Viewpoint-independent Action Recognition Method using Depth Image

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

Abstract: In this paper we propose action recognition methods that use depth images for recognizing action independently on the viewpoints. We propose a method for reducing the amount of influence from changes in orientation of the people being observed, while suppressing the required quantity of training samples for dealing with the viewpoint changes between when learning and when recognizing. We first create the three-view drawings expansion by virtually changing the viewpoints within a predetermined range from the training samples when learning, and learn the weak classifier candidates respective to each viewpoint for discriminating the action categories. Then, we learn a strong classifier from these weak classifier candidates and limited number of training samples that is suitable for the viewpoint during recognition. Furthermore, we propose a method for accepting cases when the camera is too close to the people and some part of their body, i.e., an arm or leg, become visually deficient because it protrudes outside the given viewing angle in order to enlarge the region coverage for the action recognition. In this method, arbitrary motion features that are outside a given viewing angle are compensated for before discriminating the action categories by using a regression estimate that is based on a correlation between the motion features of the body parts outside the viewing angle and that of full body images. The experimental results showed that the action categories could be successfully recognized using the proposed methods even under influences from some changes in orientation or visual deficits of the people, when compared to conventional action recognition methods.

1. Introduction

There has been a steady increase in the use of monitoring systems using video recognition technology for video surveillance cameras that automatically detect moving or abandoned objects. Some research has already been done in search of methods for recognizing human actions and comprehending human behavior such as violence and accidents [7][8] for creating more advanced systems, and some of them have already been put into practical use [15]. The use of these kinds of human behavior comprehension techniques can reduce the monitoring burden of security personnel semantically summarizing video surveillance footage.

Any change in appearance of the people being observed due to changes in viewpoints is a serious problem when attempting to incorporate these kinds of action recognition techniques into the actual surveillance systems at many locations. A change in orientation of the people based on the relative angle between the people and the directional view of the camera is the most influent problem caused by changes in viewpoints, and this changes the orientations of the motion features used for action recognition when describing the appearance or motions of the people. Furthermore, when the camera is too close to the people, any protrusion of their bodies outside the viewing angle is also an influent problem, and this damages the motion features for action recognition. In addition, optical disturbances such as flickers in the lighting, shadow disturbances, and imaging noise in low illuminating environments should be desirably considered, which commonly occur depending on the locations of the viewpoints.

We adapted a depth image sensor for the action recognition discussed in this paper for solving these problems caused by changes in viewpoints. 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) and Light Coding. 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 their shape in a 3D space, against any changes in the viewpoints. In addition, 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.

We propose a method for suppressing the influences from changes in orientation of the people being observed by using generative learning, particularly when the viewpoint changes between when learning and when recognizing, for

[1] Hitachi, Ltd., Research & Development Group, Center for Technology Innovation - Controls, Omika 7-1-1, Hitachi, Ibaraki, 319-1292, Japan
[2] Chubu University 1200 Matsumoto-cho, Kasugai, Aichi 487-8501, Japan
[a)] ryo.yumiba.xp@hitachi.com
[b)] hf@cs.chubu.ac.jp

creating a viewpoint-independent human action recognition method. Depth image training samples from a given viewpoint when learning were used to generate three-dimensional data and the motion features from them when the viewpoint is changed, and weak classifiers are learned that can cover a given range of viewpoints within the generated range. Then, we learn the strong classifier from these weak classifiers and the limited number of training samples that is suitable for the given viewpoint during recognition. We could follow the changes in orientation of the people being observed caused by changes in viewpoints, while suppressing the amount of labor necessary for collecting the training samples from the respective viewpoints during recognition using this method.

Furthermore, we propose a method for compensating for the motion features of human bodies that are visually deficient because they are partially outside the given viewing angle for expanding the sites for applying the action recognition technique. This compensation is executed by making a regression estimate that is based on the correlation between the motion features from the visually deficient body parts, when recognizing the actions of people whose bodies are only partially within the given view. Using this method would allow for action recognition technology to be used in such sites as elevator cars or automated teller machine booths, where a depth image sensor must be placed within close proximity to the people being observed.

The former method is discussed in Section 2, the latter method is discussed in Section 3, and the findings and conclusion are described in Section 4.

2. Generative learning for action recognition using three-view drawings expansion of depth images

The conventional methods of human action recognition primarily extract the motion features that represent the motion and appearance within the localized parts of videos, and they use many different statistical learning methods for discriminating the action categories. These motion features describe the direction or magnitude of the human action motion. The motion feature characteristics when an input video is taken using a visible-light camera [2][9][16][18] or a depth image sensor [4][5][10][13][14] are common. Therefore, when a given viewpoint is changed and the orientation of the person/people being observed accordingly changes, it will change the orientations of the motion features and damage the action recognition performance if the change in the motion features is not covered on the feature level or statistical learning level. In this section, we first describe the adverse effect of the change in viewpoint in action recognition and the related works. Then, we describe the proposed method for suppressing the adverse effects by using generative learning for action recognition using the three-view drawings expansion of the depth images.

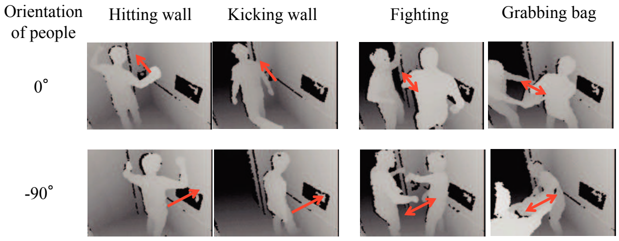
2.1 Problems of conventional methods

We describe how the changes in viewpoint influence the action recognition in this section. Then, we describe the conventional methods for following the changes in viewpoint and their problems.

2.1.1 Influence of change in viewpoint

The orientation of a person changes when their images are taken from several different viewpoints due to changes in orientation from different cameras. These changes in orientation of a person adversely affect the action recognition performance, by changing the motion features that describe the motion or appearance in the video. In addition, when the size of the person changes due to the transition of the distance between the person and camera, any kind of solution is desirable for the change in size.

Here, we include changes in the position or orientation of the person/people being observed in terms of the viewing angle of the same camera, because the change in orientation of the person due to the relative changes in position of the camera and the person change, and the size of the person in the image also changes when the distance between the camera and the person changes.



(a) Actions for which features such as centroid or shape of people can be easily extracted

(b) Actions for which features such as centroid or shape of people can be barely extracted

Fig. 1 Examples of depth images when viewpoint changes.

Examples of the depth images from different viewpoints are shown in Fig. 1. In these depth images, brighter pixels are closer to the depth image sensor, and darker pixels are farther away. In the first and second rows, the examples of the depth images of four kinds of actions are shown in two kinds of orientations: 0° and -90°. We consider a case here in which the classifier is statistically learned from the motion features of each action category obtained from the depth images at 0°, and the actions are recognized from the motion features obtained from the ones a -90°. The action recognition performance is decayed by the difference in the motion features between when learning and recognizing as long as the change in orientation of the person being observed in this case is not absorbed by the motion features or the statistical learning. In addition, the difference in size of the images should be desirably absorbed, because the people are slightly farther at 0° when learning than -90° when recognizing.

2.1.2 Related works

Several methods for using motion features that can follow changes in viewpoint in action recognition by using the

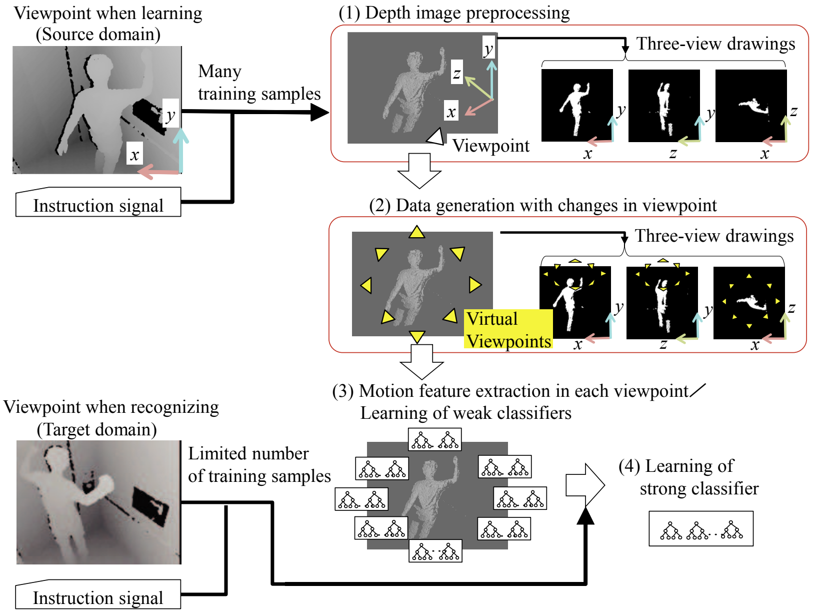


Fig. 2 Approach of proposed method.

depth information from depth images have previously been proposed [4][12][19]. These methods can roughly be divided into two approaches: one is for compensating for the motion features based on the estimated orientation of the person being observed, and the other is for making motion features invariant to the orientation of the person being viewed. Here, the size of the person under observation is simultaneously made invariant using any method because it describes the motion features in three-dimensional coordinates.

As the former approach, Xia et al. proposed a method that estimates the centroid and orientation of a person by using the output of the skeleton recognition in depth images, and extracts the motion features by using a distribution of the joint positions in a polar coordinate ($\phi - \theta$) histogram whose axes corresponds to the vertical and frontal directions of the person [12]. The motion features can follow the changes in viewpoint because the coordinates can follow the change in orientation of the person being observed using this method.

As the latter approach, Holte et al. proposed a method for extracting the invariant motion features from the person orientations by detecting the centroid of the person in the depth images first, by summing up the three-dimensional positions of every pixel in the frame subtraction image of the depth image to a polar coordinate ($\phi - \theta$) histogram. and then translating the histogram into a spherical harmonics function [4]. Weinland et al. also proposed a method for extracting the invariant motion features from the person orientations by first detecting the centroid of the person in the depth images, by calculating the Motion History Volume that records the occurrence time of the changes in the volume data, and translating the data into a Fast Fourier Transform (FFT) spectrum in a cylindrical coordinate system whose origin corresponds to the centroid of the person being observed [19].

In these conventional methods, there is an assumption that these kinds of features can be calculated as an orientation or centroid of the person being observed. However, this assumption is not satisfied when recognizing actions that involve contact between multiple people or ones that involve large changes in posture. Considering the orientation and centroid of the person are required in [12], and the centroid is required in [4] and [19] in advance, these methods probably could not be used for actions such as fighting or robbing someone of their bag, as shown in Fig. 1(b), because contact between multiple people and large changes in posture exist with these actions. These kinds of actions definitely need to be recognized by security systems.

For dealing with the changes in viewpoint, in addition to the conventional methods described above, there could be a method that previously collects training samples from the given viewpoint when recognizing and learns the classifiers for the action recognitions from the respective viewpoints by using the motion features extracted from the training samples and the instruction signals (grand truth of action recognition). However, a lot of labor is required for collecting an adequate amount of training samples from all the respective viewpoints, because there must be enough training samples for a classifier to learn so that it could recognize actions at a high recognition rate.

2.2 Approach of proposed method

The approach of the proposed method is shown in Fig. 2. The training samples are from the viewpoint when learning and are perspectively translated into three-dimensional data (x, y, z) in Step (1); the data is converted into three-view drawings expansion as seen from the virtual viewpoints at an infinite distance. These three-view drawings expansion have an advantage in that they are invariant to the position changes of the person being observed accompanied with the changes in orientation. In Step (2), the three-dimensional data and three-view drawings are generated by comprehensively changing the viewpoints within a predetermined range. The characteristic used here is that the viewpoints of the three-dimensional data can be virtually changed. In Step (3), the motion features are extracted for every viewpoint, and the weak classifiers for action recognition are learned from each viewpoint. Finally, in Step (4), the strong classifier that is suitable for the given viewpoint during recognition is learned from these weak classifiers and limited number of training samples. Here, the training samples in Steps (1) and (4) match the sets of depth images and the instruction signal.

Feature extraction such as the orientation or centroid of the person being observed are not necessary during recognition using the proposed method, and this situation is applicable when there is contact between multiple people or there is a large change in their postures. Furthermore, the labor involved in collecting training samples is reduced by cutting back on the number of training samples for each viewpoint during recognition.

The proposed method stated above is a kind of transfer learning whose source domain is the viewpoint when learning and whose target domain is one when recognizing. The required quantity of learning samples in a target domain can be suppressed by using the learning output of action recognition in a source domain.

In Section 2.3, we describe the generative learning using the three-view drawings expansion of the depth images. In Section 2.4, we describe the action recognition method using the generative learning output. In Section 2.5, we describe the experimental conditions. In Section 2.6, we describe the experimental results. In Section 2.7, we summarize Section 2.

2.3 Generative learning using three-view drawings expansion of depth images

We describe an outline of the learning flow of the proposed method shown in Fig. 3, and detail every step in this section.

The learning flow can be divided into the learning of the weak classifiers in the source domain, as shown in Fig. 3(a), and learning of the strong classifier in a target domain, as shown in Fig. 3(b). In Fig. 3(a), the depth image preprocessing, data generation by changing viewpoints, motion features extraction, and the learning of the weak classifiers are executed in order from the depth images and the action instruction signals in a source domain.

Then, depth image preprocessing and motion features extraction are executed from a limited number of depth images and instruction signals in the target domain, as shown in Fig. 3(b). Then, the strong classifier is optimally learned for the target domain from these motion features and weak classifiers learned in the source domain. In the proposed method, a strong classifier is learned that can follow the changes in viewpoint by learning from a combination of the data in the source domain and the limited number of data in the target domain. We describe every step in Fig. 3.

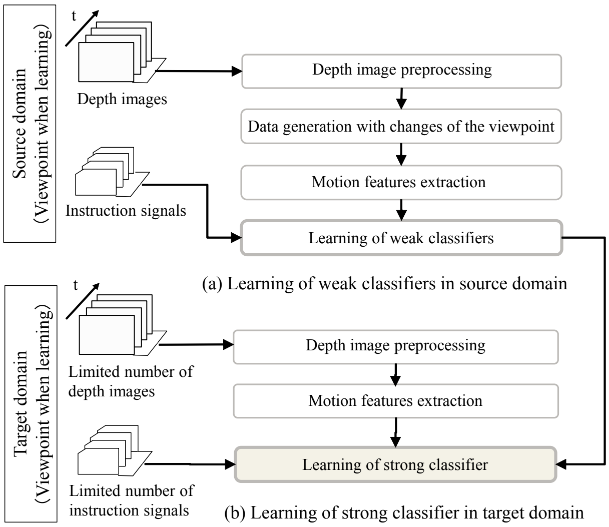
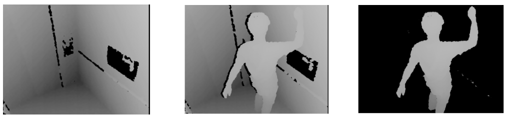

Source domain
(Viewpoint when learning)
t
Depth images
Instruction signals
Depth image preprocessing
Data generation with changes of the viewpoint
Motion features extraction
Learning of weak classifiers
(a) Learning of weak classifiers in source domain
Target domain
(Viewpoint when learning)
t
Limited number of depth images
Limited number of instruction signals
Depth image preprocessing
Motion features extraction
Learning of strong classifier
(b) Learning of strong classifier in target domain

Fig. 3 Learning flow of proposed method when learning.

2.3.1 Depth image preprocessing

In this step, human silhouettes are extracted from a depth image and are converted into three three-view drawings expansion. In this paper, a set of images obtained by projecting three-dimensional data in three directions is referred to as three-view drawings.

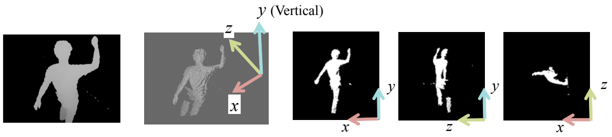
The human silhouettes in a depth image are extracted using background subtraction, as shown in Fig. 4. This background subtraction method is simple but precise because it uses the depth information [17].



(a) Background depth image (b) Input depth image (c) Human silhouette

Fig. 4 Example of background subtraction of depth image.

Then, the human silhouettes in the depth images, as shown in Fig. 5(a), are transformed into a point cloud, as shown in Fig. 5(b), using a perspective transform. Here, the viewpoint of a point cloud, as shown in Fig. 5(b), can be virtually changed to an arbitrary position and direction.



(a) Silhouette in depth image (b) Silhouette in point cloud (c) Silhouette in three-view drawings expansion

Fig. 5 Three-view drawings expansion of depth image.

Then, every point in the human silhouettes is projected onto the $x-y$, $z-y$, and $x-z$ planes from the virtual view points on the z, x, and y axes, which correspond to the depth, left, and vertical directions. The y axis depends on the depression angle of the depth image sensor. Each image of the three-view drawings contains only two-dimensional information, but a set of three-view drawings images can describe the three-dimensional information. In addition, a change in appearance of the people being observed caused by a change in their positions in a depth image can be suppressed using the three-view drawings because their viewpoints are located at an infinite distance. This characteristic is effective when the people change their positions in the images, such as when hitting or kicking a wall, as shown in Fig. 1. It should be noted that the occluded parts of the people are neglected in the three-view drawings.

2.3.2 Data generation with changes in viewpoint

We describe the method for generating data according to the virtual changes in viewpoint in this section by taking into account that the orientation of the people being observed changes in the depth images. The viewpoint of a human silhouette in a depth image can be virtually changed to an arbitrary direction and position, and then the silhouette is transformed into a point cloud. In the proposed method, the viewpoint is virtually changed for rotating the orientation of the people around a y axis (vertical direction). Here, the depth direction z' after changing the viewpoint corresponds to the virtual gaze direction. A point cloud converted from the human silhouette is converted into a three-view drawing expansion after changing the viewpoint using the same manner as described in Section 2.3.1. An example of a coordinate system (x', y, z') when the viewpoint is changed is shown in Fig. 6(a). Examples of three-view drawings that correspond to some given viewpoints are shown in Fig. 6(b).

The three-view drawings generated in this section are the approximate data generated by excluding the self occlusion, when compared to those obtained from depth images with an actual change in viewpoint. Self occlusion here represents when more distant parts of the person being observed are occluded by the closer parts from the viewpoint. Therefore, the rough features of holistic human bodies could be described from the three-view drawings expansion in this section, excluding detailed features like hand movements. For recognizing actions such as those associated with violence or stumbling, which are the recognition targets of the proposed method, the influence of this self occlusion could be estimated as small because these motions are recognized mainly from the movements of holistic human bodies.

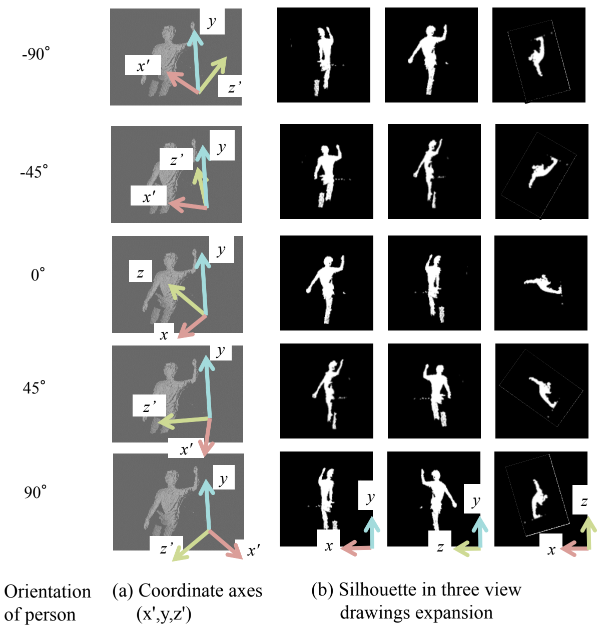

-90°
-45°
0°
45°
90°
y
x'
z'
z
x
Orientation of person
(a) Coordinate axes (x',y,z')
(b) Silhouette in three view drawings expansion

Fig. 6 Example of data generation corresponding to changes in viewpoint.

2.3.3 Motion feature extraction

We apply Motion History Image (MHI) [2] for extracting motion features that describe the appearance and motion of the people being observed in preprocessed depth images as shown in Fig. 7. MHI is a kind of feature that records the history of the motions in grayscale images. A histogram that describes the orientation of a time slice shape of MHI is calculated. Then, the size of the histogram is normalized so that the amount equals the area of the time slice. These motion features describe the direction of the 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 54-dimensional when the number of the bin is 18, and they are expanded using the time series [6]. The motion features are 324- dimensional in total when the time slice is six.

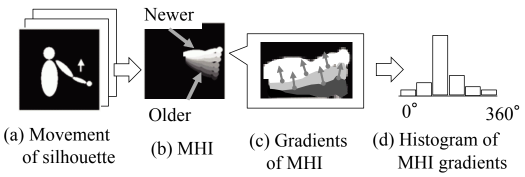


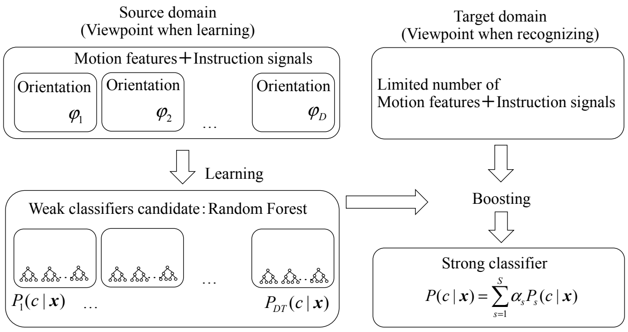
Fig. 7 Motion features using MHI.

2.3.4 Learning process of weak classifiers

The weak classifiers in a source domain are respectively learned from the motion features generated according to the viewpoint. The statistical learning method applied here is Random Forest [3]. Random Forest creates a high generalization performance by using the bootstrap method, which learns multiple classifiers from the training samples and tends to prevent over-training.

The weak classifiers are learned by using Random that is composed of multiple decision trees. Each decision tree in the Random Forest is composed of split and leaf nodes, and each leaf node poses a category posterior, which is obtained from the training samples that arrived within it while learning.

In the proposed method, the Random Forest is respectively learned from the motion features from each viewpoint, and each decision tree in the respective Random Forest is a weak classifier candidate. When the number of viewpoints is D and the number of decision trees is T, the total number of weak classifier candidates is $D \times T$. An outline for learning weak classifiers is shown in Fig. 8(a).


Source domain
(Viewpoint when learning)
Motion features＋Instruction signals
Orientation φ_1
Orientation φ_2
…
Orientation φ_D
Learning
Weak classifiers candidate：Random Forest
$P_1(c|\boldsymbol{x})$ …
…
$P_{DT}(c|\boldsymbol{x})$
Target domain
(Viewpoint when recognizing)
Limited number of
Motion features＋Instruction signals
Boosting
Strong classifier
$P(c|\boldsymbol{x}) = \sum_{s=1}^{S} \alpha_s P_s(c|\boldsymbol{x})$

(a) Learning of weak classifiers

(b) Learning of strong classifier

Fig. 8 Outline of learning weak classifier and strong classifier.

2.3.5 Learning process of strong classifiers

We describe the method for learning the strong classifier in the target domain, as shown in 8(b). First, depth image preprocessing is done to limit the number of training samples in the target domain and the motion features are extracted from the preprocessed data. Then, the strong classifier is learned from the extracted motion features, the instruction signals of the features, and the weak classifier candidates learned as described in Section 2.3.4. In the following paragraphs, we describe the method for learning the strong classifier in a Boosting manner.

A strong classifier, as shown using Eq. (7), is learned by using the learning algorithm in a Boosting manner. This learning algorithm learns the strong classifier from the motion features and instruction signals from a limited number of training samples in the target domain and the weak classifier candidates $\{P_i(c|\boldsymbol{x})_{i=1\ldots TD}\}$, which is a pool of decision trees from the Random Forest learned from the predetermined range of viewpoints. This learning algorithm selects effective decision trees in the target domain from a pool of decision trees in the Random Forest in order to learn what the effective strong classifier is in the target domain.

The strong classifier is learned by sequentially selecting a weak classifier from its pool with any criteria by using the learning algorithm in a Boosting manner. Here, the weights w for each learning sample are taken in for representing the importance of the sample when selecting the next weak classifier. These weights are changed after every step for selecting the weak classifiers, so that w gets smaller for a sample that is correctly classified and gets larger for one that is wrongly classified. These steps are repeated as long as the accuracy rate of the selected weak classifier surpasses $1/C$, which is an inverse of the number of category Cs of the recognition target, and the strong classifier is composed of a set of weak classifiers. An outline of the algorithm for learning the strong classifier is shown in **Algorithm 1**. Considering there are decision trees learned by using data from viewpoints that are virtually changed in the pool of the decision tree $\{P_i(c|\boldsymbol{x})_{i=1\ldots TD}\}$, an accurate strong classifier could be learned from a limited number of samples in the target domain by selecting the optimal weak classifiers in the target domain using this algorithm.

Algorithm 1: Learn strong classifier in Boosting manner.

```
1. Input: $J$ pieces of training samples $\{\boldsymbol{x}_1, y_1\} \ldots \{\boldsymbol{x}_J, y_J\}$ are collected.
   $\boldsymbol{x}$ is the motion features, $y_i \in \{1, \ldots, C\}$ is the class label (instruction signal).

2. Initialize: weights $w$ of the learning samples are initialized.
   $w_{j,1} = 1/J$    (1)
   Coefficients of weak classifiers $\{\alpha_1, \ldots, \alpha_S\}$ are initialized as zero 0.

3. Learning:
 For $s = 1, \ldots, S$    // Learning round
     · From all the weak classifier candidates $\epsilon_s$selects a classifier $P_s(c|\boldsymbol{x})$ whose error ratio is minimal.
         · Calculation of error ratio $\epsilon_s$
   $\epsilon_s = \sum_{j \in \arg\max_c P_s(c|\boldsymbol{x}_j) \neq y_j}^{J} w_{j,s}$    (2)
         · Calculate weight $\alpha_s$ of weak classifier $P_s$
   $\alpha_s = \frac{1}{2} \log \left( \frac{(C-1)(1-\epsilon_s)}{\epsilon_s} \right)$    (3)
     · Determination of end of learning
   if $\alpha_s < 0$ then break    (4)
     · Update weights of training samples $w$.
   $w'_{j,s+1} = \begin{cases} w_{j,s} \exp(+\alpha_s) & \text{if } \arg\max_c P_s(c|\boldsymbol{x}_j) \neq y_j \\ w_{j,s} \exp(-\alpha_s) & \text{otherwise} \end{cases}$    (5)
     · Normalization of weights of training sample $w$.
   $w_{j,s+1} = \frac{w'_{j,s+1}}{\sum_{j=1}^{J} w'_{j,s+1}}$    (6)
   End for

4. Output: strong classifier
   $P(c|\boldsymbol{x}) = \sum_{s=1}^{S} \alpha_s P_s(c|\boldsymbol{x})$    (7)
```

2.4 Action recognition using output of generative learning

In this section, we describe an outline of the action recognition, as shown in Fig. 9, and the details for every step of the algorithm.

In Fig. 9, the depth images during recognition are preprocessed first, and then the motion features are extracted. After that, the action categories are discriminated by using the strong classifier learned in Section 2.3.5. Finally, the action categories are filtered using the time series. In the following, we describe the details for every step excluding the depth image preprocessing and motion features extraction, which are the same as when learning.

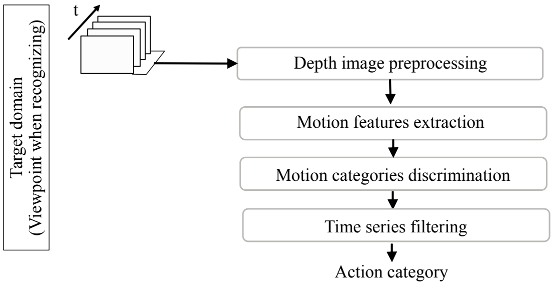

Target domain
(Viewpoint when recognizing)
t
Depth image preprocessing
Motion features extraction
Motion categories discrimination
Time series filtering
Action category

Fig. 9 Outline of action recognition of proposed method.

2.4.1 Action category discrimination

An outline of the action category discrimination is shown in Fig. 10. First, the motion features $\boldsymbol{x}$ during recognition are input into each decision tree. In each decision tree, the leaf node corresponding to the motion features $\boldsymbol{x}$ is searched for by tracing the nodes from the top to the leaf using the splitting functions learned during the learning process. Then, the action category posterior probability of the strong classifier is calculated from the weighted sum of coefficient α_s and the posterior probability $P_s(c|\boldsymbol{x})$ at each leaf node, as shown in Fig. 10. Finally, an action category is selected whose posterior probability $P(c|\boldsymbol{x})$ is maximal, as shown in Eq. (8).

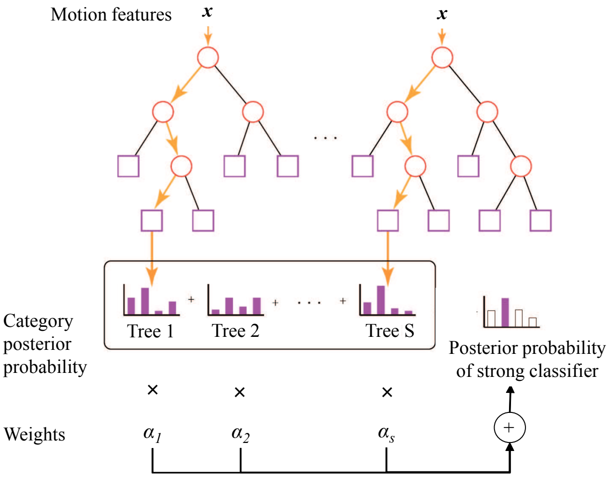


Fig. 10 Outline of action recognition using strong classifier made by Random Forest.

$$\arg\max_c P(c|\boldsymbol{x}) \left(= \arg\max_c \sum_{s=1}^{S} \alpha_s P_s(c|\boldsymbol{x}) \right) \quad (8)$$

2.4.2 Time series filtering using posterior probability

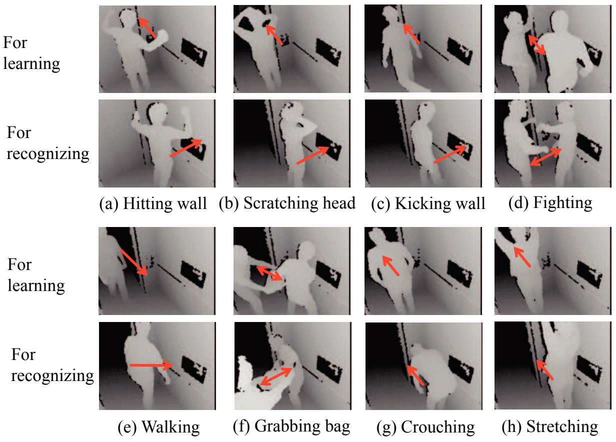
For the time series filtering, action category c is chosen as the latest recognition result that maximizes the posterior probability given by Eq. (9), which is calculated from the motion features in the latest K frames $\{\boldsymbol{x}_k\}_{k=\{1...K\}}$.

$$\arg\max_c \prod_{k=1}^{K} P(c|\boldsymbol{x}_k) \quad (9)$$

2.5 Experimental conditions

We describe the experimental conditions for the evaluation experiment in this section. The depth sensor used for the experiments was a Kinect$^{\mathrm{TM}}$ device from Microsoft Co., Ltd. This depth sensor was place on the ceiling and tilted downward. The evaluation data obtained using this sensor include the people being observed who face several different directions and that do a predetermined category of actions.

We used eight kinds of actions, as shown in Fig. 11. For the hitting wall action, the person continues to hit a wall in front of them with their hands. For the scratching the head action, the person continues to scratch their head. For the kick wall action, the person continues to kick the wall directly in front of them with their feet. For the fighting action, two people face each other and exchange blows. For the walking action, the person comes in through the entrance and turns back. For the grabbing a bag action, two people face each other and one of them grabs and tries to take the other person ' s bag. For the crouching action, the person crouches to the ground from an upright posture, and continues crouching. For the stretching action, the person stretches their arms upward, and then let the arms down. We took the data from the viewpoint when learning under the condition in which the people being observed were around an entrance and their orientations were basically toward it (0° in Fig. 1). We took data from the viewpoint when testing under the condition that the people were close to the right wall in the depth image and their orientations were basically toward this wall (−90° in Fig. 1). However, the orientations drifted sometimes during the fighting and grabbing bag actions in which the movements were intense, and the orientations of the people were both toward the entrance for the actions of walking, crouching and stretching. Every action was done by the same three people. There were 48 total data samples, which were the products for both the learning and testing, eight kinds of actions, and three people. There were a data total of 1,235 frames for the learning,



(a) Hitting wall (b) Scratching head (c) Kicking wall (d) Fighting
(e) Walking (f) Grabbing bag (g) Crouching (h) Stretching

Fig. 11 Examples of action data.

and 1,242 for the testing.

The parameters in the algorithm are set by conducting the following preliminary experiment. For the generation of data in Section 2.3.2, the range of orientations of the people being observed was $-90^\circ \sim +90^\circ$, and their interval was 5°. The range of orientations above is determined because the orientations differ by about 90° between when learning and testing, as shown in Fig. 11. For the parameters of the Random Forest for learning the weak classifiers described in Section 2.3.4, there were 100 trees T, the depth of a tree D was 10, the dimension of the motion features was 324 (as described in Section 2.3.3), there were 19 feature selections, 50 selecting thresholds, and the number of samples for a subset was set to 75% of the number of learning samples. The data in the source domain equals of the number of learning samples. There were 15 weak classifier selections. The length of time series K was 7 for the time series filtering of the action categories.

The data for the learning samples from the viewpoint when recognizing varied by the parameter τ against the overall data for testing. Randomly selected τ% data from the testing data was used as the learning samples in the target domain. The remaining $100 - \tau$% is used for evaluating the action recognition performance.

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

2.6 Experimental results

We describe our evaluation of the experimental results using the proposed method. The graphs of the F-measure for the proposed method and four other methods mentioned below are shown in Fig. 12. The horizontal axis of the graphs is τ, which is the percentage of the target domain data out of the total amount of testing data.

- Method 1: Learn the Random Forest from the target domain data.
- Method 2: Learn the Random Forest from the target data and source domain data.
- Method 3: Learn the Random Forest from the target data, source domain data, and generated data from the changed viewpoints.
- Method 4: Learn the Random Forest from the human silhouettes in the depth images.

The Random Forest is learned from the predetermined data using Methods 1-4, and the action category is discriminated by using the Random Forest learned and the methods shown in Fig. 9. Method 4 is equivalent to the method excluding depth image preprocessing in Section 2.3.1 and data generation with changes in viewpoint discussed in Section 2.3.2 from the proposed method. In Method 4, the depth values of the depth images are treated as gray scale values of camera images.

When comparing the proposed method and Method 1, the former surpasses the latter in all ranges in the graph. The F-measure difference between them is large when τ is small, gets smaller as τ gets larger, and they are almost equivalent when τ is 15%. When comparing the proposed method and Method 2, the former surpasses the latter when $\tau < 10$%, and the former gets inferior when $\tau \leq 10$%. These results show that the proposed method performs quite well especially when the quantity of data is small in the target domain when compared to Methods 1 and 2, which follow the changes in orientation of the people being observed by adding training samples to the target domain.

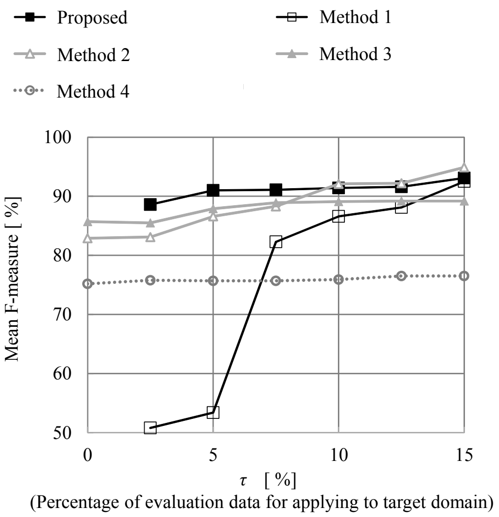


Fig. 12 Graphs of experimental results.

Table 1 F-measure comparison when $\tau = 5$%.

Method	Recall [%]	Precision [%]	F-measure[%]
Proposed	90.7	91.3	91.0
Method 1	52.5	54.2	53.4
Method 2	84.4	88.9	86.6
Method 3	87.2	88.5	87.9
Method 4	75.6	75.9	75.7

A comparison of the F-measure of each method when τ is the representative value 5% is classified in Table 1. In this table, the proposed method surpasses all four method by 37.6%, 4.4%, 3.1%, and 15.3%. The F-measure difference of 3.1% between the proposed method and Method 3 indicates the effect of the strong classifier learned in the Boosting manner (Section 2.3.5). The F-measure difference of 1.3% between Method 3 and 2 indicates the effect of the data generation with the changes in viewpoint (Section 2.3.2). The F-measure difference of 10.9% between Method 2 and 4 indicates the effect of the depth image processing using three-view drawings expansion (Section 2.3.1).

We consider the meaning of $\tau = 5$% in Table 1. When comparing the 91.0% F-measure of the proposed method in Table 1 and 82.9% F-measure when $\tau = 0$% of Method 2 in Fig. 12, which learns the Random Forest from only the data in the source domain, the former surpasses the latter by 8.1%. When considering that when $\tau = 5\%$ the proposed method surpasses Method 1, which learns the Random Forest from the data in the source and target domains, by as much as 37.6%, the proposed method improved the F-measure by 8.1% when the orientations of the people being observed changed by 90° while significantly suppressing the quantity of the training samples in the target domain. This F-measure improvement could be achieved by generating data with changes in viewpoints (difference between Methods 2 and 3) and learning of the strong classifier in the Boosting manner (difference between the proposed method and Method 3), when we take the evaluation results of the components mentioned above into consideration.

2.7 Summary

We proposed a method for suppressing the influence of the changes in orientation of the people being observed in this section for cases in which the viewpoint changes between when learning and when recognizing. In the proposed method, we first generate three-view drawings expansion by virtually changing the viewpoints to within a predetermined range from the training samples from the viewpoint when learning, and learn the weak classifier candidates respective to each viewpoint. Then, we learn the strong classifier from these weak classifier candidates and limited number of training samples that is suitable for the viewpoint when recognizing. In the proposed method, we acknowledged the performance of the proposed method and its components depth image processing using three-view drawings expansion, data generation with changes in viewpoint, and the strong classifier learned in the Boosting manner.

3. Compensation Method of Motion Features with Regression for Deficient Depth Image

In Section 2, we describe a method for suppressing the influence of the changes in orientation change of the people being observed for cases when the viewpoint changes between when learning and when recognizing. However, this method could not be used when a depth image sensor is too close to the people, because parts of their bodies are partially protruding outside the viewing angle of the depth image sensor. The assumption for this method is not satisfied if the entire bodies of the people are not inside the viewing angle. In this section, we first describe the adverse effect of these protrusions on the action recognition. Then, we describe the proposed method for suppressing the adverse effect by compensating for the motion features.

3.1 Problem of conventional methods

We will also explain the adverse effect from body parts only slightly protruding outside a given viewing angle on the conventional methods. This protrusion will occur more frequently when a depth image sensor is used for monitoring purposes instead of a surveillance camera, because the viewing angle of the former is narrower than that of the later for the formers' specific optical system, e.g., the range of the horizontal viewing angle of the former is ordinarily 40∼70° while that of the later is ordinarily 40∼110°.

There is an assumption that the entire bodies of the people being observed are inside the given viewing angle, for most of the conventional methods used for recognizing actions, including the method described in Section 2. This assumption could be comparatively easily satisfied if the actions occur in the center of the given viewing angle of a depth image sensor, but would barely be satisfied around the edge of the given viewing angle of the sensor. When parts of the bodies of the people are visually partially deficient because they are protruding outside the given viewing angle, the motion features become partially deficient, and the ability of the action recognition when using these features will seriously decline. This problem could be avoided by limiting the positioning of the people being observed in advance when targeting a specific gesture [4] or actions within the specified positions [5], but could not be avoided when targeting human actions whose positions could not be limited in advance, like that for human behavior comprehension.

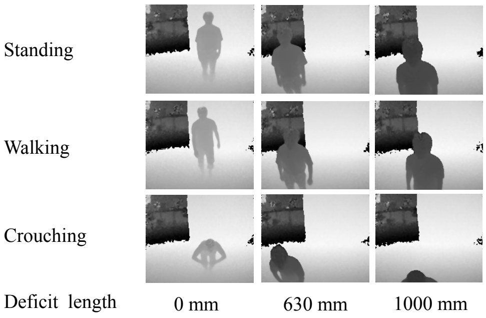


Fig. 13 Examples of depth images at different deficit lengths.

Examples of depth images are shown in Fig. 13, and the bottom part of the people being observed in some of the images is visually deficient. In the second and third columns in this figure, the human bodies from the legs down or the lower stomachs down are visually deficient because they protrude outside the given viewing angle. These visually deficit parts of the people occurs when a depth image sensor is too close to the people and thus they are positioned closer in the images. Considering that in cases where surveillance cameras had to be placed within close range to the monitored people (e.g., in elevator cars or automated teller machine booths) and the positions of the surveillance cameras would be acceptable for users who use depth image sensors for monitoring purposes, a solution to this deficit would be necessary for expanding the applicable locations of the monitoring technology.

3.2 Approach of proposed method

We discuss the approach of the proposed method in Section 3. We used an approach for diminishing the affect of

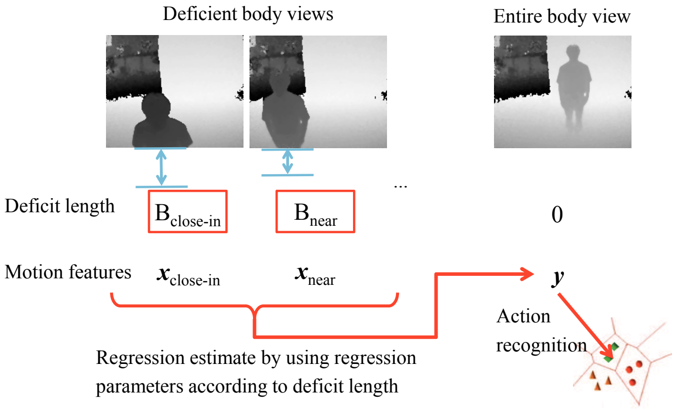


Fig. 14 Approach of proposed method.

body protrusions outside the given viewing angle, by creating the motion features from human bodies that are visually deficient that are close to the ones from entire body views, as show in Fig. 14. We propose a method for compensating for the 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 that are visually deficient, when recognizing the actions of people whose bodies are only partially within the given view. For the regression estimate, we choose regression coefficients according to the degree of deficit, which depends on a position of the person being observed. This regression estimate approach is effective and more efficient than a method for learning the classifier for each degree of deficit (discussed later in Section 3.6.1).

In Section 3.3 that follows, we describe the motion features compensation using the regression estimate. In Section 3.4, we describe an action recognition method using the compensated for motion features. In Section 3.5, we describe the experimental conditions. In Section 3.6, we describe the experimental results. In Section 3.7, we summarize Section 3.

3.3 Motion features compensation with regression estimate

We describe a method for compensating for the motion features that are outside the given viewing angle by using a regression estimate for cases when parts of a person ' s body are partially outside the view in depth images. We present an outline of the method in Fig. 15. First, the motion features are calculated from the depth images. Simultaneously, the person ' s position and deficit length according to their position are calculated. Second, the regression coefficients according to the deficit length are selected, and the regression estimates of the motion features for an entire human body are made from the ones of the partially deficient human body views.

3.3.1 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.

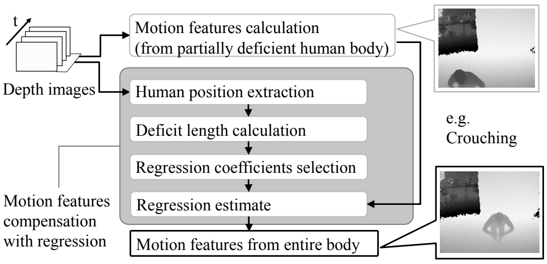

t
Depth images
Motion features calculation
(from partially deficient human body)
Human position extraction
Deficit length calculation
Regression coefficients selection
Regression estimate
Motion features compensation with regression
Motion features from entire body
e.g.
Crouching

Fig. 15 Outline of motion features compensation.

3.3.1.1 Human position extraction

Human silhouettes in the depth images are extracted using background subtraction. This background subtraction method is precise because it uses the depth information [17].

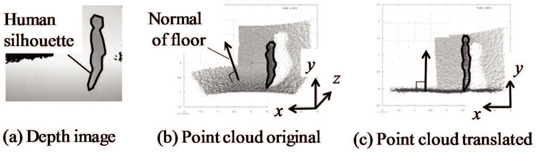


Fig. 16 Example of coordinate transformation of point cloud.

For calculating the position of a human within a given view, the pixels in the silhouette are first converted into a point cloud, and the coordinates are transformed so that the normal of the floor is vertical, as in Fig. 16. Then, the centroid for the $x - z$ plane of the floor is calculated, which helps to determine the human position. Here the x, y, z axes correspond to the left, upper, and depth directions.

3.3.1.2 Deficit length calculation

The deficit length B in Eq. (10) is calculated using the distance L on the floor between the depth image sensor and the person, as shown in Fig. 17, by using a geometric model of the vertical viewing angle in a depth image. When $B = 0$, the silhouette is within the viewing angle and is not deficient.

$$B = \max\Big(0, Y_C - L\Big/\tan(90-\theta-\omega/2)\Big) \qquad (10)$$

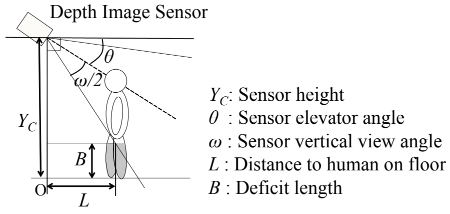

Depth Image Sensor
θ
ω/2
Y_C
B
O
L
Y_C: Sensor height
θ : Sensor elevator angle
ω : Sensor vertical view angle
L : Distance to human on floor
B : Deficit length

Fig. 17 Model of human position and deficit length.

3.3.2 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 that is selected for its deficit length. When the given deficit length is B, B_i is selected first using Eq. (11), which is the closest to B among N kinds of deficit length sets $\{B_1, B_2, \dots B_N\}$ prepared beforehand.

$$\underset{i\in\{1,2,\dots N\}}{\arg\min} \parallel B - B_i \parallel \qquad (11)$$

Second, the regression coefficients A_i corresponding to B_i are selected among the sets $\{A_1, A_2, \dots A_N\}$ that is also prepared beforehand. Last, a regression estimate is done using Eq. (12), so that the explanatory variable $\boldsymbol{x}$ is the motion features from a partially deficient body part view and objective variable $\boldsymbol{y}$ is that from an entire body view. $\boldsymbol{c}_i$ is a constant term for the regression in Eq. (12).

$$\hat{\boldsymbol{y}} = A_i\boldsymbol{x} + \boldsymbol{c}_i \qquad (12)$$

3.3.3 Calculation procedure of regression coefficients according to deficit length

Every element in the regression coefficients set $\{A_1, A_2, \dots A_N\}$ is calculated beforehand using the depth image samples corresponding to the deficit length set $\{B_1, B_2, \dots B_N\}$. 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. The regression coefficients A_i are calculated as shown in Eq. (13) from a sum of the squared deviations $S_{xx,i}$ of the motion features from a view of the 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.

$$A_i = S_{xy,i}\, S_{xx,i}^{-1} \qquad (13)$$

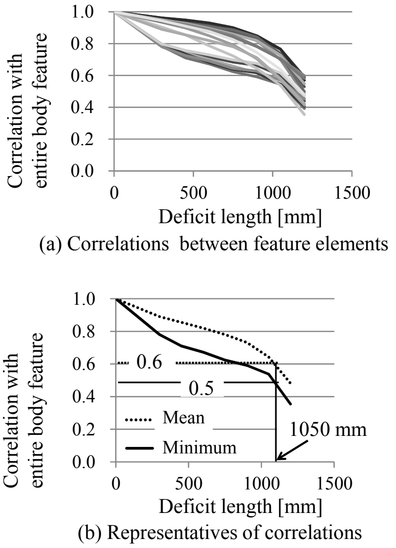
The constant term is calculated by using Eq. (14), which represents a formula for the constant term of the regression. In Eq. (14) $\boldsymbol{\mu}_{x,i,}$ and $\boldsymbol{\mu}_y$ are the mean of the motion features whose deficit lengths are B_i and 0.

$$\boldsymbol{c}_i = \boldsymbol{\mu}_y - A_i\boldsymbol{\mu}_{x.i} \qquad (14)$$

Here, $\hat{\boldsymbol{y}}$ in Eq. (12) is a statistically optimal estimated value in the least-square manner when the changes in the objective variable $\boldsymbol{y}$ according to the ones for explanatory variable $\boldsymbol{x}$ are linearly approximated. In this regression estimate, we assume that there is a correlation between the motion features from an image in which parts of the subject ’ s body are 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.

3.3.4 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 lengths and ones from an entire human body image are shown in the graph in Fig. 18(a), which are calculated from our experimental data discussed in Section 3.5. The motion features are 18-dimensional ones described in Section 2.3.3. In Fig. 18(b), the average and minimum values of each dimension of the motion features are shown as representative values. The range in the correlation coefficients is from 0 to 1, where 0 means there is no correlation and 1 is a perfect correlation. The linear regression in Eq. (12) can be precisely done when this correlation is high; the ideal correlation should be 1 when every element of the two variables locates on one line, and the regression error increases as the 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. Every correlation coefficient in Fig. 18(a) uniformly decreases as the deficit length increases, but the degree of decrease in each case is gradual. This result shows that there is a correlation between the motion features from a partially deficient



(a) Correlations between feature elements

(b) Representatives of correlations

Fig. 18 Motion features’ correlations at some deficit lengths.

human body view and the ones from an entire human body view. In Fig. 18(b), the minimum correlation is 0.5 and the average one is 0.6 when the deficit length is as much as 1050 mm. We assume that this correlation should come from the co-occurrence of motions between the partially deficient human body view and the entire human body view, like for the up and down movements of the upper body and for the bending and stretching movements of the entire body when a person repeats crouching and standing.

3.4 Action recognition using motion features compensation

An outline of the proposed action recognition method including the motion features compensation is shown in Fig. 19. 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.

In these steps, the depth image preprocessing is the same as that described in Section 2.3.1, and the motion feature extraction is the same as that described in Section 2.3.3 although the dimensions of the features are different. In this section, there are a total of 108 dimensions of the motion features, which is a product of the 18 MHI bins, 3 planes of three-view drawings, and 6 time slices. We also describe the detail of the action category discrimination and time series filter.

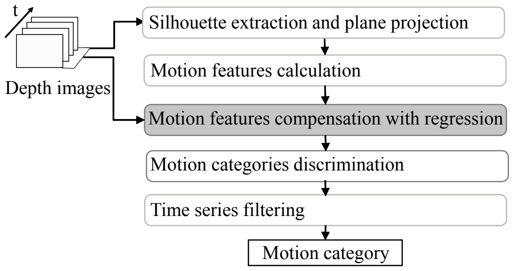

t
Depth images
Silhouette extraction and plane projection
Motion features calculation
Motion features compensation with regression
Motion categories discrimination
Time series filtering
Motion category

Fig. 19 Outline of proposed action recognition method.

3.4.1 Action category discrimination

Dimensionality reduction using Linear Discernment Analysis (LDA) and the kNN method are used for discriminating the action categories from the motion features [6]. The dimensionality reduction aims 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. (15). In Eq. (15), $\boldsymbol{y}$ is the motion features, $\boldsymbol{v}_{cm}$ is the m-th element of the representative vectors $\{\boldsymbol{v}_c\} = \{\boldsymbol{v}_{c_1}, \boldsymbol{v}_{c_2}, \dots \boldsymbol{v}_{c_M}\}$ belonging to action category $\boldsymbol{c} \in \{1, 2, \dots C\}$.

$$\arg\min_{c} \| \boldsymbol{y} - \boldsymbol{v}_{cm} \| \tag{15}$$

The representative vectors here are calculated beforehand from the learning samples using the LBG method [11]. These learning samples are the data from entire human body views.

3.4.2 Time series filtering using posterior probability

Action category c is chosen as the latest recognition result for the time series filtering that maximizes the posterior probability given by Eq. (16), for diminishing the number of erroneous discriminations from instant turbulence in the motion features.

$$\arg\max_{c} \prod_{k=1}^{K} P_{B_i}(c|\boldsymbol{v}_k) \tag{16}$$

$$P_{B_i}(c|\boldsymbol{v}_k) = \frac{S_{c,i}}{\sum_{j=1}^{C} S_{j,i}} \tag{17}$$

In Eq. (16), K is the history length, which is 18 because we target continuous actions in this paper, $\boldsymbol{v}_k$ is the representative vector chosen for the k-th in the history when using Eq. (15), and $P_{B_i}(c|\boldsymbol{v}_k)$ is the posterior probability of action category c around representative vector $\boldsymbol{v}_k$ when the deficit length is B_i. This posterior probability is calculated beforehand using Eq. (15) for every representative vector and every deficit length $\{B_1, B_2, \dots B_N\}$. $S_{c,i}$ in Eq. (17) is the number of training samples whose nearest neighbor is representative vector $\boldsymbol{v}$ when the compensated for 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.

3.5 Experimental conditions

We describe our evaluation of the experimental results using the proposed method. We used a standard TOF device, which is called the SR 4000 provided from Mesa Imaging AG, as the depth sensor used for the experiments. The horizontal and vertical viewing angles of the device were 41° and 36°. The device was mounted 2.2 m off the ground and tilted at 25°.

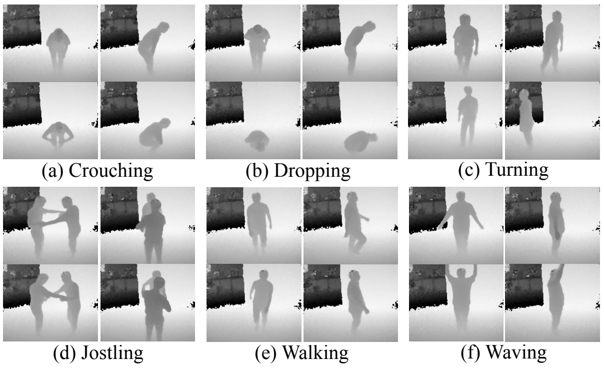

(a) Crouching (b) Dropping (c) Turning
(d) Jostling (e) Walking (f) Waving

Fig. 20 Examples of action data.

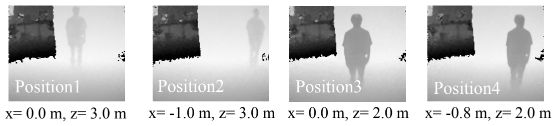


Fig. 21 Examples of depth images at given positions.

We used six different actions, crouching, dropping, turning, jostling, walking, and waving, in our experiments, as shown in Fig. 20. For the crouching action, the person stretched and bent their knees several times into a kneeling position. For the dropping action, the person fell to the ground from an upright position. For the jostling action, two facing people grasp each others' arms and jostled. The person stood and then looked back for the turning action. The person marched in the same position for the walking action. For the waving action, the person shook and raised both arms from horizontal to straight up several times. In using these actions, the jostling and dropping are examples of abnormal actions that are violent and accidental, respectively, and the remaining 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 the products of the 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, three action experimenters, and the six positions. Two of the positions are given in the second and third columns in Fig. 13, where the people were only partially within view, and four of them are shown in Fig. 21, where the entire person's body was shown. Around 3,000 action data frames were taken at each position: 3,190, 3163, 2,924, 2,892, 3,177, and 2,892 in action order.

There were three kinds of motion features: MHI, CHLAC [8], and ST-Patch [16]. 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. 5. The dimensions of ST-Patch are expanded using the six accumulated frames [6].

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.

3.6 Evaluation results

We describe our evaluation of the experimental results using the proposed method. The experiments were done under the three following conditions.

3.6.1 Evaluation results using simulated deficit

For a fundamental evaluation, only the compensation process in the proposed method was evaluated using synthetically deficient depth images under the condition that the set of deficit lengths is dense, the calculation of the deficit length according to the human positions (Section 3.3.1) is omitted, and the deficit length of the regression coefficients (Section 3.3.2) is set to the deficit length of the synthesized data. For synthesizing the deficit depth images, the points whose height was from the ground level to the deficit length were omitted. The set of deficit lengths is incremented by 150 mm from 300 to 1,200 mm. The data for positions 2 and 3 in Fig. 21 were used for training, and the ones for positions 1 and 4 were 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 person being observed were diminished by the three-view drawings expansion, which is the projection transformation from the viewpoints at infinite distances.

Graphs of the F-measure averaging for every action are shown in Fig. 22. The ones without motion feature compensation, the ones with motion feature compensation by [1], and the ones with learning with deficient images are also shown in Fig. 22. Reference [1] describes 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. 5. The restored images by [1] are not dealt as deficient (0 mm deficient) and are used for recognizing actions with the

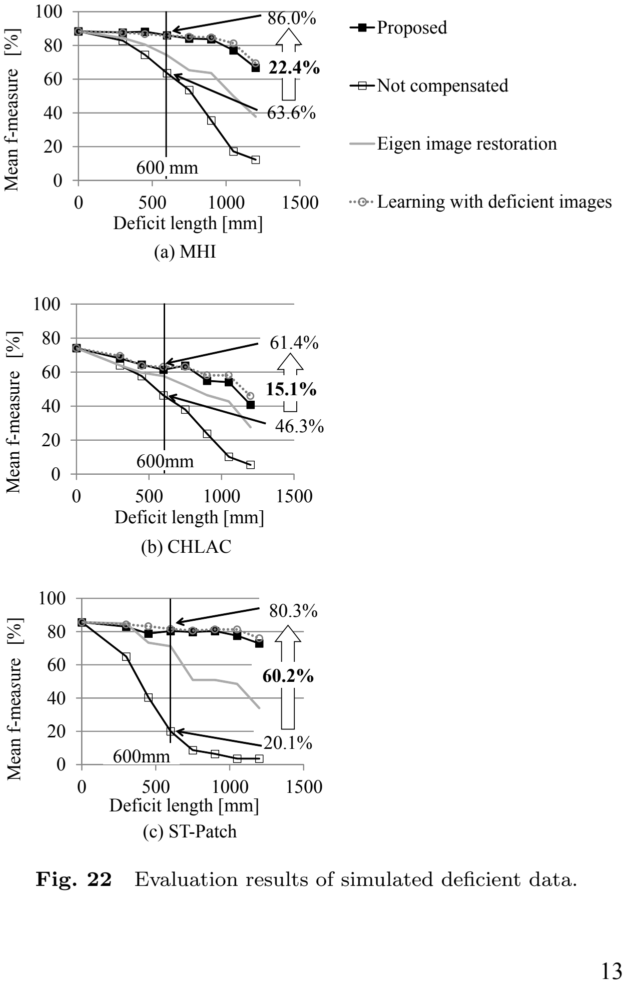


Fig. 22 Evaluation results of simulated deficient data.

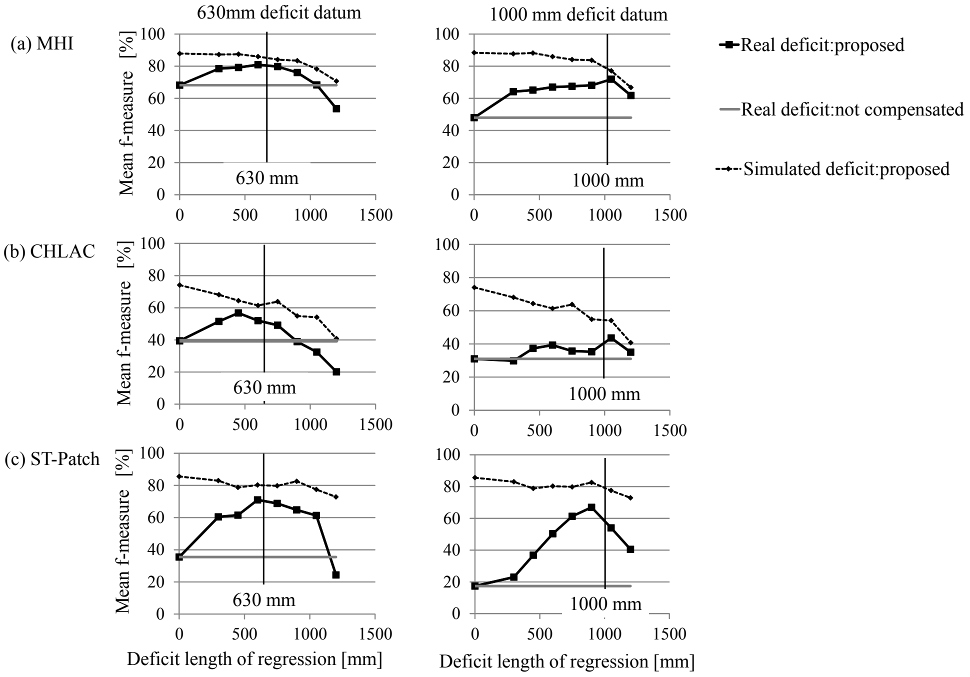


Fig. 23 Evaluation results of actual deficient data.

method described in Section 3.4. Learning with deficient images is a useful method for learning the classifiers of action categories from several levels of deficient images in the manner described in Section 3.4 and for recognizing action categories without motion feature compensation by selecting a classifier according to deficient length of the target data.

The compensation in the proposed method was valid because the F-measure when the motion features were compensated for was consistently higher than the one when not compensated for within the range of 0 to 1,200 mm deficient. When choosing to focus on the 600 mm deficient case, improvement of the F-measure was 22.4% when using MHI, 15.1% using CHLAC, and 60.2% using ST-Patch. The 600 mm deficient length here 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 the people being observed come closer to the sensor by up to 29% under the experimental conditions, which is closer by a distance of 0.7-2.4 m, where the deficit length is just zero. The result in which the degree of improvement of the ST-Path is especially large in this paper 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 the parts of the people being viewed protrude outside the viewing angle.

When comparing the proposed method and [1], 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, their graphs both nearly overlap but the latter did slightly surpass the former; it quantitatively surpassed it by 1.6% on an average of 3 kinds of motion features and 7 levels of deficient lengths. This difference should be small comparing with the 33.2% average under the same conditions between the proposed method and when not compensated for. Whereas the difference in the F-measure is small, the proposed method is superior in terms of the memory usage. The memory usages of the proposed method and learning with deficient images were calculated using Eqs. (18) and (19).

$$NFDw + CM(D + C)w \tag{18}$$

$$NCM(D + C)w \tag{19}$$

In Eqs. (18) and (19), N is the levels of the deficit length, F and D are the dimensions of the motion features before and after LDA, w is the byte length per data element, C is the action categories, and M is the number of representative vectors for an action category. When using the MHI features, N is 7, F is 108, D is 54, w is 8, C is 12, M is 100, and the memory usages of the proposed method and learning with deficient images are 0.9 MBytes and 4.4 MBytes, so the former needs 80% less memory than the latter. This difference in memory usage is significant for low cost embedded processors, which are widely used for monitoring purposes.

3.6.2 Evaluation results using actual deficiency

The proposed method was evaluated for actual deficient depth images using the data from the 2nd and 3rd columns in Fig. 13. The median calculated deficit length was 630 mm for the data in the 2nd column in Fig. 13 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 3.6.1.

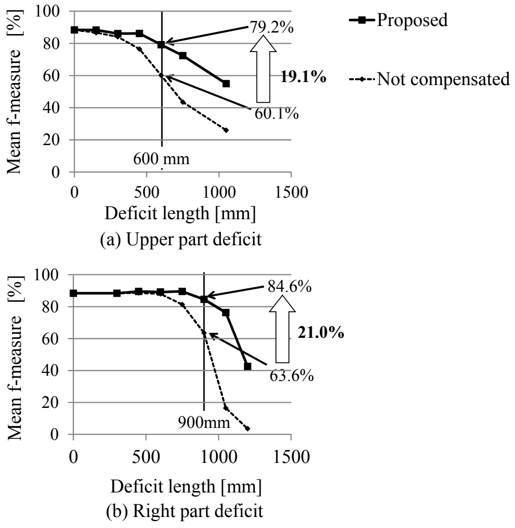
Graphs of the experimental results are shown in Fig. 23. The horizontal axes of the graphs in this figure represent the deficit lengths of the regression coefficients (Section 3.2). There are three graphs in this figure, when actually deficient data is compensated for, when actually deficient data is not compensated for, and when the simulated deficient data described in Section 3.6.1 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 all the motion features when the regression coefficients were used for the deficit length 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 deficient data are compensated for in Fig. 23, the graphs were almost identical when regression coefficients that were close to the actual deficit length were used, and the differences in F-measure between them were at most 10.5% when the regression coefficients closest to the actual deficit length were used.

3.6.3 Evaluation results with altered deficient positions

Two cases were used for evaluating whether or not the deficient parts could be altered from the 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 were simulated deficiencies like those described in Section 3.6.1. For the upper deficit position, the deficit length was set to 0 - 2,150 mm from the floor, which corresponds to the overall height of the person whose arms are raised above them. For the right deficit position, the deficit length was set to 0 mm at a position 900 mm to the right of the centroid of the people being observed, which corresponds to the maximum length of an arm lifted horizontally. MHI was used as the motion features. When we take the manner of motion for each action category in Fig. 20 into consideration, there should be a correlation between the motion features from the entire body and the ones from the partially deficient upward and to the right body views.



(a) Upper part deficit

(b) Right part deficit

Fig. 24 Evaluation results when deficit parts are altered.

The experimental results are shown in Fig. 24. 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. These results show that the proposed compensation method for the motion features could be applied to deficit parts other than the lower ones when the deficit position and length could be calculated.

3.7 Summary

We proposed a method in this section that helps to compensate for the motion features that are outside a given viewing angle by using a regression estimate, for enlarging the target area for action recognition for monitoring purposes when using a depth image sensor. This compensation method is composed of human position extraction, deficit length calculation, regression coefficients selection, and regression estimation. We acknowledged the action recognition performance when the compensation was applied from three kinds of experimental results: the simulated deficit and actual deficit in the lower part, and the simulated data with altered deficient positions.

4. Conclusion

In this paper, we proposed action recognition methods for recognizing human action views-invariantly, which are essential for applying action recognition technology to actual security systems. In Section 2, we proposed a generative learning method for action recognition by using the three-view drawings expansion of the depth images. With this method, we could follow the changes in orientation of the people being viewed that are caused by changes in viewpoints between when learning and recognizing, while suppressing the quantity of training samples from the viewpoint when recognizing. In Section 3, we proposed a compensation method of the motion features with regression for deficient depth images. With this method, the affect of the protrusion of any of the body parts of the person being viewed outside the viewing angle can be diminished, so that the depth image sensor can be placed close to the people being viewed.

References

[1] Amano, T., Hiura, S., Yamaguchi, A. and Iguchi, S.: Eigenspace Approach for a Pose Detection with Range Images, *Trans. IEICE C*, Vol. J80-D-2, No. 5, pp. 1136–1143 (1997).

[2] Bradski, G. and Davis, J.: Motion Segmentation and Pose Recognition with Motion History Gradients, *IEEE Winter Conference on Applications of Computer Vision*, pp. 238–244 (2000).

[3] Breiman, L.: Random Forests, *Mach. Learn.*, Vol. 45, No. 1, pp. 5–32 (online), DOI: 10.1023/A:1010933404324 (2001).

[4] Holte, M., Moeslund, T. and Fihl, P.: Fusion of range and intensity information for view invariant gesture recognition, *Workshop on Time-of-Flight based Computer Vision*, pp. 1–7 (2008).

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

[6] Kazui, M., Miyoshi, M. and Muramatsu, S.: Incoherent Motion Detection using a Time-series Gram Matrix Feature, *International Conference on Pattern Recognition*, pp. 1–5 (2008).

[7] Ke, Y., Sukthankar, R. and Hervert, M.: Event Detection in Crowded Videos, *IEEE International Conference on Computer Vision*, pp. 8–15 (2007).

[8] Kobayashi, T. and Otsu, N.: Action and Simultaneous Multiple-Person Identification Using Cubic Higher-Order Local Auto-Correlation, *International Conference on Pattern Recognition*, pp. 741–744 (2004).

[9] Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B., Rennes, I., Grenoble, I. I. and Ljk, L.: Learning realistic human actions from movies, *IEEE Computer Vision and Pattern Recognition*, pp. 1–8 (2008).

[10] Li, W., Zhan, Z. and Liu, Z.: Action Recognition Based on A Bag Of 3D Points, *Workshop on CVPR for Human Communicative Behavior Analysis (in conjunction with CVPR 2010)*, pp. 9–14 (2010).

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

[12] L.Xia, Chen, C. and J.Aggarwa: View Invariant Human Action Recognition Using Histograms of 3D Joints, pp. 20–27 (2012).

[13] 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 (in conjunction with ICCV 2011)*, pp. 1147–1153 (2011).

[14] Schwarz, L., Mateus, D. and Navab, N.: Manifold learning for ToF-based human body tracking and activity recognition, *British Machine Vision Conference*, pp. 80.1–80.11 (2010).

[15] Seki, M., Hayashi, K., Taniguchi, H., Hashimoto, M. and Sasagawa, K.: Violent Action Detector for Elevator, *Symposium on Sensing via Image Information*, pp. 273–278 (2004).

[16] Shechtman, E. and .Irani, M.: Space-Time Behavior-Based Correlation OR How to Tell If Two Underlying Motion Fields Are Similar Without Computing Them?, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 11, pp. 2045–2056 (2007).

[17] 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, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1297–1304 (2011).

[18] Wang, H., Klaser, A., Schmid, C. and Liu, C.-L.: Action Recognition by Dense Trajectories, *CVPR*, pp. 3169–3176 (2011).

[19] Weinland, D., Ronfard, R. and Boyer, E.: Free viewpoint action recognition using motion history volumes, pp. 249–257 (online), DOI: 10.1016/j.cviu.2006.07.013 (2006).