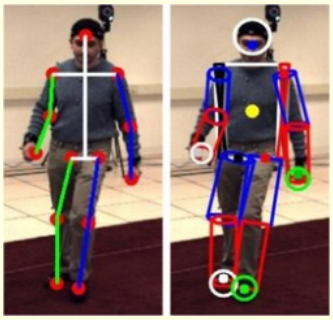


Fig. 4. Left: 2D skeleton and locations of 13 joints. Right: corresponding 3D model.

III. Feature Conversion and Classification

After extracting the 3D coordinates of human joints from the estimated human poses, these 3D coordinates are further converted into just one symbol at each frame for dimensionality reduction and discrimination increase. Two operations are considered: individual GRF conversion [28], [29] and k-means clustering [39]. The dimensionality reduction is shown in Fig. 5.

After the input video sequences are converted into one-dimensional feature vectors, these feature vectors of different types of actions with various repeated cycles are used to train the corresponding CHMMs [38], with one action being modeled by one CHMM, based on the Baum–Welch algorithm [35]. Subsequently, a switching graphical model is designed to switch between different CHMMs for a long video sequence concatenated from different types of human actions and to recognize the continuous combined human actions based on the maximal likelihood of the observation sequence, computed by forward/backward and Viterbi algorithms [35].

1. GRF Conversion

The GRF descriptor defines the degree of the relative position of the body parts in 15 dimensions, as described in Table 1. Two types of features are included in the GRF descriptor: distance-related features (F_1 to F_9) and angle-related features (F_{10} to F_{15}). The features F_1 to F_4 define the geometrical relation between a point and a plane. Take F_1 as an example, which is the measurement of the distance between the right hand and the plane formed by the right shoulder, left pelvis, and right pelvis. The illustration of F_1 is shown in Fig. 6. The features of F_5 and F_6 define the geometrical relation between two vectors. One vector is defined as (R-Shoulder, R-Hand) or (L-Shoulder, L-Hand), and the other is a unit vector, defined as (middle of L-Pelvis and R-Pelvis, Head). The inner product of these two vectors is used to provide the signed distance, to indicate the degree of R/L-Hand above or below R/L-Shoulder. The feature of F_7 is defined as the distance between the centroid of the body and the lowest foot

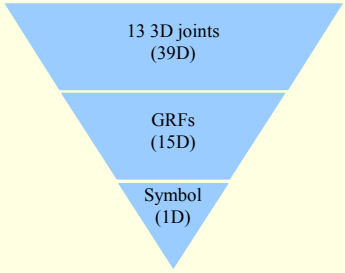


Fig. 5. Dimensionality reduction.

Table 1. GRF descriptors.

Feature	Description
$F_{1,2}$	Signed distance between R/L-hand and the plane defined by R/L-shoulder, L-pelvis, and R-pelvis.
$F_{3,4}$	Signed distance between R/L-foot and the plane defined by L-shoulder, R-shoulder, and R/L-pelvis.
$F_{5,6}$	Signed distance between vector of R/L-hand and R/L-shoulder, and unit vector of the middle of pelves and head.
F_7	Distance between the centroid of the body and the lowest foot (y-coordinate).
F_8	Distance between R-foot and L-foot (y- coordinate).
F_9	Accumulated distance between the centroid of the body at current frame and the centroid of body at first frame.
$F_{10,11}$	Angle between upper and lower R/L-arm.
$F_{12,13}$	Angle between upper and lower R/L-leg.
F_{14}	Angle of the body bending vertically (x-coordinate).
F_{15}	Angle change of body rotation horizontally between the previous frame and the current frame (y-coordinate).

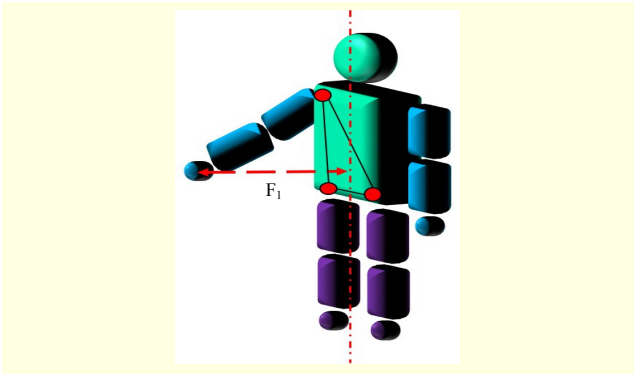


Fig. 6. Illustration of F_1 feature in GRF descriptor.

in a vertical direction (y-coordinate). F_7 might display large changes due to the actions of sitting down and getting up. The feature of F_8 is defined as the distance between R-Foot and L-Foot in a vertical direction (y-coordinate). The feature of F_9 is defined as the accumulated distance between the centroid of the body at the current frame and the centroid of the body at the first frame. Feature F_9 can hint at whether the body acts always at the original location or acts with some movement, such as in spinning versus circling. Moreover, the angle-related features of F_{10} to F_{13} measure the degree of the angle (in the unit of π) of a body part. Take F_{10} as an example, which is the measurement of the angle between the upper and the lower right arm, F_{10} is the arccosine of the inner product of the two unit vectors of the upper and the lower right arm. The feature of F_{14} is defined as the degree of the angle of the body bending vertically (x-coordinate). The feature of F_{15} is defined as the change of the angle of the body rotation horizontally (y-coordinate) between the previous frame and the current frame. The effectiveness of the definition of GRFs will be proved in experiments later.

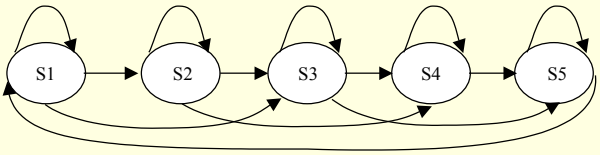


Fig. 7. Transition graph of CHMM.

2. K-Means Clustering

The 15-dimensional GRF vectors are further clustered into k centroids (codewords) by the k-means algorithm. Each GRF vector can be represented by one of the k codewords (that is, a symbol) based on the nearest centroid. Therefore, each frame is eventually converted into one symbol out of k possible values, and the input monocular video sequence can now be represented as a one-dimensional feature vector.

3. CHMM

An HMM is a doubly stochastic process with an underlying hidden stochastic process and an observed stochastic process. Many human actions, such as walking and waving, exhibit quasi-periodicity, where repeated movements are not identical in each cycle and the number of cycles is also indefinite without a predefined value. This motivates our use of a CHMM (see Fig. 7 [38]) — an HMM that has left-to-right structure with a return transition from the last state to the first state, so as to have the capability to model actions with multiple cycles. Each CHMM is trained using action sequences (labelled according to type) containing variable multiple cycles. The action sequences in the testing dataset can be recognized by finding the maximal likelihood based on the forward–backward algorithm.

4. Switching Graphical Model

Based on the trained single-action CHMMs, a graphical model, based on the GMTK [40], is designed for recognition of continuous human actions that are concatenated from several different human actions.

IV. Experiments and Evaluation

The proposed system is performed on two different datasets: the self-recorded dataset and the IXMAS dataset [17]. In each dataset, the following two experiments are performed: single-

action recognition and continuous-action recognition. In single-action recognition, each action type is trained as a single-action CHMM. In continuous-action recognition, different action types are combined to form long video sequences. The long video sequences with distinct types of actions are recognized by the designed switching graphical model, which is concatenated from different types of trained single-action CHMMs. The evaluations for 3D pose estimation and classification algorithms are also given at the end.

1. Self-recorded Dataset

In our self-recorded dataset {person A, person B, person C, person D}, each of the four people performed each of the five different actions (boxing, kicking, throwing, waving, and standing) five times. Therefore, in total there were 100 monocular video sequences for each of the four people.

A. Single-Action Recognition

In the experiments of single-action recognition, each action sequence (short sequence), not necessarily of the same length, is concatenated from one of the five different actions performed by the same person. Since each action was performed five times, there are 15 variable-cycle sequences for each action: one 5-cycle (1-2-3-4-5), two 4-cycle (1-2-3-4, 2-3-4-5), three 3-cycle (1-2-3, 2-3-4, 3-4-5), four 2-cycle (1-2, 2-3, 3-4, 4-5), and five 1-cycle (1, 2, 3, 4, 5) action sequences. Therefore, for each person there is a total of 75 short sequences (15 short sequences of variable size × 5 actions) — each being of variable length. Two experiments are performed for single-action recognition.

The first experiment is the recognition generalization of the unknown person — that is, every single-action CHMM is trained by using the actions from 3 individuals and recognized by the actions performed by the fourth person. Therefore, 225 short sequences (15 short sequences × 3 people × 5 actions) are used for training — that is, five CHMMs for five actions. Seventy-five short sequences (15 short sequences × 1 person × 5 actions) are used for recognition. There are four subexperiments for the four people in the testing dataset. The average recognition rate, 92.7%, shows the favorable performance for the unknown person recognition.

The second experiment is the cross-validation for the mixture of persons. For each type of action, out of 60 short sequences, 48 short sequences are randomly selected for training and the remaining 12 short sequences are used for testing in each of the CHMMs — that is, in total, 240 short sequences for training and 60 short sequences for testing. There are five independent subexperiments that are performed for cross-validation. The average recognition rate, 100%, shows perfect performance.

B. Continuous-Action Recognition

In the recognition of continuous actions experiment, each continuous action sequence (long sequence) is concatenated from different actions performed by the same person. Each long sequence consists of six stages, and for each stage we randomly selected one of the four actions (boxing, kicking, throwing, or waving). Additionally, at each stage the corresponding standing action, for the person in question, was inserted right after each randomly chosen action. Thus, gradually and sequentially we could compose a long sequence in such a manner. Therefore, there are a total of 4,096 (4×4×4×4×4×4) long sequences for each person. We randomly sampled 100 of the total number of long sequences for testing in each subexperiment. Based on the designed switching graphical model, which dynamically switches between CHMMs, each of the underlying six actions is recognized in the 100 randomly sampled long sequences. In total, 600 actions are recognized in each subexperiment.

Based on whether standing action sequences or a standing model is used, seven different subexperiments are conducted, as shown in Table 2. In subexperiment 1, no standing sequences are appended to each of the six continuous actions. In subexperiment 2, three to five frames of standing actions are appended to each of the six continuous actions. The overall recognition rates for subexperiments 1 and 2 are 70.5% and 65.75%, respectively. Moreover, we carried out adding standing frames combined with one of the boxing, kicking, throwing, or waving actions to help with the training of CHMMs. That is, three to five extra standing frames and seven to ten extra standing frames individually in subexperiments 3 and 4, respectively. The overall recognition rates for

Table 2. Seven subexperiments for continuous-actions recognition.

	Standing sequences	Standing CHMM	Recognition rate
Subexp. 1	0 frame	No	70.5%
Subexp. 2	3–5 frames	No	65.75%
Subexp. 3	3–5 frames	No, but trained in other CHMMs	67%
Subexp. 4	7–10 frames	No, but trained in other CHMMs	66%
Subexp. 5	30–40 frames	Yes, trained in standing CHMM	100%
Subexp. 6	7–10 frames	Yes, trained in standing CHMM	99.55%
Subexp. 7	3–5 frames	Yes, trained in standing CHMM	100%

subexperiments 3 and 4 are 67% and 66%, respectively. Furthermore, the extra standing frames are trained independently in a standing CHMM. In subexperiments 5, 6 and 7, individually thirty to forty standing frames, seven to ten standing frames, and three to five standing frames are used to train a standing CHMM. With the explicitly standing CHMMs appended right after each action, the recognition rates for subexperiments 5, 6, and 7 almost achieved 100%. The conclusion is drawn that the recognition rate can reach almost 100% if the standing frames are used to train the standing CHMM and are inserted between actions, regardless of the number of standing frames. On the other hand, the insertion of standing frames into only the test sequences can significantly degrade recognition performance, especially when CHMM training is done without the insertion of standing frames.

2. IXMAS Dataset

To further justify the effectiveness of our proposed system, we also tried the IXMAS dataset [17], where 11 types of actions are performed by 12 actors, including "check_watch," "cross_arm," "scratch_head," "sit_down," "get_up," "turn_around," "walk," "wave," "punch," "kick," and "pick_up." The two experiments, single-action recognition and continuous-action recognition, are performed in the IXMAS dataset using camera 3.

A. Single-Action Recognition

In the single-action recognition experiments recorded in the IXMAS dataset, the estimated 3D coordinates of the human body are converted into GRFs as features, and the CHMM method is used as the recognition algorithm (exactly the same as that used in section IV-1A). To show the effectiveness of the GRFs, we compare the recognition performance based on features of the absolute 3D coordinates of joints and the relative 3D coordinates of joints. The absolute 3D coordinates are the 3D joints with respect to the world's origin, while the relative 3D coordinates are the 3D joints with respect to the centroid of the human body. Moreover, we also compare the proposed method with three other up-to-date methods using the IXMAS dataset. The first method is proposed by Weinland and others [17] to represent 3D exemplars, the 3D occupancy grids obtained from the projections of multiple cameras, as features and to use HMMs as the recognition algorithm. The second method is proposed by Junejo and others [18] to represent SSM plus HoG and optical flow (OF) as features and to use SVMs as the recognition algorithm. The third method is proposed by Laptev and others [4]–[5] to represent STIPs as features and to use SVMs as the recognition algorithm. A comparison of the recognition rates is shown in Table 3, and the confusion matrix of the proposed method is shown in Table 4. As shown in Table 3, our proposed method of GRFs

Table 3. Comparison of recognition rates under IXMAS dataset.

Method	Recognition rate
3D exemplar + HMM [17]	80.5%
SSM_HoG_OF + SVM [18]	71.2%
STIP + SVM [4], [5]	85.5%
Absolute 3D + CHMM	74.2%
Relative 3D + CHMM	78.6%
GRF + CHMM	91.7%

Table 4. Confusion matrix of proposed method under IXMAS dataset.

Single_action (%)	Check_watch	Cross_arm	Scratch_head	Sit_down	Get_up	Turn_around	Walk	Wave	Punch	Kick	Pick_up
Check_watch	92	0	0	0	0	0	0	0	8	0	0
Cross_arm	0	92	0	0	0	0	0	0	8	0	0
Scratch_head	0	0	92	0	0	0	0	0	8	0	0
Sit_down	0	0	0	100	0	0	0	0	0	0	0
Get_up	0	0	0	0	100	0	0	0	0	0	0
Turn_around	0	0	0	0	0	75	25	0	0	0	0
Walk	0	0	0	0	0	0	100	0	0	0	0
Wave	8	0	8	0	0	0	0	83	0	0	0
Punch	0	0	0	0	0	0	0	0	92	8	0
Kick	0	0	0	0	0	0	0	0	0	100	0
Pick_up	0	0	0	17	0	0	0	0	0	0	83

Table 5. Confusion matrix for continuous-actions recognition in four actions under IXMAS dataset.

Four actions (%)	Check_watch	Cross_arm	Scratch_head	Sit_down	Get_up	Turn_around	Walk	Wave	Punch	Kick	Pick_up
Check_watch	88.8	0.9	0.7	1.0	0.1	3.6	0.3	0.5	2.0	1.8	0.4
Cross_arm	1.3	84.6	0.4	1.4	0.4	2.1	0.7	0.8	5.5	2.0	0.6
Scratch_head	3.0	2.5	82.5	0.5	0.5	1.5	0.6	1.2	5.3	1.3	1.0
Sit_down	2.5	1.9	0.5	86.0	0.1	1.6	0.2	1.0	1.6	2.7	1.9
Get_up	1.2	0.1	0.3	0.3	96.5	0.3	0.5	0.2	0.3	0.2	0.1
Turn_around	3.2	0.7	0.4	0.4	0.3	84.7	6.2	0.7	0.8	2.4	0.2
Walk	0.8	0.5	0.1	0.8	0.3	5.5	88.9	0.6	0.4	1.5	0.6
Wave	6.1	0.8	8.8	0.7	0.2	3.8	1.0	76.6	0.7	1.0	0.5
Punch	2.7	0.8	0.5	0.5	0.3	2.1	1.6	0.7	88.0	2.3	0.5
Kick	1.6	1.1	0.8	0.5	0.3	2.8	0.9	1.2	1.6	88.6	0.6
Pick_up	1.3	0.3	0.6	6.2	0.4	0.9	0.9	0.2	1.8	2.0	85.3

with 91.7% outperforms the performance of absolute 3D coordinates with 74.2% and relative 3D coordinates with 78.6%. It shows the more effective and discriminative representation of the GRFs. In addition, the proposed method of GRFs + CHMM with 91.7% also outperforms 3D exemplar + HMM [17] with 80.5%, SSM_HoG_OF + SVM [18] with 71.2%, and STIP + SVM [4], [5] with 85.5%. The improved performance of the proposed method is most likely attributed to three main factors: the use of 3D pose estimation, which allows for the invariance of viewing perspectives, as well as mitigating the effects caused by changes in illumination and body occlusions; the effectiveness of the conversion of the GRFs, which not only decreases the dimensionality to result in better CHMM training but also increases the discrimination of the features; and the use of cyclic HMMs, which enable the same actions to be repeated a variable number of times.

B. Continuous-Action Recognition

In addition to the continuous-action recognition experiments conducted using the IXMAS dataset [17], two subexperiments are performed.

In the first subexperiment, each long sequence is formed by concatenating four actions. Each action is randomly selected from one of the 11 action types in the IXMAS dataset, resulting in a total of 14,641 (11×11×11×11) possible long sequences to be compared against. We randomly chose 100 of the 14,641 long sequences for testing with 10-fold cross-validation. The confusion matrix is shown in Table 5. The average recognition rate is 86.4%, which is a satisfactory rate.

In the second subexperiment, each long sequence is formed by concatenating two, three, or four actions with actions again being selected from the 11 types of actions in the IXMAS dataset. Compared to the first subexperiment, the long sequence is no longer restricted to four actions, rather two, three, or four actions; making the recognition more difficult. Totally, there are 16,093 (11×11×11×11+11×11×11+11×11) possible long sequences to be compared against. We also randomly chose 100 of the 16,093 long sequences for testing with 10-fold cross-validation. Because the number of types of actions is variable, Levenshtein distance (that is, edit distance) [48] is applied for measuring the error distance between a ground-truth sequence and a recognized sequence. The Levenshtein distance is defined as the minimum number of edits (insertion, deletion, or substitution) between two sequences. For example, the Levenshtein distance is equal to 1 between sequence A={a,b,c} and sequence B={a,c} with only one edit (deletion of b in A). The recognition rate for two, three, or four actions is defined in (1). The results of the recognition rates for continuous four actions and continuous two, three, or four actions are shown in Table 6. The recognition rate for two, three, or four actions (85.1%) is 1.3% less than the recognition rate for four actions (86.4%). Therefore, it shows that the recognition rate is still satisfactory even when the variable number of action types is concatenated.

$$\text{Recognition rate} = 1 - \frac{Sum\ of\ edit\ distance\ for\ all\ seq}{Sum\ of\ actions\ for\ all\ truth\ seq}. \quad (1)$$

3. Evaluation

In this section, two main parts of our proposed system are discussed — namely, 3D pose estimation and CHMM. First, the 3D pose estimation is compared with other state-of-the-art techniques. Second, with the same sequential data (that is, the GRF features), CHMM is compared with the Dynamic Time

Table 6. Recognition rates for continuous-actions recognition.

	Continuous 4 actions	Continuous 2/3/4 actions
Recognition rate (%)	86.4	85.1

Table 7. Comparison of mean error on HumanEvaII (in pixels).

Subject/camera	Mean (std) [23]	Mean (std) [24]	Mean (std) proposed
S2C1	16.96 (4.83)	12.98 (3.5)	10.43 (2.68)
S2C2	18.53 (5.97)	14.18 (4.38)	9.82 (2.18)

Table 8. Comparison of DTW and CHMM under IXMAS dataset.

	GRFs + DTW	GRFs + CHMM
Recognition rate (%)	68.2	91.7

Warping (DTW) to show the advantage of the use of CHMM.

A. Evaluation for 3D Pose Estimation

To evaluate the 3D pose estimation, our proposed method is compared with two other state-of-the-art methods [49]–[50] using the well-known HumanEvaII dataset [51]. In [49], Rogez and others use spatiotemporal 2D models to fit shape–skeleton features. In [50], Rogez and others apply random forest classifiers on HOG features. Table 7 shows the experiment results of the pose estimation for [49]–[50] and our proposed method on the HumanEvaII dataset, including the videos on subject 2 of camera 1 (S2C1) and subject 2 of camera 2 (S2C2). Table 7 shows our proposed system outperforms [49] and [50].

B. Evaluation for Classification Algorithms

With the same sequential data (that is, GRF sequences) CHMM is compared with DTW under the IXMAS dataset. The comparison of the DTW and CHMM is provided in Table 8, in which DTW and CHMM obtain recognition rates of 68.2% and 91.7%, respectively. It shows our proposed GRFs method is much more appropriate for CHMM than for DTW.

V. Conclusion

The proposed system is to recognize both single and continuous combined human actions. First of all, with the input monocular video sequences, 3D human modeling is applied to extract the 3D coordinates of 13 joints. With GRF conversion and k-means clustering, the 39-dimensional feature vectors are converted into 1D feature vectors. The 1D feature vectors associated with one type of action are used to train a CHMM, corresponding to the type of action. Moreover, the switching graphical model is designed to switch CHMMs based on a long observation sequence. Besides this, the proposed system is tested using two different datasets: the self-recorded dataset and the IXMAS dataset, for both single-action recognition and continuous-action recognition. In the IXMAS dataset, the proposed method is favorably compared with three other up-to-date methods in single-action recognition.

References

[1] M. Blank et al., "Actions as Space-Time Shapes," *IEEE Int. Conf. Comput. Vis.*, Beijing, China, vol. 2, 2005, pp. 1395–1402.

[2] Y. Ke, R. Sukthankar, and M. Hebert, "Spatio-Temporal Shape and Flow Correlation for Action Recognition," *IEEE CVPR*, 2007.

[3] W. Kim, C. Jung, and C. Kim, "Spatiotemporal Saliency Detection and its Applications in Static and Dynamic Scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 4, Apr. 2011, pp. 446–456.

[4] I. Laptev, "On Space-Time Interest Points," *Int. J. Comput. Vis.*, vol. 64, no. 2–3, Sept. 2005, pp. 107–123.

[5] I. Laptev et al., "Learning Realistic Human Actions from Movies," *Proc. IEEE CVPR*, Anchorage, AK, USA, June 23–28, 2008, pp. 1–8.

[6] S. Kumari and S.K. Mitra, "Human Action Recognition Using DFT," *Proc. NCVPRIPG*, Dec. 15–17, 2011, pp. 239–242.

[7] D.G. Lowe, "Object Recognition from Local Scale-Invariant Features," *IEEE Int. Conf. Comput. Vis.*, Kerkira, Greece, vol. 2, 1999, pp. 1150–1157.

[8] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, 2004, pp. 91–110.

[9] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *IEEE CVPR*, San Diego, CA, USA, vol. 1, June 25, 2005, pp. 886–893.

[10] C. Lin, F. Hsu, and W. Lin, "Recognizing Human Actions Using NWFE-Based Histogram Vectors," *EURASIP J. Advances Signal Proc.*, vol. 2010, no. 9, Feb. 2010.

[11] B.D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," *Int. Joint Conf. Artif. Intell.*, Vancouver, Canada, 1981, pp. 674–679.

[12] J. Shi and C. Tomasi, "Good Features to Track," *IEEE CVPR*, Seattle, WA, USA, June 1994, pp. 593–600.

[13] P. Scovanner, S. Ali, and M. Shah, "A 3-Dimensional SIFT Descriptor and its Application to Action Recognition," *Proc. Int. Conf. Multimedia*, Augsburg, Germany, 2007, pp. 357–360.

[14] W. Lu and J.J. Little, "Simultaneous Tracking and Action Recognition Using the PCA-HOG Descriptor," *Canadian Conf.*

Comput. Robot Vis., June 7–9, 2006, p. 6.

[15] H. Kataoka and Y. Aoki, "Symmetrical Judgment and Improvement of CoHOG Feature Descriptor for Pedestrian Detection," *IAPR Conf. Mach. Vis. Appl*., Nara, Japan, June 13–15, 2011, pp. 536–539.

[16] X. Lu, Q. Liu, and S. Oe, "Recognizing Non-rigid Human Actions Using Joints Tracking in Space-Time," *Int. Conf. ITCC*, Las Vegas, NW, USA, vol. 1, Apr. 5–7, 2004, pp. 620–624.

[17] D. Weinland, E. Boyer, and R. Ronfard, "Action Recognition from Arbitrary Views Using 3D Exemplars," *IEEE Int. Conf. Comput. Vis*., Rio de Janeiro, Brazil, Oct. 14–21, 2007, pp.1–7.

[18] I.N. Junejo et al., "View-Independent Action Recognition from Temporal Self-similarities," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 33, no. 1, Jan. 2011, pp. 172–185.

[19] G. Rogez, J.J. Guerrero, and C. Orrite, "View-Invariant Human Feature Extraction for Video-Surveillance Applications," *IEEE Conf. AVSS*, London, UK, Sept. 5–7, 2007, pp. 324–329.

[20] M. Lee and R. Nevatia, "Body Part Detection for Human Pose Estimation and Tracking," *IEEE Workshop Motion Video Comput*., Austin, TX, USA, Feb. 2007.

[21] M. Lee and R. Nevatia, "Human Pose Tracking in Monocular Sequence Using Multilevel Structured Models," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 31, no. 1, Jan. 2009, pp. 27–38.

[22] W. Gilks, S. Richardson, and D. Spiegelhalter, Markov Chain Monte Carlo in Practice, London, UK: Chapman and Hall, 1996.

[23] S. Zhu, R. Zhang, and Z. Tu, "Integrating Bottom-up/Top-down for Object Recognition by Data Driven Markov Chain Monte Carlo," *IEEE CVPR*, Hilton Head Island, SC, USA, vol. 1, 2000, pp. 738–745.

[24] S. Ke et al., "Real-Time 3D Human Pose Estimation from Monocular View with Applications to Event Detection and Video Gaming," *IEEE Int. Conf. AVSS*, Boston, MA, USA, Aug. 29–Sept. 1, 2010, pp. 489–496.

[25] S. Ke et al., "View-Invariant 3D Human Body Pose Reconstruction Using a Monocular Video Camera," *ACM/IEEE ICDSC*, Ghent, Belgium, Aug. 22–25, 2011, pp. 1–6.

[26] L.W. Campbell et al., "Invariant Features for 3-D Gesture Recognition," *Proc. Int. Conf. Autom. Face Gesture Recogn*., Killington, VT, USA, Oct. 14–16, 1996, pp. 157–162.

[27] M. Muller, T. Roder, and M. Clausen, "Efficient Content-Based Retrieval of Motion Capture Data," *ACM SIGGRAPH*, Los Angeles, CA, USA, vol. 24, no. 3, July 2005, pp. 677–685.

[28] H. Thuc et al., "Human Action Recognition Based on 3D Body Modeling from Monocular Videos," *Frontiers Comput. Vis. Workshop*, Kawasaki, Japan, Feb. 2–4, 2012, pp. 6–13.

[29] H. Thuc, P. Tuan, and J. Hwang, "An Effective 3D Geometric Relational Feature Descriptor for Human Action Recognition," *IEEE Int. Conf. Comput. Commun. Tech. RIVF*, Ho Chi Minh City, Vietnam, Feb. 27–Mar. 1, 2012, pp. 1–6.

[30] L. Rabiner and B. Juang, Fundamentals of Speech Recognition, 1st ed., Upper Saddle River, NJ: Prentice Hall, 1993.

[31] S. Sempena, N.U. Maulidevi, and P.R. Aryan, "Human Action Recognition Using Dynamic Time Warping," *Int. Conf. Electr. Eng. Informat*., Bandung, Indonesia, July 17–19, 2011, pp. 1–5.

[32] V.N. Vapnik, Statistical Learning Theory, New York: Wiley, 1998.

[33] V.N. Vapnik, S.E. Golowich, and A.J. Smola, "Support Vector Method for Function Approximation, Regression Estimation and Signal Processing," In M. C. Mozer, M. I. Jordan, and T. Petsche editors, *Advances in Neural Information Processing Systems* 9, Cambridge, MA: MIT Press, 1997.

[34] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," *Proc. IEEE Int. Conf. Pattern Recogn*., Cambridge, UK, vol. 3, Aug. 23–26, 2004, pp. 32–36.

[35] L. Rabiner and B. Juang, "An Introduction to Hidden Markov Models," *IEEE ASSP Mag*., vol. 3, no. 1, Jan. 1986, pp. 4–16.

[36] J. Yamato, J. Ohya, and K. Ishii, "Recognizing Human Action in Time-Sequential Images Using Hidden Markov Model," *IEEE CVPR*, Champaign, IL, USA, June 15–18, 1992, pp. 379–385.

[37] M. Umeda, "Recognition of Multi-font Printed Chinese Characters," *Proc. IEEE CVPR*, Las Vegas, NV, USA, 1982, pp. 793–796.

[38] H. Thuc et al, "Quasi-Periodic Action Recognition from Monocular Videos via 3D Human Models and Cyclic HMMs," *Int. Conf. ATC*, Hanoi, Vietnam, Oct. 10–12, 2012, pp. 110–113.

[39] J.A. Hartigan and M.A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Appl. Statistics*, New York: Wiley, 1979, pp. 100–108.

[40] J. Bilmes and G. Zweig, "The Graphical Models Toolkit: An Open Source Software System for Speech and Time-Series Processing," *IEEE ICASSP*, Orlando, FL, USA, vol. 4, May 13–17, 2002, pp. 3916–3919.

[41] T. Horprasert, D. Harwood, and L.S. Davis, "A Statistical Approach for Real-Time Robust Background Subtraction and Shadow Detection," *IEEE ICCV, Frame-Rate Workshop*, Greece, Sept. 1999, pp. 1–19.

[42] S.A. Al-Shehri, "A Simple and Novel Method for Skin Detection and Face Locating and Tracking," *Proc. APCHI*, Rotorua, New Zealand, 2004, pp. 1–8.

[43] J. Canny, "A Computational Approach to Edge Detection," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 8, no. 6, Nov. 1986, pp. 679–698.

[44] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach toward Feature Space Analysis," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 24, no. 5, May 2002, pp. 603–619.

[45] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-Based Object Tracking," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 25, no. 5, May 2003, pp. 564–577.

[46] G. Welch and G. Bishop, "An Introduction to the Kalman Filter,"

Technical Report TR 95-041, Department of Computer Science, University of North Carolina at Chapel Hill, 1995.

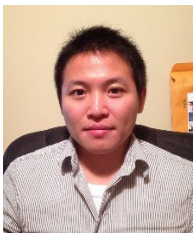
[47] W. Press et al., "Numerical Recipes in C++: The Art of Scientific Computing," *Pearson Education*, 1992.

[48] R.A. Wagner and M.J. Fischer, "The String-to-String Correction Problem," *J. ACM*, vol. 21, no. 1, Jan. 1974, pp. 168–173.

[49] G. Rogez, C. Orrite, and J. Martínez, "A Spatio-Temporal 2D-Models Framework for Human Pose Recovery in Monocular Sequences," *Pattern Recogn.*, vol. 41, no. 9, Sept. 2008, pp. 2926–2944.

[50] G. Rogez et al., "Randomized Trees for Human Pose Detection," *IEEE CPVR*, Anchorage, AK, USA, 2008, pp. 1–8.

[51] L. Sigal, A. Balan, and M.J. Black, "HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion," *Int. J. Comput. Vis.*, vol. 87, no. 1–2, Mar. 2010, pp. 4–27.



Shian-Ru Ke received his BS degree in civil engineering from National Central University, Taoyuan, Taiwan, in 2003 and his MS degree in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 2005. He then received his MS and PhD degrees in electrical engineering from the University of Washington, Seattle, WA, USA, in 2010 and 2014, respectively. He currently works as a senior engineer at Qualcomm, San Diego, CA, USA. His research interests concentrate on computer vision, machine learning, and human behaviour recognition. He is a student member of IEEE.

Hoang Le Uyen Thuc received her BE degree in electronics from Danang University of Technology (DUT), Vietnam, in 1994 and her ME degree in electronics communications from Hanoi University of Technology, Vietnam, in 1997. She is currently working toward her PhD degree under the joint program between Electronics and Telecommunications Engineering (ETE) department, DUT, Vietnam and Information Processing Lab, Electrical Engineering department, University of Washington, USA. She is now a lecturer of ETE department, DUT, Vietnam. Her research interests include signal processing, pattern recognition, and human behaviour understanding.

Jenq-Neng Hwang received his BS and MS degrees, both in electrical engineering, from National Taiwan University, Taipei, Taiwan, in 1981 and 1983, respectively. He then received his PhD degree from the University of Southern California, Los Angeles, CA, USA. In the summer of 1989, he joined the Department of Electrical Engineering of the University of Washington in Seattle, WA, USA, where he has been promoted to full professor since 1999. He is currently the associate chair for Research in the EE Department. He has served as associate editor for IEEE T-SP, T-NN and T-CSVT, T-IP, and Signal Processing Magazine (SPM). He is currently on the editorial board of ETRI, IJDMB, and JSPS journals. He was the program co-chair of ICASSP 1998 and ISCAS 2009. He has been a fellow of IEEE since 2001.

Jang-Hee Yoo received his BSc in physics from Hankuk University of Foreign Studies (HUFS), Seoul, Rep. of Korea, in 1988 and his MSc in computer science from HUFS in 1990. He received his PhD in electronics and computer science from the University of Southampton, UK, in 2004. Since November 1989, he has been at the Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea, as a principal researcher. He has also been an adjunct professor with the Department of Information Security Engineering at the University of Science and Technology, Daejeon, Rep. of Korea. His current research interests include embedded computer vision, biometric systems, human motion analysis, intelligent video surveillance, HCI, and intelligent robots. He is a member of the IEEE and the IEEK.

Kyoung-Ho Choi received his BS and MS degrees in electrical and electronics engineering from Inha University, Incheon, Rep. of Korea, in 1989 and 1991, respectively. He received his PhD degree in electrical engineering from the University of Washington, Seattle, WA, USA, in 2002. Since January 1991, he had been with the Electronics and Telecommunications Research Institute, where he was a leader of the telematics-content research team. He was also a visiting scholar at Cornell University, Ithaca, NY, USA, in 1995. He went on sabbatical leave from the University of Washington from 2011 to 2013. In March 2005, he joined the department of information and electronic engineering at Mokpo National University, Muan, Rep. of Korea. He is a program cochair of the 19th Korea–Japan Joint Workshop of Frontiers of Computer Vision (FCV 2013). His research interests include multimedia signal processing, human behavior recognition, MPEG-HEVC, sensor networks, and audio-to-visual conversion. He is a senior member of the IEEE.