Copyright ⓒ2017 KSII

Spatial-temporal texture features for 3D human activity recognition using laser-based RGB-D videos

Yue Ming[1,*], Guangchao Wang[1], Xiaopeng Hong[2]
[1]School of Electronic Engineering, Beijing University of Posts and Telecommunications
Beijing 100876, P. R. China
[2]Department of Computer Science and Engineering, University of Oulu, Finland
[e-mail: myname35875235@126.com]
*Corresponding author: Yue Ming

Received September 24, 2015; revised July 23, 2016; revised August 29, 2016; accepted September 20, 2016; published March 31, 2017

Abstract

The IR camera and laser-based IR projector provide an effective solution for real-time collection of moving targets in RGB-D videos. Different from the traditional RGB videos, the captured depth videos are not affected by the illumination variation. In this paper, we propose a novel feature extraction framework to describe human activities based on the above optical video capturing method, namely spatial-temporal texture features for 3D human activity recognition. Spatial-temporal texture feature with depth information is insensitive to illumination and occlusions, and efficient for fine-motion description. The framework of our proposed algorithm begins with video acquisition based on laser projection, video preprocessing with visual background extraction and obtains spatial-temporal key images. Then, the texture features encoded from key images are used to generate discriminative features for human activity information. The experimental results based on the different databases and practical scenarios demonstrate the effectiveness of our proposed algorithm for the large-scale data sets.

Keywords: Spatial-template texture features, 3D human activity recognition, RGB-D videos, depth information, Maximum Outline of the History Behavior Binary Image (MOHBBI)

The work presented in this paper was supported by the National Natural Science Foundation of China (Grants No. NSFC-61402046), President Funding of Beijing University of Posts and Telecommunications (Grants No. 2013XZ10). The previous version of this manuscript has been presented in Interntion Conference on Human System Interactions and the current version is more than 50 percent substantial new contributions added in the manuscript. The new contributions include Section 1, Section 2, Section 3, parts of Section 4, parts of Section 5, and parts of Section 6.

https://doi.org/10.3837/tiis.2017.03.019 **ISSN : 1976-7277**

1. Introduction

Human activity recognition has gradually penetrated into intelligence video surveillance, body sensor interaction, virtual reality and other fields. Due to its broad application prospects, it has grown to be an important research topic in computer vision and machine learning. Hence, the accuracy has significantly improved for 2D human activity recognition [1]. However, the traditional video capturing method for RGB videos is a challenging problem because it dramatically changes the subject's appearance under varying illumination, shadow or shaking, etc, especially for the applications where the subject is non-cooperative [2, 3, 4].

The video capture landscape has been substantially changed by the proliferation of depth information acquisition device (such as Kinect and Leap motion). The depth data captured from the lowered cost of 3D somatosensory camera and smart terminal can collect RGB-D videos in real-time and can effectively overcome the influence of illumination and occlusions, which also significantly increase the size of data [5]. As a result, a large number of RGB-D video processing methods have been proposed for human activity analysis, which can collect different geometric shapes based on camera distances for preserving more discriminative information. 3D human activity recognition can obtain the distinctive advantages for improving the recognition accuracy [6]. However, fusing the depth channel will dramatically increase the data dimension, which causes a rapid increasing in computational complexity and a rapid decline in operational efficiency. How to find an efficient feature representation algorithm different from the traditional videos, which can simultaneously reduce the computational load and improve the recognition accuracy, is a critical problem for human activity recognition in the age of Big Data.

Based on the above discussion, we expand the LBP (Local Binary Pattern) description into a three-dimensional spatial-temporal space for fusing depth and its corresponding color videos. The novel human activity recognition method includes the following steps: video acquisition based on laser-based projector, spatial-temporal key images extraction in 3D human activity videos, spatial-temporal texture feature extraction, feature classification and activity recognition.

The main process is shown in **Fig. 1**. Firstly, we collect human activity videos by Microsoft Kinect camera to obtain the RGB and depth videos; then, we construct volume activity model based on the modified ViBe method from depth and RGB videos, which generates the mixed spatial domain MOHBBI for spatial-temporal key images extraction that can be characterized for this activity video. Furthermore, we perform feature extraction based on spatial-temporal texture feature in space and time domains for different activity videos, therefor, the local image feature of the video model is obtained. Afterwards, the classification and recognition by different methods is performed on these activity feature data. Experimental results demonstrate that our method can dramatically reduce the size of data and overcome the challenging issues of human activity recognition.

We concentrate on the key issues closely related to 3D human activity recognition and propose a new framework. The main contributions of our novel framework can be summarized as follows:

1. Reliability: A modified ViBe (Visual Background extractor) and volume activity model have been combined in spatial domain for RGB-D video preprocessing. Different from previous research, we extract the spatial motion key images for generating Y-T and X-T

activity orthogonal planes for better describing the activity changing trends and effectively eliminating the noises around the moving objects. When light shade, shadow and jitter noise happens, the recognition rate will reduce substantially. The introduction of depth distance information and our improved spatial-temporal preprocessing method for 3D human activity videos can make up for the shortage of RGB videos, which can greatly reduce the interference of shade shadow, light change, color and other factors.

2. Efficientness: we derive a novel feature for 3D human activity recognition, named spatial-temporal texture feature. The logical relation between frames and variation trends between spatial-domain volume changes can be described by time-domain key images. A hybrid spatial-temporal texture feature integrated of RGB-D videos lays a solid foundation for the higher level data analysis. The proposed spatial-temporal texture feature suggests that the local pattern texture features not only reduce the information amount of the feature data, but also can effectively characterize motion information of human activity from videos with accomplishing promising recognition result. Therefore, the features based on local pattern texture characteristics are gradually applied to the field of human activity recognition from videos.
3. Universality: The collected RGB-D video datasets include the major challenges for human activity recognition, such as the different individuals, different times, different distances, and so on. According to the performance evaluation based on different databases and different practical scenarios, the performance improvement of our framework has demonstrated the universality with a wide range of distances, poses, colors and complex backgrounds. The novel framework for 3D human activity recognition proposed in this paper obtains better performance for the different datasets and scenarios, which has the sufficient universality based on the different datasets and practical applications.

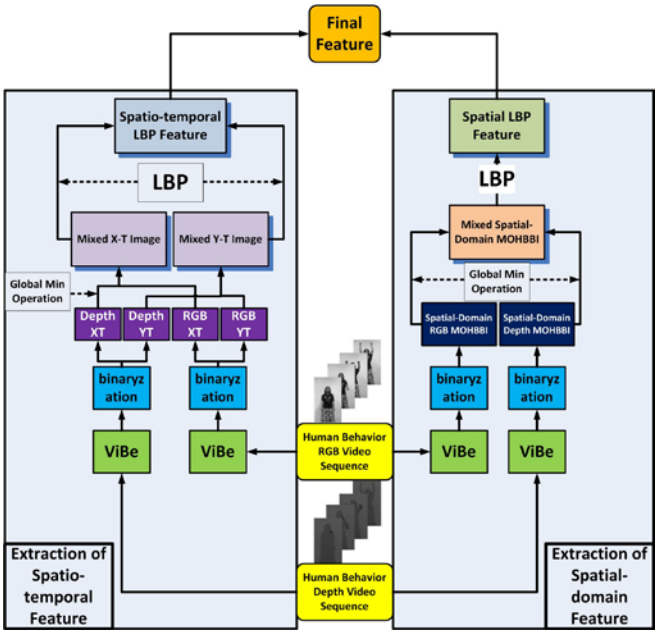

Fig. 1. The framework of our proposed 3D human activity recognition method.

The rest of this paper is organized as follows. Related work about human activity recognition is introduced in Section 2. Section 3 describes the procedure of the optical video capturing. Spatial-temporal preprocessing of 3D human activity videos is described in Section

4. Section 5 proposes our novel framework of spatial-temporal feature extraction for 3D human activities. Then, experimental results are given in Section 6. Finally, we conclude the paper in Section 7.

2. Related Work about Human Activity Recognition

The recent focus on human activity recognition has resulted in a variety of approaches. Sufficient broad investigations have been achieved in the literature [1, 5, 6, 7]. Here, we thoroughly analyze the related work into three parts: activity recognition in 2D and 3D spaces, feature extraction, and spatial-temporal feature descriptors.

2.1 Activity Recognition in 2D and 3D Spaces

Most previous researches on human activity recognition are based on RGB videos and great strides have been made in recent years. Aggarwal and Ryoo [7] summarized the related research progress of human activity analysis. Evert et al. [8] proposed spatial-temporal interest points (STIPs) for the recognition of realistic human activities. Lin et al. [9] modeled human trajectories as series of heat sources and introduced a thermal diffusion process to represent the group activities. The high-level descriptions and interactive phrases [10] can be learned for recognizing human interaction activities. However, these algorithms only focus on the color-related recognition. Thus, it is very difficult to handle the challenging issues of illumination and occlusions, especially for unconstrained wild scenarios.

With the advent of the laser-based video capturing methods, a large range of activity recognition algorithms in 3D space have been proposed [5], which showed the superiorities of the depth information for activity feature extraction. 3D centroid trajectory as a naïve human activity feature can represent a human subject as a point, which was used to indicate the 3D location of the subject in the visual scene [11]. However, these representations are only suitable for the individual that only occupies a small region in the video frame. Human shape information is another representation for activities based on RGB-D data, of which the most representative is 3D human silhouette [12]. 3D human models are the third category for recognizing human activities in 3D space [13, 14]. The discrimination and robustness of human activity recognition algorithms mentioned above relies substantially on the foreground segmentation and human tracking, which are hard-to-solve challenges due to camera dithering, lighting, occlusions and pose changes [15, 16].

2.2 Feature Extraction

First, HOG (Histogram of Oriented Gradient) and HOF (Histogram of Optical Flow) features have been widely introduced into human activity analysis [17]. However, these features based on the motion and gradient are directly calculated by the whole RGB or depth frames, which is difficult to describe the local fine-motion trends. The traditional features also include MEI (Motion Energy Information) [18] and MHI (Motion History Information) [19] and their expansions. However, these methods rely heavily on supervised learning [20] and require a large number of hand annotations and training samples.

In recent years, a large range of novel features are specially proposed for RGB-D data. Our previous work extended the traditional feature MoSIFT into 3D MoSIFT by adding the depth information, which can significantly improve the recognition accuracy [21]. Then, 3D EMoSIFT (3D Enhanced Motion SIFT) [20] and 3D SMoSIFT (3D Sparse Motion SIFT) extended our previous work to spatial-temporal domain by fusing RGB-D information. The

features not only have the properties of the rotation and scale invariance, but also have more compact and richer visual representation [22]. However, feature extraction combined RGB and depth video information will dramatically increase the data storage capacity and the step of feature extraction is time-consuming.

2.3 Spatial-temporal Feature Descriptors

As a promising representation algorithm, spatial-temporal features have important theoretical and practical value in the relevant fields, such as computer vision, machine learning and robotics communities. Zhao et al. [23] proposed the LBP for dynamic texture recognition and applied to facial expression analysis, which showed the advantages of local processing, robustness to monotonic changes and simple computation. Although spatial-temporal features used human activity recognition have demonstrated superior performance based on color information from 2D videos, most of previous spatial-temporal features do not make use of one important piece of information that is now available depth. Further, they extended texture feature description to a three-dimensional space [24] and proposed LBP-TOP description. Then, Shao et al. [25] introduced LBP-TOP texture features to human activity recognition and obtained good recognition results.

3. Video Acquisition based on IR Camera and Laser Projector

One laser-based IR projector combined with two IR cameras can be used to be composed of RGB-D video capture device. First, a fixed pattern of light and dark speckles are sent out from a set of diffraction. The pattern is memorized at a known depth. The memorized pattern at that pixel compare the local pattern with each pixel in a new IR image in a 9 * 9 correlation window. An offset from the known depth to the best match is called disparity. Suppose we know the known depth of the memorized pattern, and the disparity, an estimated depth for each pixel in the IR image can be obtained by triangulation.

The video capturing device can return the RGB video and its corresponding depth video. However, the offset between RGB and depth video is small and fixed, since there is a small baseline separated by the IR cameras and laser-based IR projector. Stereo calibration based on chessboard data can be used to determine the 6 DOF transform and calibrate the two videos,which is called as RGB-D videos.

4. Spatial-temporal Preprocessing for 3D Human Activity Videos

Video pre-processing can effectively remove irrelevant backgrounds and reduce noise interference, since human activity videos are easily affected by the lighting conditions, occlusions and complex backgrounds. In this paper, we adopt ViBe (Visual Background extractor) [26] background substraction method for each activity video, and calculate video spatial images binarization based on a universal threshold. The Vibe algorithm have already achieved an excellent performance on extracting object from traditional RGB videos, which is fast and accurate. In order to ensure the accuracy of mix of binary image in depth and RGB videos and reduce the deviation error that caused by using different extracting methods in different videos, we also apply the Vibe algorithm on depth videos to extract objects. Then, the mixed binary image in depth and RGB videos is used for constructing the volume activity model in spatial domain.

4.1 Video Sequence Preprocessing

We first use the first frame of the video to construct the initial background model. In the first frame, each pixel (x, y) can be treated as the center to form a circle, whose radius is set as R. We randomly select N pixels in the circle, denoted as the background sample of the pixel $S_R(x, y) = \{p_0, p_1, p_2, \cdots, p_{N-1}\}$, where p_i is the selected pixel value. The selected pixels were stored in a dataset in turn to generate the updated background model. Then, we sum S_R of the whole frame. If W and H are set as width and height of the frame, the union can construct the background sample $\bigcup_{x,y}^{W,H} S_R(x, y)$. For the following frames of the video, if the pixel values are close to the corresponding sample values in the background sample, the pixels will be identified as background points. Obviously, if a new pixel belongs to the background, the observed values of the points should be similar to sampling values of corresponding pixel in the background model. The background sample continuously updates to adapt to the changing of the background of the sequence images, and then we obtain background subtraction in time domain.

Donate that $Pix_n(x, y)$ is pixel value of (x, y) in the n-th frame, and $P_{n,R}(x, y) = \{s_0, s_1, s_2, ..., s_{N-1}\}$ is the sample set sampled by R at pixel (x, y) in the n-th frame. If $M_n(x, y) = P_{n,R}(x, y) \cap S_R(x, y) > M$, where M is a threshold and we take $M = 2$ in this paper, then the pixel (x, y) belongs to background point, otherwise, it belongs to foreground point. In the same time, we perform binary operation on the background points and foreground points. The binarization processing of foreground pixels and background pixels can be calculated as following:

$$Pix_n(x, y) = \begin{cases} 255, (x, y) \in \text{Foreground pixels} \\ 0, (x, y) \in \text{Background pixels} \end{cases} \tag{1}$$

In order to better adapt to the changes in the background, the background model needs to constantly update. The background update strategy is used in this paper by testing and counting the occurrence of foreground pixel. For example, if a pixel point is detected T times continuously as foreground pixel, then assume the pixel as the background pixel and update the background model. On the contrary, if less than T times, the pixel is still assumed as foreground pixels. $T = 20$ is defined in this paper. Besides, the self-update method for background model in this paper applies the probability updating method, where $\varphi = 16$, which makes the probability equals to $1/\varphi$ for each background pixel to update its modeling sample value itself.

While updating the sample set, we randomly select a sample from all N samples to update, therefore the probability for one sample to be updated is $1/N$ and the probability of not being updated is $(N-1)/N$. After dt, the probability for sample value staying constant is $P(t, t+dt) = (\frac{N-1}{N})^{(t+dt)-t}$, i.e., $P(t, t+dt) = e^{-\ln(\frac{N}{N-1})dt}$, where t is irrelevant.

Then, the original video can be transformed to RGB foreground activity video and its corresponding depth foreground activity video as shown in **Fig. 2** and **Fig. 3**. We can see that the noises around the moving object can be effectively eliminated by Vibe algorithm. Through the above preprocessing, the motion region can be effectively separated from the background,

which reduces the effects of noise and jitter, and reduces the size of data.

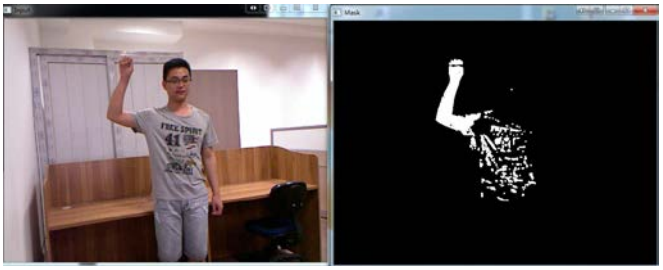

Fig. 2. Effects of RGB video before (left) and after (right) background substraction.

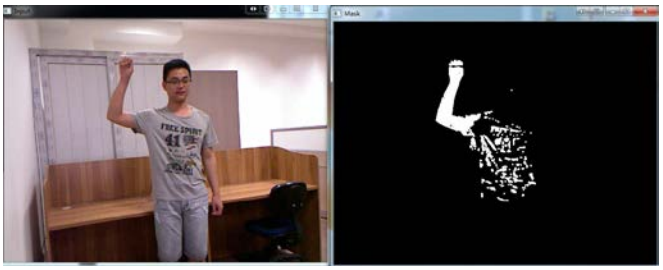

Fig. 3. Effects of depth video before (left) and after (right) background substraction.

4.2 Volume Activity Model in Spatial Domain

In order to describe the activity from the video in 3D space and time domains, this subsection will introduce Volume Activity Model in Spatial Domain for modeling the moving object region extracted by the subsection 4.1. The method extracts the spatial-temporal activity change images of video that can characterize the human activity from the video. We introduce the LBP-TOP feature [25] into the foreground video pairs. The sequence of video frames can constitute a three-dimensional volume video over time T. Then, the human activities in videos can be described by a spatial volume model as shown in **Fig. 4**.

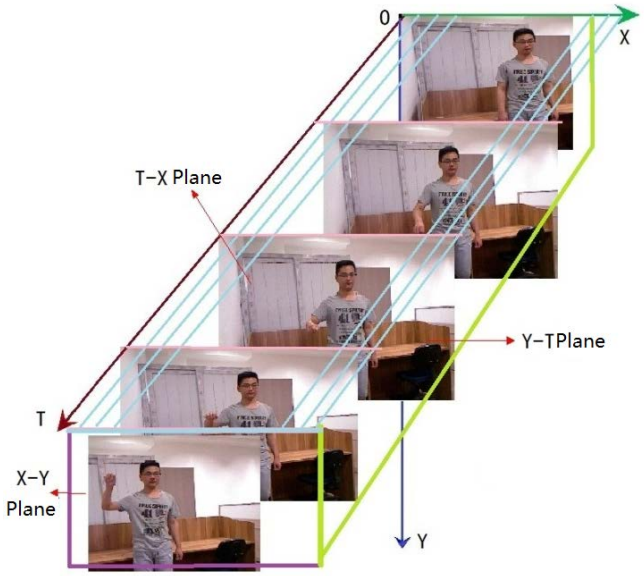

Fig. 4. Spatial volume model of RGB human activity video.

The volume activity model of activity video in spatial domain not only contains the action change in time domain, but also includes the action change in the orthogonal three-dimensional space domain. Compared with texture feature extraction in the pure time-domain, it largely conserves the action change information in human activity videos.

Then，image sequence composed by X-Y plane images represents the activity changes in time domain. Y-T and T-X orthogonal planes represent the changes of human activities in a three-dimensional space. Therefore, we extract the spatial motion key images from Y-T and T-X planes, and then generate Y-T and T-X activity orthogonal plane in the spatial domain for better describing the activity changing trends.

We can use the robustness of depth video to slight distance to effectively remove the image regions that do not change very much on the surface of motion and the noise pixels caused by jittering, such as the cloth color change caused by motion and color difference caused by different wearing. Compared to simple RGB video, using RGB-D video presents more generative meanings. Simple RGB video could easily be influenced by the non-edge pixels and jitter noises caused by background subtraction. Thus, once the depth video is introduced, the mixed image can better remove irrelevant pixels.

5. Spatial-temporal Texture Feature Extraction

Here, spatial-temporal texture feature extraction is introduced to our 3D human activity recognition. First, time-domain key image extracted by volume activity model is used to represent the history outline images of the human activity. Then, we extract motion key images of Y-T and T-X orthogonal planes to describe spatial-domain volume changes. Finally, we concatenate the local feature of spatial domain in spatio-temporal domain as the final spatial-temporal texture features.

5.1 Time-domain Key Image Extraction

In order to describe the features that change along with the activity of human in videos, this subsection details the time-domain key image extraction method based on volume activity model. It basically extracts the feature which can represent the history outline images of the human activity in the video. After video pre-processing, binary foreground videos contain no color information. Developed from our previous method [27], we establish a time-domain image model and calculate the maximum pixel value in the corresponding location frame by frame, called Maximum Outline of the History Behavior Binary Image (MOHBBI). Due to color pixel jitter in RGB videos, depth videos can effectively avoid the influence of noise pixels. As a result, we extract MOHBBI from depth foreground videos to remove residual change (**Fig. 5 A, B**). Meanwhile, MOHBBI generated by RGB foreground video can also remove depth pixel noises produced by video capturing (**Fig. 6 C**).

Fig. 5. RGB behavior video and its time-domain MOHBBI.

Fig. 6. Depth behavior video and its time-domain MOHBBI.

In order to take full advantages of the complementarity of RGB and its corresponding depth videos, the fusion strategy is used to describe the time-domain MOHBBI, which can effectively remove non-edge pixels and noise pixels and maximally preserve the outline of the history activity.

$$P_{mix}(x,y) = \max\left\{P_{depth}(x,y), P_{rgb}(x,y)\right\} \tag{2}$$

where $P_{rgb}(x,y)$ and $P_{depth}(x,y)$ are respectively pixel values of MOHBBI generated by RGB and depth foreground videos in (x,y). $P_{mix}(x,y)$ is the pixel value of Maximum MOHBBI in (x,y). The fusion result is shown in **Fig. 7**. The process not only eliminates interference, but also preserves human history activity outline. The fused activity history outline model images have shaper edges and can better describe the outline feature and reduce the interference of non-activity outline, which is favorable for the feature extraction latter on, compared to the non-fused images.

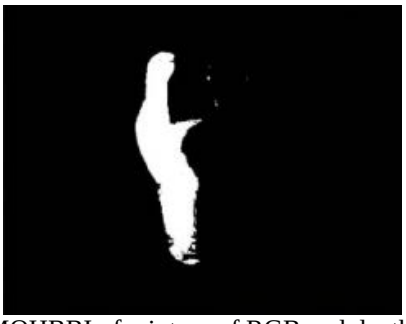

Fig. 7. MOHBBI of mixture of RGB and depth videos.

5.2 Image Extraction based on Spatial-domain Volume Changes

After building a model for activity volume video in spatial domain, we should extract motion key images of Y-T and T-X orthogonal planes. As illustrated in **Fig. 8**, the row of pixels in Y-T orthogonal plane is decided by the row of pixels of frames in X-Y plane. The row of pixels in X-T orthogonal plane is decided by the column of pixels of frames in X-Y plane. Then, we respectively sum row and column pixels of each frame of time-domain original RGB and depth videos. The pixel values in Y-T orthogonal plane are defined as global binarization pixels. The threshold is set as the mathematical expectation M in the frames of X-Y plane,

$$P_{Y-T}(y,t)=\begin{cases}0,(\delta_{y,t}<M_t)\\255,(\delta_{y,t}>M_t)\end{cases}$$
$$M_t=\frac{1}{Width\cdot Hight}\sum_{i=1}^{Width}\sum_{j=1}^{Hight}pix_t(x_i,y_j),1\le t\le F \quad (3)$$
$$\delta_{y,t}=\frac{1}{Width}\sum_{i=1}^{Width}pix_t(x_i,y)$$

where $P_{Y-T}(y,t)$ is the binarization pixel value in the location of (y,t) in Y-T orthogonal plane. $\delta_{y,t}$ is the pixel mean of row y in the t^{th} frame in X-Y plane. M_t is the global pixel mean of the t^{th} frame in X-Y plane. $pix_t(x,y)$ is the value of pixel (x,y) in the t^{th} frame and F is the total number of video frame. Width and height are respectively width and height of video images in X-Y plane.

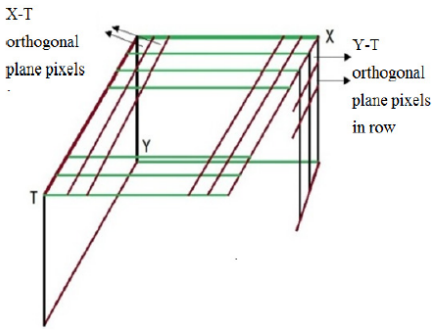

Fig. 8. Volume video model of space-domain activity video sequences.

Pixel points in X-T orthogonal plane are defined as similar as the pixel points in Y-T plane, which are calculated as follows,

$$P_{X-T}(x,t)=\begin{cases}0,(\omega_{x,t}<M_t)\\255,(\omega_{x,t}>M_t)\end{cases}$$
$$\omega_{x,t}=\frac{1}{Hight}\sum_{j=1}^{Hight}pix_j(x,y_j) \quad (4)$$

where $P_{X-T}(x,t)$ is the binarization pixel value in the location of (x,t) in X-T orthogonal plane. $\omega_{x,t}$ is the pixel mean of column x in the t^{th} frame in the X-Y plane. The definition of M_t is the same as the above formula. It can better verify that the three dimensional space-time motion change images generated by orthogonal space activity video model have relatively high texture discrimination. The experiments based on Weizmann database show that the activities, which are difficult to be distinguished, such as jump and walk, texture distinguishing effect is relatively obvious, as shown in **Fig. 9** and **Fig. 10**. Realizing the modeling of human activity in the three dimensional space domains from depth and RGB videos and generating the three dimensional space-time motion change images of the model are the conditions for the follow-up information extraction based on local model coding.

Fig. 9. Jump of Daria (left) and Denis (right), respectively changing images in X-T and Y-T planes.

Fig. 10. Walk of Daria (left) and Denis (right), respectively changing images in X-T and Y-T planes.

5.3 Spatial-temporal Texture Feature Extraction

In this subsection, we apply LBP [28] and its variant algorithms, including Rotation Invariant LBP [29] and Uniform LBP [23] for extraction and description of temporal-spatial features obtained by the above subsections. In our spatial-temporal texture feature extraction, we first introduce the depth and RGB information into traditional LBP and its derivative algorithms as shown in equation (5),

$$f_{spatial} = LBP\left\{\min_{RGB,Depth}\left[\max\left(P_{\substack{RGB\\1\le i\le M}}(x,y,i)\right),\max\left(P_{\substack{Depth\\1\le i\le M}}(x,y,i)\right)\right]\right\} \tag{5}$$

Then, local features from spatio-temporal behavior volume change image is extracted by LBP and its variant algorithms [30] as the equation (6), which are used to describe the spatio-temporal features of human activities.

$$f_{spat-temp} = \left[\underset{pix\in RGB}{LBP}\left(I_{Y-T}\right), \underset{pix\in RGB}{LBP}\left(I_{X-T}\right), \underset{pix\in Depth}{LBP}\left(I_{Y-T}\right), \underset{pix\in Depth}{LBP}\left(I_{X-T}\right)\right] \tag{6}$$

Finally, we concatenate the local feature of spatial domain with the local feature of spatio-temporal domain as the final spatial-temporal texture features.

$$f_{final} = \left[f_{spatio}, f_{spat-temp}\right] \tag{7}$$

Our proposed spatial-temporal texture feature not only reduces the information amount of the feature data, but also can effectively generates discriminative features for human activity information from RGB-D videos with accomplishing promising recognition result. The following experiments demonstrate that the features based on spatial-temporal texture characteristics are gradually applied to the field of human activity recognition from RGB-D videos.

6. Experimental Results and Analysis

In order to verify the feasibility of our proposed framework for 3D human activity recognition, experiments will be carried out in several private and public RGB-D datasets, including Local Human Activity datasets and DHA dataset. We also test the performance of our proposed framework in practical scenarios in subsection 6.3. The experimental results based on the different databases and practical scenarios demonstrate the effectiveness and universality of our proposed algorithm.

6.1 Experiments with the Local Human Activity Datasets

To verify the effectiveness of activity recognition method proposed in this paper, two video datasets were recorded, including RGB and depth videos. The one was recorded in an ordinary environment, and the other was recorded under complex lighting conditions as illustrated in **Fig. 11** and **Fig. 12**, denoted as Local-Simple and Local-Complex human activity databsets, respectively. The following experiments were carried out to verify the feasibility and effectiveness of recognizing 3D human activities by fusing RGB and depth information.

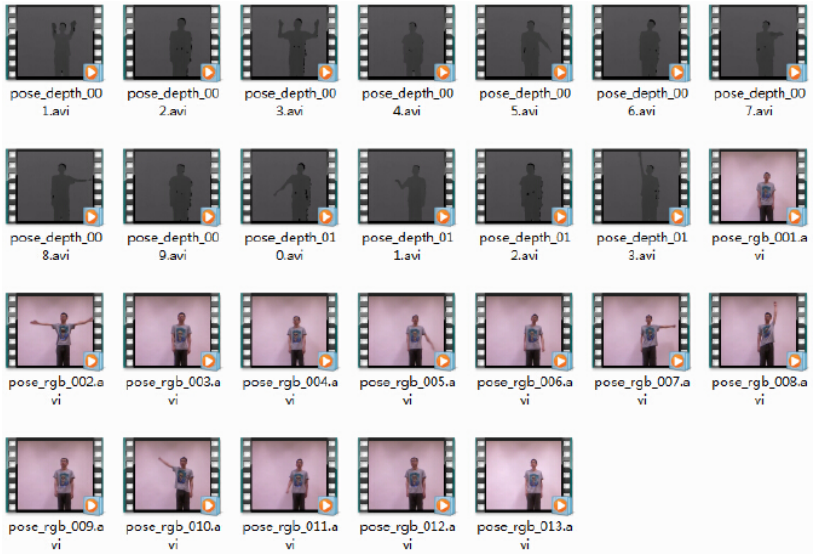

Fig. 11. The local-simple human activity dataset.

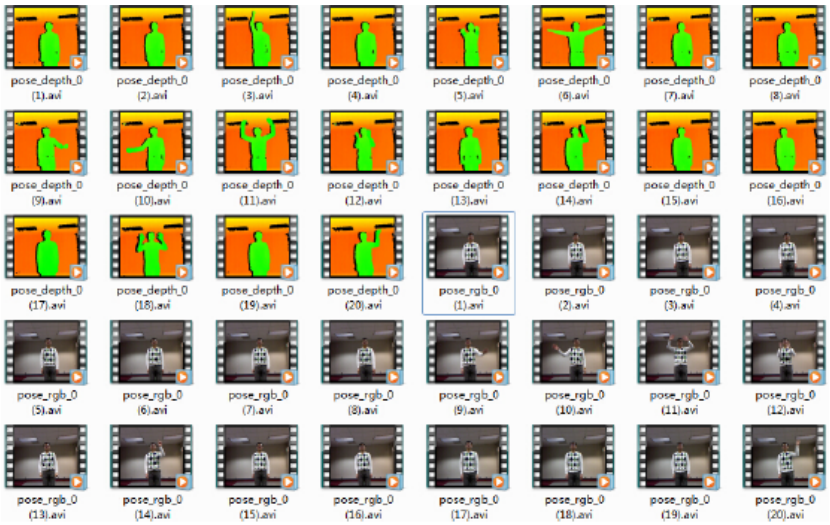

Fig. 12. The local-complex human activity dataset.

First, Local-Simple activity databset was composed of 13 categories of human activities. Each of them was completed by 10 individuals. The leave-one-out method was employed [31] to separate all the samples into 10 parts. One part was left out for testing with the remaining nine as the training set. We take turns to conduct experiments and calculate the average recognition rate. We introduced LBP, RILBP and Uniform-LBP as feature extraction methods

for evaluating the performance of our proposed method. KNN and HMM were used for classification. The results on local 3D human activity data were set in **Fig. 13** and **Table 1**.

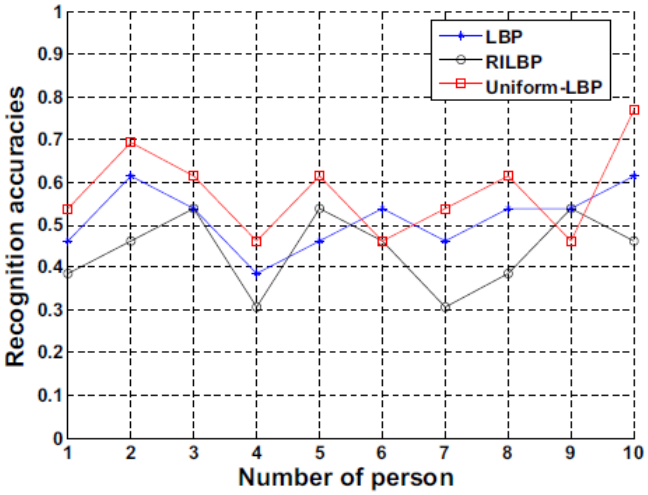

Fig. 13. Recognition rate of KNN classification algorithm in activity video database.

Table 1. Human activity recognition rates of different characteristics and classification algorithms

Methods	KNN	HMM
LBP	51.54%	47.69%
RILBP	43.85%	39.23%
Uniform-LBP	57.69%	53.85%

On the local activity dataset, the recognition strategy proposed in this paper achieved satisfied results. Although the recognition accuracy is less than public activity video database, our method has been proved to be feasible and universal for human activity recognition under normal circumstances, so it can be widely applied in 3D human activity analysis.

Furthermore, to evaluate the performance of human activity recognition in complex environments (illumination variances, background jitter and other interference factors), Local-Complex human activity databset, comprised of 20 categories of human activities including single and double hands pointing, waving and clapping. Each activity was completed by seven individuals, and three kinds of collection distances were used (behavior individual to somatosensory distance).

First, according to the different distances, three categories are divided into 2m, 3m and 5m. Each category contains 7 people and 20 kinds of activities per person. The leave-one-out method is employed at each distance to leave out one individual as the test set, while the remaining six individuals constitute the training set. We evaluate the recognition rates for performance comparison. LBP, RILBP and Uniform-LBP are selected as the extracted features. KNN and HMM algorithms are used for classification. The recognition results of KNN method at different distances are shown in **Table 2** and **Table 3**. We obtain the best result at the distance of 3m. Thus, we further evaluate the performance of two classification algorithms and different features as shown in **Fig. 14**. The recognition rates of Uniform-LBP feature are selected as the extracted features. KNN and HMM algorithms are used for classification. The recognition results of KNN method at different distances are shown in **Table 2**. The recognition rates of Uniform-LBP feature are shown in **Table 3** with the different classification methods and different distances. The experimental results have proven

that 3D human activity recognition method proposed in this paper can achieve a good recognition performance in complex environments. It also verifies that 3D spatial-temporal texture features can effectively respond to environment interference with stable recognition effects. The experiments fully demonstrate that our proposed method has effectiveness and robustness to light variations and complex backgrounds, which can be applied to 3D human activity recognition in complex environments.

Table 2. Recognition results at different distances under the KNN classification algorithms

Person number	2m	3m	5m
A	0.45	0.5	0.25
B	0.35	0.45	0.35
C	0.3	0.35	0.25
D	0.35	0.45	0.2
F	0.4	0.55	0.35
G	0.35	0.4	0.3

Table 3. Recognition rates under different classification algorithms and different collection distances in Uniform-LBP features

Classification algorithms	2m	3m	5m
KNN	35.71%	44.28%	27.14%
HMM	36.43%	37.86%	30.71%

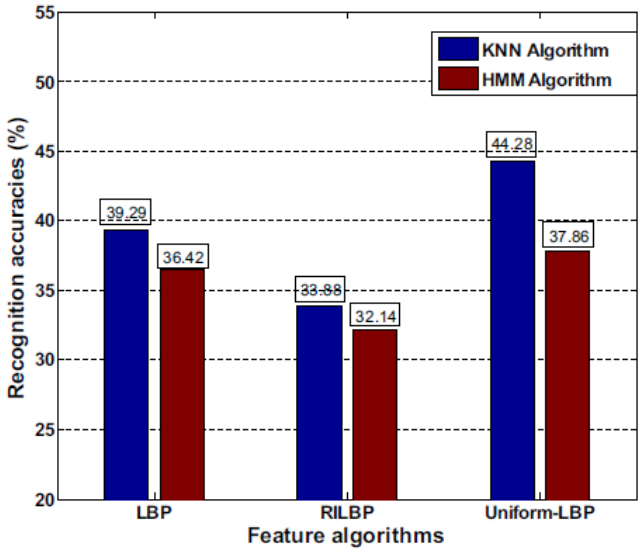

Fig. 14. Recognition rates of different rates at two classification algorithms at the distance of 3m.

6.2 Experiments with DHA Behavior Dataset

We also test the algorithm performance with international common DHA behavior dataset [32]. This dataset collects 17 types of human activities from 17 different individuals. It includes depth videos and their corresponding RGB videos. Leave-one-out method is introduced to evaluate the experimental results. We divide all samples into 21 parts. One part is treated as the test set of activity sample and the other ones as training sets. The test set is changed in turn and the rest samples are used for new training sets. The average recognition

rates can be calculated and evaluate the performance of our proposed method for the common dataset. Feature extraction methods include LBP, Edge-LBP, RILBP and Edge-RILBP [23]. Classification algorithms are KNN and HMM. The recognition results are shown in **Table 4**.

Table 4. Recognition accuracies of four algorithms experiments on DHA behavior database

Algorithms	Edge-RILBP	Edge-LBP	RILBP	LBP
HMM	93.15%	86.07%	83.78%	76.63%
KNN	86.58%	82.13%	74.05%	64.97%

It can be concluded that the recognition rate is remarkably improved by adding depth information. The linear classification KNN is weaker than HMM classification algorithm. The results show that the recognition proposed in this paper has higher recognition accuracy than the traditional methods. Its recognition rate is up to 93%, which demonstrates that our recognition algorithm has excellent performance on 3D human activity recognition.

In **Table 5**, we illustrate the time consuming of our proposed recognition framework, including the steps of the pre-processing runtime, feature extraction time and classification time. Compared with Edge-LBP, our method has longer extraction time and shorter classification time with higher recognition accuracy. Our proposed method has relatively smaller feature data compared with other methods. Thus, our method efficiently reduces the runtime and improves the recognition speed, which has been widely applied in 3D human activity recognition.

Table 5. Running time(s) of two feature algorithms

Running time(s)	Preprocessing	Feature Extraction	Recognition
Edge-LBP	63.17	0.21	0.14
Ours	63.17	0.28	0.13

6.3 Implementation of Visual Perception System

To further investigate the performance of our framework in a practical application, we apply the proposed framework into our developed visual perception system. The system software is based on the MFC framework and written in Microsoft Visual Studio 2010 development environment. It can run on Windows PC. The programs of image processing and algorithm involve open source libraries of Opencv and OpenNI. The hardware of the device is a Kinect camera containing depth information, which is designed to collect human activity videos; the computer with excellent performance that can perform parallel GPU computing is designed to demonstrate the video image processing algorithm and the platform.

In this system, based on the proposed human activity recognition algorithm mentioned in section 4 and 5, we construct the human activity perception system and realize the individual activity control service in the real scene. The system mainly consists of video acquisition module, preprocessing module, feature extraction and classification algorithm module, activity training model module, activity recognition and personalized service module.

In smart-home-controlled scenario, it mainly contains 10 types of physical activities and 5 kinds of personalized control services. Firstly, use Kinect somatosensory device to capture RGB video and depth video sequences, each activities performed by 8 different individuals, totally 80 pairs of activity videos. 10 types of activity model are generated by training phase of preprocessing, feature extraction and classification. The testing phase performs the same operation on the collected human activity video and then matches the features to recognize the

current activity. The personalized service module finds the service instruction for the corresponding activity and executes it through personalized service response. Personalized service response is shown in **Fig. 15**, such as open TV when people make a clapping, then capture the video and recognize it.

Fig. 15. Simulation demonstration effect of personalized service of the system for opening TV.

In motion street scene, it contains 3 kinds of body movements and 3 kinds of personalized service; each activity is performed by 8 different individuals, totally 24 pairs of activity videos. When people start to move, the corresponding system will recognize the activity and simulate the street scene showing corresponding service to the activity. The activity and service under this scenario is shown in **Fig. 16**.

Fig. 16. Demonstration of simulation effect of the running motion in the system.

The difficulties in human activity recognition under complicated environment mainly converge on the impact that interference around the human movement have on analysis of human activities. Experimental tests under the practical scenarios show that our proposed framework of spatial-temporal texture feature provides strong technical support for 3D human activity recognition of personalized services.

7. Conclusion

This paper presents an effective three-dimensional human activity recognition method based on spatial-temporal texture feature. Compared with the previous methods, our method fuses

depth information with RGB video, which has effectively reduced color noises, light and other interference factors. Using behavioral modeling and spatial-temporal key images can decrease the data dimensions and improve the recognition rate. Experimental results show that our method is simple and can be widely used in three-dimensional human activity recognition, which is an efficient way for feature representations from Big Data.

References

[1] P. Borges, N. Conci, and A. Cavallaro, "Video-based human behavior understanding: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 11, pp. 1993-2008, 2013. Article (CrossRef Link)

[2] Jinpyung Kim, Gyujin Jang, Gyujin Kim and Moon-Hyun Kim, "Crowd activity recognition using Optical Flow Orientation Distribution," *KSII Transactions on Internet and Information Systems*, vol. 9, no. 8, pp. 2948-2963, 2015. Article (CrossRef Link)

[3] B. Ben Amor, J. Su and A. Srivastave, "Action recognition using rate-invariant analysis of skeletal shape trajectories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 99, pp. 1-12, 2015. Article (CrossRef Link)

[4] Jinseok Lee, Shung Han Cho, Sangjin Hong, Jaechan Lim and Oh Seong-Jun, "Object tracking in 3D space with passive acoustic sensors using particle Filter," *KSII Transactions on Internet and Information Systems*, vol. 5, no. 9, pp. 1632-1652, 2015. Article (CrossRef Link)

[5] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," *Pattern Reccognition Letters*, vol.34, no.15, pp. 1995-2006, 2013. Article (CrossRef Link)

[6] S.S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1-54, 2015. Article (CrossRef Link)

[7] J. Aggarwal, and M. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys*, vol.43, no.3, pp. 1-47, 2011. Article (CrossRef Link)

[8] I. Everts, J. van Gemert, and T. Gevers, "Evaluation of color spatio-temporal interest points for human action recognition," *IEEE Transactions on image processing*, vol.16, no.2, pp. 1569-1580, 2014. Article (CrossRef Link)

[9] W. Lin, Y. Chen, J. Wu, H. Wang, B. Sheng, and H. Li, "A new network-based algorithm for human activity recognition in videos," *IEEE Transactions on Circuits and Systems I*, vol.24, no.5, pp. 826-841, 2013. Article (CrossRef Link)

[10] Y. Kong, Y. Jia, and Y. Fu, "Interactive phrases: Semantic descriptions for human interaction recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.36, no.9, pp. 1775-1788, 2014. Article (CrossRef Link)

[11] O. Brdiczka, M. Langet, J. Maisonnasse, and J. Crowly, "Detecting human behavior models from multimodal observation in a smart home," *IEEE Transactions on Automation Science and Engineering*, vol. 6, no. 4, pp. 588-597, 2009. Article (CrossRef Link)

[12] M. Singh, A. Basu, and M. Mandal, "Human activity recognition based on silhouette directionality," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 9, pp. 1280-1292, Sept. 2008. Article (CrossRef Link)

[13] J. Y. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from RGBD images," in *Proc. of AAAI Conference on Artificial Intelligence Workshops*, August 7-11, 2011. Article (CrossRef Link)

[14] L. Schwarz, D. Mateus, V. Castaneda, and N. Navab, "Manifold learning for tof-based human body tracking and activity recognition," in *Proc. of British Machine Vision Conference*, August 31 - September 3, 2010. Article (CrossRef Link)

[15] H. Zhang, C.M. Reardon, and L.E. Paker, "Real-time multiple human perception with color-depth cameras on a mobile robot," *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1429-1441, 2013. Article (CrossRef Link)

[16] Hao Zhang and Lynne E. Parker, "CoDe4D: Color-depth local spatio-Temporal features for human activity recognition from RGB-D videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 3, pp. 1280-1292, 2016. Article (CrossRef Link)
[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 886-893, June20-25, 2005. Article (CrossRef Link)
[18] B. Liang and L. Zheng, "Gesture recognition from one example using depth images," *Lecture Notes on Software Engineering*, vol. 1, no. 4, 2013. Article (CrossRef Link)
[19] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," IEEE T PAMI 23(3), 257-267, 2001. Article (CrossRef Link)
[20] Jun Wan, Guodong Guo, and Stan Z. Li, "Explore efficient local features from RGB-D data for one-shot learning gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.38, no.8, pp. 1626-1639, 2016. Article (CrossRef Link)
[21] Y. Ming, and Q. Ruan, "Activity recognition from kinect with 3d local spatiotemporal features," in *Proc. of IEEE International Conference on Multimedia and Expo*, pp. 344-349, July 9-13, 2012. Article (CrossRef Link)
[22] J. Wan, Q. Ruan, W. Li, and S. Deng, "One-shot learning gesture recognition from rgb-d data using bag of features," *Journal of Machine Learning Research*, vol.14, no.1, pp. 2549-2582. 2013.
[23] G. Zhao, and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.29, no.6, pp. 915-928, 2007. Article (CrossRef Link)
[24] G. Zhao, T. Ahonen, J. Matas, and M. Pietikainen, "Rotation invariant image and video description with local binary pattern features," *IEEE Transactions on Image Processing*, vol.21, no.4, pp. 1465-1467, 2012. Article (CrossRef Link)
[25] R. Mattivi, and L. Shao, "Human action recognition using lbp-top as sparse spatio-temporal feature descriptor," *Computer Analysis of Images and Patterns*, vol.16, no.2, pp. 641-648, 2009. Article (CrossRef Link)
[26] O. Barnich, and M. V. Droogenbroeck, "Vibe: A universal background substraction algorithm for video sequences," *IEEE Transactions on Image Processing*, vol.20, no.6, pp. 1709-1724, 2011. Article (CrossRef Link)
[27] Yue Ming, Guangchao Wang, Chunxiao Fan, "Uniform Local Binary Pattern based Texture-Edge Feature for 3D Human Behavior Recognition," *Plos One*, vol.5, no.10, 2015. Article (CrossRef Link)
[28] D. He, and L. Wang, "Texture classification using texture spectrum," *Pattern Recognition*, vol.23, no.8, pp. 905-910, 1990. Article (CrossRef Link)
[29] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.24, no.7, pp. 971-987, 2002. Article (CrossRef Link)
[30] N. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol.19, no.3, pp. 175-185, 1992. Article (CrossRef Link)
[31] Y. Lin, M. Hu, and W. Cheng, "Human action recognition and retrieval using sole depth information," in *Proc. of the ACM international conference on Multimedia*, pp. 168-197, 1997.
[32] Yan-Ching Lin, Min-Chun Hu, Wen-Huang Cheng, Yuang-Huan Hsieh, and Hong-Ming Chen, "Human action recognition and retrieval using sole depth information," in *Proc. of 20th ACM International Conference on Multimedia*, pp. 175-186, 2012. Article (CrossRef Link)

Yue Ming received the B.S. degree in Communication Engineering, and the M.S. degree in Human-Computer Interaction Engineering, and Ph.D. degree in Signal and Information Processing from Beijing Jiaotong University, China, in 2006, 2008, and 2013, respectively. She worked as a visiting scholar in Carnegie Mellon University, U.S., between 2010 and 2011. Since 2013, she has been working as a faculty member at Beijing University of Posts and Telecommunications. Her research interests are in the areas of biometrics, computer vision, computer graphics, information retrieval, pattern recognition, etc.

Guangchao Wang received the B.S. degree in Communication Engineering, and the M.S. degree in Electronic Science and Technology from Beijing University of Posts and Telecommunications, China, in 2012 and 2015, respectively. His research interests are in the areas of biometrics, computer vision, pattern recognition, etc.

Xiaopeng Hong received his BEng, MEng, and Ph.D. degree in computer application from Harbin Institute of Technology, Harbin, P. R. China, in 2004, 2007, and 2010 respectively. He has been a scientist researcher in the Center for Machine Vision Research, Department of Computer Science and Engineering, University of Oulu since 2011. He has authored or co-authored more than 10 peer-reviewed articles in journals and conferences, and has served as a reviewer for several journals and conferences. His current research interests include pose and gaze estimation, texture classification, object detection and tracking, and visual speech recognition.