IPSJ Transactions on Computer Vision and Applications Vol.6 103–110 (Sep. 2014)

[DOI: 10.2197/ipsjtcva.6.103]

Research Paper

A Compensation Method of Motion Features with Regression for Deficient Depth Image

Ryo Yumiba[1,a)] Hironobu Fujiyoshi[2,b)]

Received: December 1, 2013, Accepted: July 14, 2014, Released: September 25, 2014

Abstract: In this paper, we propose a method for compensating for motion features that are outside a given viewing angle by using a regression estimate that is based on a correlation between the motion features from human bodies deficient visually, when recognizing the actions of people whose bodies are only partially within the given view. This compensation is good for use in situations where parts of a person's body are partially protruding outside the edges of the viewing angle, and contributes to enlarging the region coverage for action recognition. The motion features and position of the acting person in a depth image are calculated first in the proposed method. Second, the deficit length protruding outside the view angle is calculated, according to the position of the person. Finally, the motion features from the entire body are estimated using a regression estimate from the motion features by selecting the regression coefficients according to the deficit length. The method for improving the effectiveness of the F-measure is confirmed using three kinds of motion features in a fundamental laboratory experiment. We found from the experimental results that the F-measure was improved by more than 12.5% when using motion feature compensation compared to without compensation when the person within the viewing angle cannot actually be seen from the floor to 630 mm above it.

Keywords: action recognition, depth images

1. Introduction

There has been a steady increase in the use of monitoring systems that automatically detect moving or abandoned objects using video recognition technology for video surveillance cameras. Some research has already been done in search of methods for recognizing human actions and comprehending human behavior such as violence and accidents [1], [2] in order to create more advanced systems, and some of them have already been put into practice [3]. The use of these kinds of human behavior comprehension techniques would reduce the monitoring burden of security personnel semantically summarizing video surveillance footage.

Conventional methods of human action recognition primarily use motion features that represent the appearance and motion within the local parts of the videos. However, robustness against outside disturbances is one of the problems needing consideration when penetrating these kinds of motion recognition techniques. Even if we are restricted to indoor environments, we must reduce the amount of influence on the motion features resulting from flickers in the lighting, shadow disturbances, and imaging noise in low illuminant environments and so on when operating systems at many locations.

The use of a depth image sensor looks like a promising way to create a robust human action technique. A depth image sensor is a device that measures the range of every pixel in an image using a specific optical system. There are several kinds of depth image sensors such as Time of Flight (TOF) [4] and Light Coding [5]. The depth information from a depth image sensor has an advantage in that it is less likely to be affected by disturbances from the video camera feeds such as external light or shadows. In addition, using the depth information seems to be a promising way to improve the recognition performance by using it to precisely measure the human position and shape in a 3D space.

However, the narrow viewing angle of a depth image sensor for its specific optical system could be a serious problem when used for the action recognition in indoor environment monitoring. Surely, there are some kinds of depth image sensor with wide viewing angle like laser scanners, but those of common products currently are not wide both horizontally and vertically, which is a required condition for monitoring purpose. A depth image sensor would only capture a limited range of images around a given spot on the floor when placed on the ceiling at tilt angle just like that for conventional surveillance cameras, which is the most acceptable location for general customers. The conventional techniques for recognizing human actions, represented by using the skeleton recognition [6] and action recognition methods [7], [8], [9], [10], [11], [12], [13], by and large assume that the human positions would be restricted to around the center of the viewing angle. On the other hand, for human behavior comprehension targeting voluntary human actions, the positions of humans are hard to restrict and as large an area as possible is required for recognizing a target. For these reasons, the conventional methods used at the settings just mentioned can barely recognize the actions of a person whose body parts are somewhat

[1] Hitachi, Ltd., Hitachi Research Laboratory, Hitachi, Ibaraki 319–1292, Japan

[2] Department of Computer Science, Chubu University, Kasugai, Aichi 487–8501, Japan

[a)] ryo.yumiba.xp@hitachi.com

[b)] hf@cs.chubu.ac.jp

protruding outside the viewing angle, such as when the person starts to leave the viewing angle.

In this paper, we propose a method as an efficient and effective solution that compensates for the motion features by using a regression estimate that is based on a correlation between the motion features of body parts outside the viewing angle and that of full body images. The effects of the body parts outside the viewing angle could be diminished by making the motion features from them closely match the ones from an entire body using the proposed method.

2. Conventional Action Recognition Methods Using Depth Images

We will briefly introduce some conventional action recognition methods that use depth images in this section. We will also explain the adverse effect on the conventional methods from body parts only slightly protruding outside the viewing angle.

The conventional methods can be divided into two kind of approaches, one that extracts the motion features directly from the depth images that represent the appearance and motion of the local parts of the body and the one that preliminarily recognizes human skeletal structures using methods such as Ref. [6] and use the skeletal properties like the joint positions for the motion features. As a representative method of the former approach, Holte et al. extracted motion features from the spatial distribution of the pixel subtraction from the depth images' frame subtraction, and recognized the gestures by categorizing the features using the Edit Distance method [7]. Li et al. extracted motion features from the outline shapes of a human silhouette projected onto some planes, and recognized the fundamental actions such as crouching by using an action state transition model of the feature [8]. Ikemura et al. extracted motion features using the most frequent depth values from small areas in the depth images, and recognized the picking up action from store shelves by categorizing the features using Joint-Boosting [9]. Ni et al. extracted the motion features using the Bag of Features (BOF) method and the moment features from the Motion History Image (MHI), and recognized daily actions like cleaning by categorizing the features using Support Vector Machines (SVM) [10]. Schwarz et al. extracted the spatial coordinates of the corners of a human silhouette extracted with the background subtraction of the depth images as the motion features, and recognized fundamental actions such as waving arms by categorizing the features using a state transition model with a manifold [11].

As representative methods of the latter approach, which extracts motion features from human skeletons, Masood et al. used the distances between the joint positions in adjacent frames, and recognized fundamental actions such as walking by categorizing the features by using the similarities between the representative frames of each action previously defined [12]. Wang et al. extracted the displacement of the joint positions between the frames and the distribution of the pixel values around the skeleton, and recognized the actions accompanied with an instrument such as playing a musical instrument by categorizing the features using Multiple Kernel Learning (MKL) [13].

The conventional methods above assumed that the entire body of the actors was within the viewing angle. Based on the narrow viewing angle of depth image sensors, this assumption is satisfied only when the positions of the actor are around the center of the viewing angle, and not satisfied when the positions are close to the edges of the viewing angle. It is difficult to recognize the actions because of the body parts partially protruding outside the edges of the viewing angle, when the positions of the acting persons are close to the edges of the viewing angle, making the persons partially deficient visually in the images. This problem could be avoided by previously limiting the positions of the humans when targeting a gesture [7] or actions at specified positions [8], but could not be avoided when targeting human actions whose positions could not be previously limited, like that for human behavior comprehension. Considering that in cases surveillance cameras could not choose but to be settled at close range to monitoring persons (e.g., in elevator cars) and positions of surveillance cameras would be acceptable for users who apply depth image sensors to monitoring purpose, solution to this deficit would expand applicable locations of the application.

3. Motion Features Compensation with Regression Estimate

We describe a method for compensating for motion features that are outside the given viewing angle by using a regression estimate for cases when part of a person's body is partially outside the view in depth images. We present an outline of the method in **Fig. 1**. First, the motion features are calculated from the depth images. Simultaneously, the person's position and deficit length according to the position are calculated. Second, regression coefficients according to the deficit length are selected, and regression estimates of the motion features for an entire human body are made from the ones of the partially deficient human body views.

3.1 Deficit Length Calculation to Human Positions

The position of a human within an image is calculated by extracting the person's silhouette from the depth image and by calculating the depth value from the pixels within the silhouette. The deficit length is calculated from this position and a geometric model representing the set position, set angle, and the viewing angle of the depth image sensor.

3.1.1 Human Position Extraction

Human silhouettes in depth images are extracted using background subtraction. This background subtraction method is precise because it uses the depth information [6].

For calculating the position of a human within a given view, the

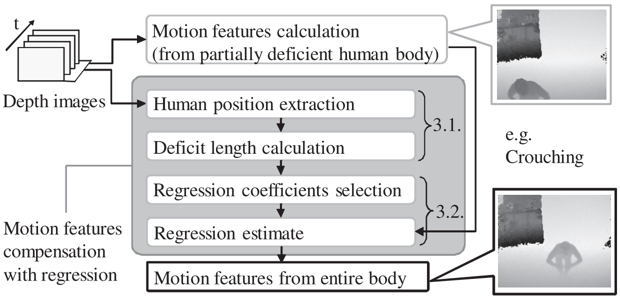
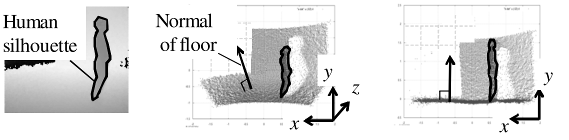

t
Motion features calculation
(from partially deficient human body)
Depth images
Human position extraction
Deficit length calculation
3.1.
Regression coefficients selection
Regression estimate
3.2.
Motion features compensation with regression
Motion features from entire body
e.g.
Crouching

Fig. 1 Outline of motion features compensation.



(a) Depth image (b) Point cloud original (c) Point cloud translated

Fig. 2 Example of coordinate transformation of point cloud.

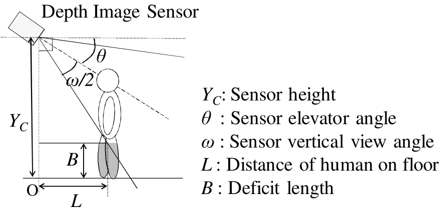

Depth Image Sensor
θ
ω/2
Y_C
B
O
L
Y_C: Sensor height
θ : Sensor elevator angle
ω : Sensor vertical view angle
L : Distance of human on floor
B : Deficit length

Fig. 3 Model of human position and deficit length.

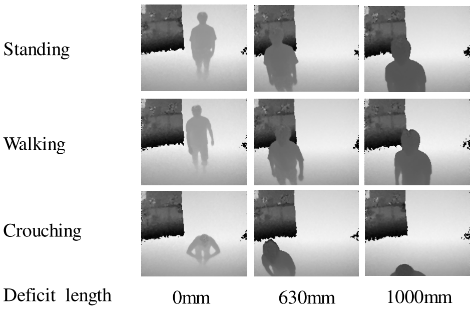


Fig. 4 Examples of depth images at different deficit lengths.

pixels in the silhouette are first converted into a point cloud [14], and the coordinates are transformed so that the normal of the floor is vertical, as in **Fig. 2**. Then, the gravity center for x-z plane of the floor is calculated, which helps to determine the human position.

3.1.2 Deficit Length Calculation

The deficit length B in Eq. (1) is calculated using the distance L on the floor between the depth image sensor and the person as **Fig. 3**, by using a geometric model of the vertical viewing angle in a depth image.

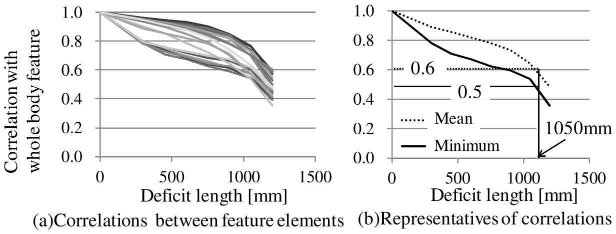
$$B = \max\Big(0, Y_C - L\Big/\tan(90 - \theta - \omega/2)\Big) \tag{1}$$

When $B = 0$, the silhouette is within the viewing angle and not deficient. Examples of deficiencies are in **Fig. 4**.

3.2 Regression Estimate of Motion Features According to Deficit Length

The motion features from an entire human body can be estimated by using the regression coefficients from a regression estimate selected for its deficit length. When the given deficit length is B, B_i is selected first, which is the closest to B among N kinds of deficit length sets $\{B_1, B_2, \ldots, B_N\}$ prepared beforehand. Second, the regression coefficients A_i corresponding to B_i are selected among sets $\{A_1, A_2, \ldots, A_N\}$ that is also prepared beforehand. Last, a regression estimate is done using Eq. (2), so that explanatory variable $\mathbf{x}$ is the motion features from a partially deficient body part view and objective variable $\mathbf{y}$ is that from an entire body view. $\mathbf{c}_i$ is a constant term of regression in Eq. (2).

$$\hat{\mathbf{y}} = A_i\mathbf{x} + \mathbf{c}_i \tag{2}$$



(a)Correlations between feature elements (b)Representatives of correlations

Fig. 5 Motion features' correlations at some deficit lengths.

3.3 Calculation Procedure of Regression Coefficients According to Deficit Length

Every element of the regression coefficients set $\{A_1, A_2, \ldots, A_N\}$ is calculated beforehand using depth image samples corresponding to the deficit length set $\{B_1, B_2, \ldots, B_N\}$. Here, the depth image samples are composed in a pseudo manner from the depth image samples of an entire human body by omitting the parts in the depth image whose heights are less than B_i. Regression coefficients A_i are calculated as shown in Eq. (3) from a sum of the squared deviations $S_{xx,i}$ of the motion features from a view of partially deficit body parts whose deficit length is B_i and $S_{xy,i}$ between the motion features from an entire body view and ones from a deficit body view.

$$A_i = S_{xy,i}\, S_{xx,i}^{-1} \tag{3}$$

Here, $\hat{\mathbf{y}}$ in Eq. (2) is a statistically optimal estimated value in the least-square manner when changes in objective variable $\mathbf{y}$ according to the ones for explanatory variable x are approximated linearly. In this regression estimate, we assume that there is a correlation between the motion features from an image with parts of the subject's body only partially within view and ones from an entire human body within view. For example, in a situation where the legs of a crouching and stretching person are not within full view, this assumption is filled because the upper body movement described by the former is synchronized with the crouching and stretching movement of the entire body described by the latter.

3.4 Validation of Correlation of Motion Features

We validated the correlation between the motion features from a partially deficient human body view and ones from an entire human body view. The correlation coefficients between the motion features at prescribed deficit length and ones from an entire human body image are shown in **Fig. 5** (a), which are calculated from our experimental data later in Section 5.1. The motion features are 18 dimensional ones described later in Section 4.2. In Fig. 5 (b), the average and minimum values of each dimension of the motion features are shown as representative values. The range in correlation coefficients is from 0 to 1, where 0 means no correlation and 1 is a perfect correlation. Linear regression in Eq. (2) can be done precisely when this correlation is high; ideally correlation should be 1 when every element of two variables locates on one line, and regression error increase as correlation declines from 1 to 0. When the deficit length is 0 the entire human body is shown in the image, and as the deficit length increases from zero the parts of the human body that go outside the viewing angle from the ground enlarge. In Fig. 5 (a) every correlation coefficient uniformly decreases as the deficit length increases, but the degree of decrease for each case is gradual. This result shows that there is a correlation between the motion features from a partially deficient human body view and the ones from an entire human body view. In Fig. 5 (b), the minimum correlation is 0.5 and the average one is 0.6 when the deficit length is as much as 1,050 mm. We assume that this correlation should come from co-occurrence of motions between the partially deficient human body view and the entire human body view, like ups and down movement of upper body, and bending and stretching movement of an entire body when a person repeats crouching and standing.

4. Action Recognition Using Motion Features Compensation

An outline of the proposed action recognition method including the motion features compensation is shown in **Fig. 6**. First, a human silhouette is extracted from a depth image and is transformed by the projection. Then, the motion features representing the appearance and motion of the silhouette are calculated. Then, the motion features are compensated for. Finally, the action categories are discriminated from the motion features, and they are filtered using the time series.

4.1 Depth Image Preprocessing

For the preprocessing, the human silhouettes in the image are extracted using background subtraction, as shown in **Fig. 7** (a). Then, a point cloud in the human silhouettes is coordinate transformed so that the floor plane is vertical, and the point cloud is projected onto three planes as shown in Fig. 7 (b).

These projected images correspond to the virtual images from the virtual viewpoints located at infinite distances along the z, y, and x axes respectively, and the posture changes of the person can be described more easily using it. It is particularly effective for posture changes along the optical axis, e.g., that shown in Fig. 7 (b) where a part of an arm is enlarged and its motion feature is easily extracted. In addition, it diminishes the differences in human appearance located at different positions in a depth image.

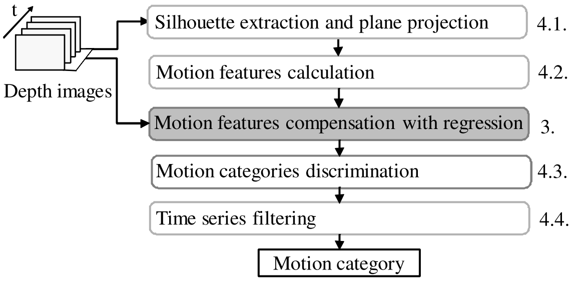
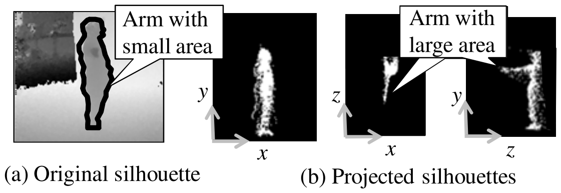

t
Depth images
Silhouette extraction and plane projection
4.1.
Motion features calculation
4.2.
Motion features compensation with regression
3.
Motion categories discrimination
4.3.
Time series filtering
4.4.
Motion category

Fig. 6 Outline of proposed action recognition method.



(a) Original silhouette (b) Projected silhouettes

Fig. 7 Preprocessing of depth images with projection.

4.2 Motion Feature Extraction from Depth Image

Arbitrary motion features could be used that can be used to describe the appearance and motion of human silhouettes in depth images for the proposed method. Here for example, we describe the motion features using MHI in **Fig. 8**, which satisfy the condition above. There are other kinds of features available, such as CHLAC [1] and ST-Patch [2] which are used in the experiment in Section 5.

MHI is a kind of feature that records the history of the motions in grayscale images [15]. For calculating the motion features using MHI, a histogram that describes the orientation of a time slice shape of MHI is at first calculated. Second, amount of the histogram is normalized so that the amount equals the area of the time slice. These motion features describe the direction of appearance and the motion and magnitude of the motion from the moving parts in the depth images.

The motion features using MHI are respectively calculated from three projections. The motion features are actually 18-dimensional when the number of the bin is 6, and they are expanded using the time series [17]. The motion features are 108-dimensional when number of the time slice is 6.

4.3 Action Category Discrimination from Motion Features

Dimensionality reduction using Linear Discernment Analysis (LDA) and the kNN method are used for discriminating action categories from the motion features [17]. The dimensionality reduction is aimed at enhancing the discrimination performance by pruning the dimensions not contributing to the category discrimination. The criterion for the feature dimensions after reduction is a 95% cumulative contribution ratio of the LDA eigen values. The kNN method is used so that the distance is minimized between a given motion feature and the representative vectors shown in Eq. (4). In Eq. (4), $\mathbf{y}$ is the motion features, $\mathbf{v}_{cm}$ is the m-th element of the representative vectors $\{\mathbf{v}_c\} = \{\mathbf{v}_{c_1}, \mathbf{v}_{c_2}, \ldots, \mathbf{v}_{c_M}\}$ belonging to action category $\mathbf{c} \in \{1, 2, \ldots, C\}$.

$$\arg\min_{c} = \| \mathbf{y} - \mathbf{v}_{cm} \| \tag{4}$$

Here, the representative vectors are calculated beforehand from the learning samples using the LBG method [16]. These learning samples are the data from entire human body views.

4.4 Time Series Filtering by Posterior Probability

For time series filtering, action category c is chosen as the latest recognition result that maximizes the posterior probability given by Eq. (5), for diminishing erroneous discriminations from instant turbulence in the motion features.

$$\arg\max_{c} = \prod_{k=1}^{K} P_{B_i}(c|\mathbf{v}_k) \tag{5}$$

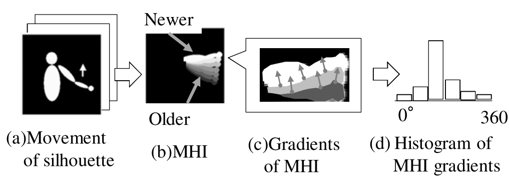


Fig. 8 Motion features using MHI.

In Eq. (5), K is the history length and the length is 18 because we target continuous actions in this paper, $\mathbf{v}_k$ is the representative vector chosen for the k-th in the history when using Eq. (4), and $P_{B_i}(c|\mathbf{v}_k)$ is the posterior probability of action category c around representative vector $\mathbf{v}_k$ when the deficit length is B_i. This posterior probability is calculated beforehand using Eq.(4) for every representative vector and every deficit length $\{B_1, B_2, \ldots, B_N\}$. $S_{c,i}$ in Eq. (6) is the number of training samples whose nearest neighbor is representative vector $\mathbf{v}$ when the compensated values are calculated for all the learning samples whose deficit length is B_i. This posterior probability is the most proper value for every compensated for motion feature of each deficit length.

$$P_{B_i}(c|\mathbf{v}_k) = S_{c,i} \Big/ \sum_{j=1}^{C} S_{j,i} \tag{6}$$

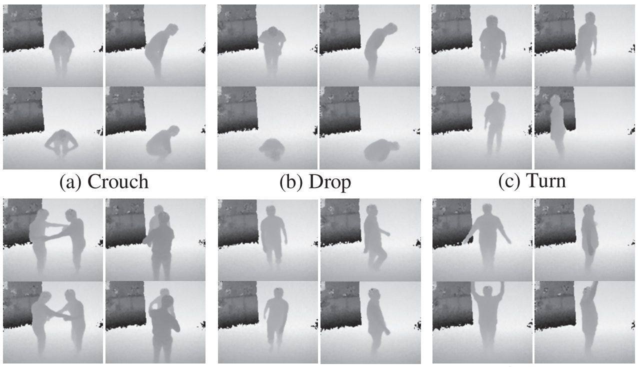
5. Experimental Results

We describe our evaluation of the experimental results using the proposed method.

5.1 Experimental Conditions

The depth sensor used for the experiments was a standard TOF device [4]. The horizontal and vertical viewing angles of the device were 41 and 36 degrees. The device was mounted 2.2 m off the ground and tilted at 25 degrees.

We used six different actions, crouch, drop, turn, jostle, walk, and wave, in our experiments, as shown in **Fig. 9**. For the crouch, the person stretched and bent their knees several times in a kneeling position. For the drop action, the person fell to the ground from an upright position. For the jostle action, two facing people grasp each others' arms and jostled. The person stood and looked back for the turn action. The person marched in the same position for the walk action. For the waving action, the person shook both arms from horizontal to straight up several times. With these actions, the jostle and drop are examples of abnormal actions which are violent and accidental respectively, and the rest actions are examples of daily actions. There are two types of actions that affect the action recognition performance from the direction of the people toward the depth image sensors: frontal and sideways. There are 12 action categories, which are products of six actions and two directions. There were two people for the jostling and only one for the rest. 216 action data in total were taken, which is a combination of the 12 action categories and three action experimenters, and the six positions. Two of the positions is given in the second and third columns in Fig. 4, where the people were only partially within view, and four of them are shown in **Fig. 10**, where the entire person's body was shown.


(a) Crouch (b) Drop (c) Turn

(d) Jostle (e) Walk (f) Wave

Fig. 9 Examples of action data.

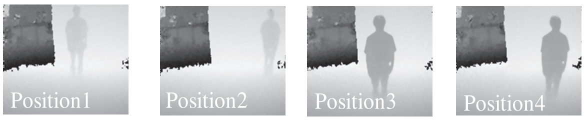
There were three kinds of motion features: MHI, CHLAC [1], and ST-Patch [2]. CHLAC is a 251-dimensional feature that makes comparisons with the binary frame subtraction using 251 local patterns. ST-Patch is a grouping of 6-dimensional features that consists of the temporal and spatial moments of the gradients of grayscale images. Each motion feature is a combined vector of the elements consisting of the three projections shown in Fig. 7. The dimensions of ST-Patch are expanded using the 6 accumulated frames [17].

The evaluation targets are the frame-wise recognition results. The evaluation indicator is an F-measure that is the harmonic average of the recall and precision. The representative indicators are the mean of the indicators of all the action categories.

5.2 Evaluation Results Using Simulated Deficit

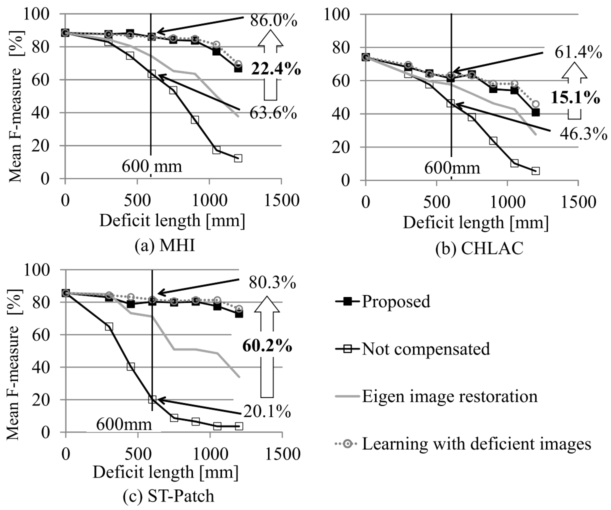
For a fundamental evaluation, only the compensation process in the proposed method is evaluated using synthetically deficient depth images on condition that the set of deficit lengths is dense, the calculation of the deficit length according to the human positions (see Section 3.1) is omitted, and the deficit length of the regression coefficients (see Section 3.2) is set to the deficit length of the synthesized data. For synthesizing deficit depth images, points are omitted whose height is from the ground level to the deficit length. The set of deficit lengths is incremented by 150 mm from 300 mm to 1,200 mm. The data in positions 2 and 3 in Fig. 10 are used for training, and the ones at positions 1 and 4 are used for the evaluation. Here, the affect from the difference in human size and the tilt angle according to the difference in the positions of the human are diminished by the preprocessing discussed in Section 4.1. using the projection transformation from the viewpoints at infinite distances.

Graphs of the F-measure averaging for every action are shown in **Fig. 11**. The ones without motion feature compensation, the ones with motion feature compensation by Ref. [18], and the ones with learning with deficient images are also shown in Fig. 11. Reference [18] is a method for restoring an entire image in an image sequence from a partial image using the eigen image method, and is used in this experiment for restoring the deficient parts of the projected depth images to the x-y and z-y planes shown in Fig. 7. The restored images by Ref. [18] are dealt as deficient less (0 mm deficient) and used for recognizing actions with the method described in Section 4. Learning with deficient images is a method for learning classifiers of action categories from several levels of deficient images in Section 4.3 manner, and recognizing action categories without motion features compensation by



x=0.0m,z=3.0m x=-1.0m,z=3.0m x=0.0m,z=2.0m x=-0.8m,z=2.0m

Fig. 10 Examples of depth images at given positions.



(a) MHI

(b) CHLAC

(c) ST-Patch

Fig. 11 Evaluation results of simulated deficient data.

selecting a classifier according to deficient length of target data.

The compensation in the proposed method was valid because the F-measure when the motion features are compensated for is consistently higher than the one when not compensated for in the range of 0 to 1,200 mm deficient. When choosing to focus on the 600 mm deficient case, improvement of the F-measure is 22.4, 15.1, and 60.2% respectively. Here, 600 mm deficient corresponds to cases when the legs of the person in the given view are rarely shown when considering the average inseam of an adult male is 800 mm. Accepting this 600 mm deficit length would let acting person come closer to the sensor by 29% on the experimental condition: closer by 0.7 m from 2.4 m distance where deficit length is just zero. Here, the result in which the degree of improvement of ST-Path is especially large comes from the appearance elements, (features in one frame) which the remaining two motion features rarely possess. The compensation of the motion features is especially effective for the appearance elements because these elements are constantly deficient (independent of the persons' motion) when parts of the people being viewed protrude out of the viewing angle.

When comparing the proposed method and Ref. [18], the former outperformed the latter because it outperformed the latter in a majority of the ranges excluding the 300 mm deficient case when using ST-Patch. When comparing the proposed method and learning with deficient images, graphs of them both nearly overlap but the latter slightly surpassed, quantitatively surpassed by 1.6% on average of 3 kinds of motion features and 7 levels of deficient length. This difference should be small comparing with 33.2% average on the same condition between the proposed method and when not compensated for. Whereas difference of F-measure is small, the proposed method is superior in term of memory usage. Memory usages of the proposed method and learning with deficient images are calculated by Eqs. (7) and (8).

$$NFDw + CM(D + C)w \tag{7}$$

$$NCM(D + C)w \tag{8}$$

In Eqs. (8) and (9), N is levels of deficit length, F and D are dimension of motion features before and after LDA, w is byte length per data element, C is action categories, M is number of representative vectors for a action category. When using MHI features, N is 7, F is 108, D is 54, w is 8, C is 12, M is 100, and memory usages of the proposed method and learning with deficient images are 0.9 MByte and 4.4 MByte, so the former needs memory less by 80% than the latter. This difference of memory usage is significant for low cost embedded processors, which are used for monitoring purposes widely.

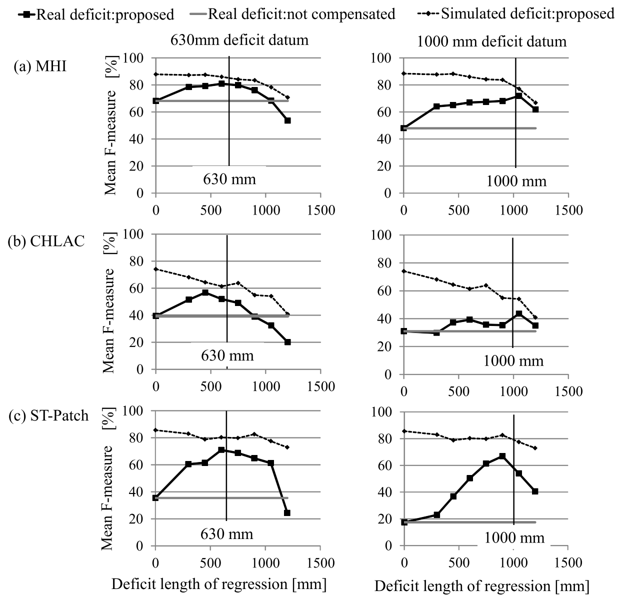


Fig. 12 Evaluation results of actual deficient data.

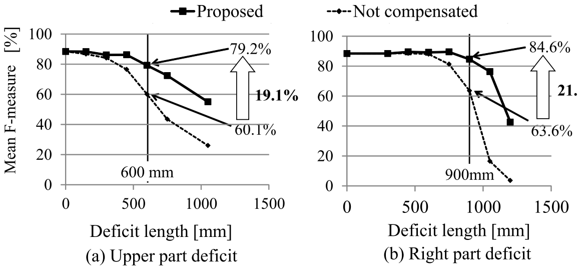
5.3 Evaluation Results Using Actual Deficiency

The proposed method was evaluated for actual deficient depth images using the 2nd and 3rd columns data in Fig. 4. The median deficit length calculated using Eq. (1) was 630 mm for the data in the 2nd column in Fig. 4 and 1,000 mm for the 3rd. The deficit length set of regression coefficients and data for training were equivalent to that in the experiment described in Section 5.2.

Graphs of the experimental results are shown in **Fig. 12**. The horizontal axis of the graphs in this figure represents the deficit length of the regression coefficients (see Section 3.2). There are three graphs in Fig. 12, when the actually deficient data is compensated for, when the actually deficient data is not compensated for, and when the simulated deficient data described in Section 5.2 is compensated for.

First, when comparing the cases when actually deficient data is and is not compensated for, the former's F-measure surpassed the latter's for any motion features when the regression coefficients were used when the deficit length was within a ±150–300 mm gap from the actual deficit length. The deficit length range should be able to tolerate the deficit length estimation for the proposed method. When comparing the cases when the regression coefficients with the nearest deficit length to the actual one, the F-measures of MHI, ST-Patch, and CHLAC are improved by 12.7, 12.5, and 35.5% for the 2nd column data using the regression coefficients for a deficit length of 600 mm, and improved by 23.9, 12.6, and 36.6% for the 3rd column data when using the regression coefficients for a deficit length of 1,050 mm.

Second, when comparing the cases when actual and simulated



(a) Upper part deficit (b) Right part deficit

Fig. 13 Evaluation results when deficit parts are altered.

deficient data are compensated for in Fig. 12, the graphs came close to each other when regression coefficients close to the actual deficit length were used, and the differences in F-measure between them are at most 10.5% when the regression coefficients closest to the actual deficit length were used.

5.4 Evaluation Results with Altered Deficient Positions

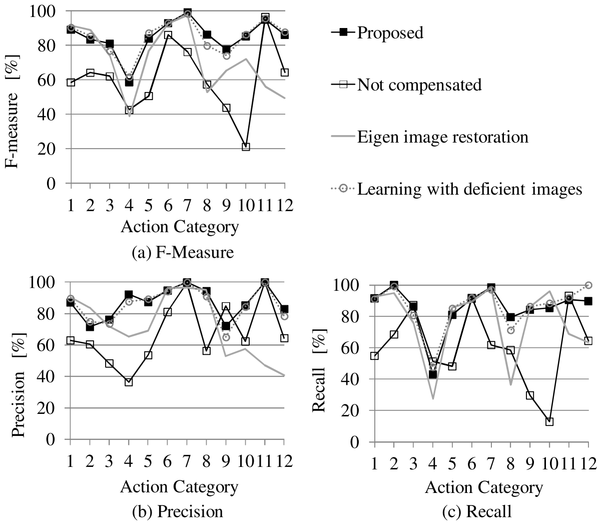
Two cases were used for evaluating whether or not the deficient parts could be altered from lower positions when the deficient positions were in the upper and right positions. The upper deficit position corresponds to cases when the body positions are farther away and the tilt angle of the depth image sensors is deep. The right deficit position corresponds to cases when the body positions are to the left end of the viewing angle of the sensor. The evaluation data are simulated deficiencies like those described in Section 5.2. For the upper deficit position, the deficit length is set to 0 mm to 2,150 mm from the floor, which corresponds to the whole height of the person within whose arms are raised above them. For the right deficit position, the deficit length is set to 0 mm at a position 900 mm to the right of gravity center of the people in the view, which corresponds to the maximum length of an arm lifted horizontally. MHI was applied as the motion features. Considering the manner of motion for each action category in Fig. 9, there should be a correlation between the motion features from the entire body, and the ones from partially deficient upward and to the right body views.

The experimental results are shown in **Fig. 13**. The graphs of the proposed method showed that it outperformed the ones without motion feature compensation when the deficit positions are both upward and to the right, and the F-measure of the former surpassed 19.1% and 21.0% when the deficit lengths were 600 and 900 mm.

6. Category-wise Evaluation Results

A category-wise evaluation was done by using simulated deficient data in Section 5.2 on condition that deficit length is 600 mm, motion features are MHI features, and using four methods evaluated in Section 5.2. Graphs of precision and recall, in addition to F-measure are shown in **Fig. 14**, each of which is average value of each action category. In Fig. 14, the horizontal axis is sequential index of action categories in Fig. 9: 1 is frontal crouch, 2 is sideway crouch, and so forth.

When comparing F-measure in Fig. 14 (a), though there are ups and downs of each graph, the proposed method outperforms when not compensated for and eigen image restoration, because graphs of the proposed method are holistically superior to ones of the others, and surpasses at 11 and 10 points out of total 12 points. The proposed method is nearly equivalent to learning with deficient images, because graphs of the both come close holistically, and the proposed method surpasses at 6 points: half of total 12.



(a) F-Measure

(b) Precision (c) Recall

Fig. 14 Category-wise evaluation results.

In Fig. 14 (b) and (c), precision and recall have same tendency to F-measure, though ups and downs are more intense. The proposed method surpasses to when not compensated for and eigen image restoration, because graphs of the proposed method are holistically superior to ones of the others, and the proposed method surpasses to the others at 18 and 16 points out of total 24 points. The proposed method is nearly equivalent to learning with deficient images, because graphs of the both come close holistically, and the proposed method surpasses at 12 points: half of total 24.

7. Conclusion

We proposed a method that helps to compensate for the motion features that are outside a given viewing angle by using a regression estimate in this paper, for the purpose of enlarging the target area for action recognition when using depth images. We acknowledged the effectiveness of the proposed method from our experimental results. In our future work, we want to improve the recognition performance by using more precise motion features such as the joint positions of human bones.

References

[1] Kobayashi, T. and Otsu, N.: Action and Simultaneous Multiple-Person Identification Using Cubic Higher-Order Local Auto-Correlation, *Proc. ICPR*, pp.741–744 (2004).

[2] Ke, Y., Sukthankar, R. and Hervert, M.: Event Detection in Crowded Videos, *Proc. ICCV*, pp.8–15 (2007).

[3] Seki, M., Hayashi, K., Taniguchi, H., Hashimoto, M. and Sasagawa, K.: Violent Action Detector for Elevator, *Proc. SSII*, pp.E-02-1-6 (2004).

[4] Mesa Imaging, Inc., available from ⟨http://www.mesa-imaging.ch/⟩.

[5] Prime Sense Ltd., available from ⟨http://www.primesense.com/⟩.

[6] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A. and Blake, A.: Real-Time Human Pose Recognition in Parts from a Single Depth Image, *Proc. CVPR* (2011).

[7] Holte, M., Moeslund, T. and Fihl, P.: Fusion of range and intensity information for view invariant gesture recognition, *Proc. Workshop on Time-of-Flight Based Computer Vision* (2008).

[8] Li, W., Zhan, Z. and Liu, Z.: Action Recognition Based on A Bag

Of 3D Points, *Proc. Workshop on CVPR for Human Communicative Behavior Analysis* (2010).

[9] Ikemura, S. and Fujiyoshi, H.: Action Classification by Joint Boosting Using Spatiotemporal and Depth Information, *Trans. IEEJ C*, Vol.130, No.9, pp.1554–1560 (2010).

[10] Ni, B., Wang, G. and Moulin, P.: RGBD-HuDaAct: A Color-Depth Video Database for Human Daily Activity Recognition, *Workshop on Consumer Depth Cameras for Computer Vision* (2011).

[11] Schwarz, L., Mateus, D. and Navab, N.: Manifold learning for ToF-based human body tracking and activity recognition, *Proc. BMVC* (2010).

[12] Masood, S., Ellis, C., Nagaraja, A., Tappen, M., LaViola, J. and Sukthankar, R.: Measuring and Reducing Observational Latency when Recognizing Actions, *Proc. Workshop on Human Computer Interaction* (2011).

[13] Wang, J., Liu, Z., Wu, Y. and Yuan, J.: Mining Actionlet Ensemble for Action Recognition with Depth Cameras, *Proc. CVPR* (2012).

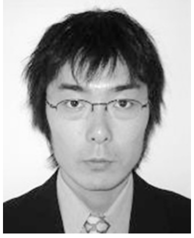
[14] Point Cloud Library, available from ⟨http://pointclouds.org/⟩.

[15] Bradski, G. and Davis, J.: Motion Segmentation and Pose Recognition with Motion History Gradients, *Proc. WACV* (2000).

[16] Linde, Y., Buzo, A. and Gray, R.: An algorithm for vector quantization design. *Trans. IBBE*, Vol.28, No.1, pp.84–94 (1980).

[17] Kazui, M., Miyoshi, M., Muramatsu, S. and Fujiyoshi, H.: Incoherent Motion Detection using a Time-series Gram Matrix Feature, *Proc. ICPR* (2008).

[18] Amano, T., Hiura, S., Yamaguchi, A. and Iguchi, S.: Eigenspace Approach for a Pose Detection with Range Images, *Trans. IEICE D-II*, Vol.J80, No.5, pp.1136–1143 (1997).



Ryo Yumiba received his B.S. and M.S. degrees, both in Electrical Engineering from Kyoto University, Japan, in 1997 and 1999. He is now an employee of Hitachi, Ltd. His research interests include computer vision, video understanding. He is a member of IEICE and IPSJ.

Hironobu Fujiyoshi received his Ph.D. in Electrical Engineering from Chubu University, Japan, in 1997. From 1997 to 2000 he was a post-doctoral fellow at the Robotics Institute of Carnegie Mellon University, Pittsburgh, PA, USA, working on the DARPA Video Surveillance and Monitoring (VSAM) effort and the humanoid vision project for the HONDA Humanoid Robot. He is now a professor of the Department of Computer Science, Chubu University, Japan. From 2005 to 2006, he was a visiting researcher at Robotics Institute, Carnegie Mellon University. His research interests include computer vision, video understanding and pattern recognition. He is a member of IEEE, IEICE, IPSJ, and IEE.

(Communicated by *Hiroshi Ishikawa*)