A Feature Extraction Method Combining Color-Shape for Binocular Stereo Vision Image

Fengfeng Duan
Hunan Normal University
China
Edffeng2010@126.com

ABSTRACT: *Feature extraction is the key and foundation of content-based retrieval of video and image. In order to realize the content-based index and retrieval of binocular stereo vision resources efficiently, the method of feature extraction based on Principal Component Analysis-Histogram of Oriented Depth Gradient (PCA-HODG) and Main Color Histograms (MCH) is proposed. In the method, on the one hand, for the depth map obtained from matching of right image and left image, the PCA-HODG algorithm is proposed to extract shape features. In the algorithm, edge detection and gradient calculation in depth map windows are performed to obtain the regional shape histogram features. Moreover, sliding window detection over a depth map is performed to extract the full features. At the same time, in feature extraction of depth map windows and full depth map, principal component analysis is used to realize dimensional reduction respectively. On the other hand, for the left image of binocular stereo vision, the improved MCH algorithm is used to extract color features. Then the shape and color descriptors can be obtained as 2-dimensional factors for similarity calculation. The experimental results show that the proposed method can detect and extract the features of binocular stereo vision image more effectively and achieve similar classification more accurately compared with the existing HOD, RSDF and GIF algorithms. Moreover, the proposed method also has better robustness.*

Keywords: Feature Extraction, Binocular Stereo Vision, Color-shape, Principal Component Analysis-Histogram of Oriented Depth Gradient (PCA-HODG), Main Color Histogram (MCH)

Received: 18 October 2017, Revised 29 December 2017, Accepted 21 January 2018

DOI: 10.6025/jmpt/2018/9/2/45-58

1. Introduction

At present, the industry and technology of stereo vision develop quickly and attract great attention. The common format is binocular stereo vision image or video which is acquired through the binocular stereo camera. It is often associated left-right images or videos and usually represented as side-by-side. With the development of Internet and new media, it becomes increasingly demanding for users to analysis and acquire the stereo vision resources and content. They also become one of the

most important research and application field of multimedia data. Because of the structural complexity, it is necessary to analyze and obtain the features of stereo vision resources from their own characteristics. In this way, the efficiency of matching and query can be improved.

In three-dimensional system, the 3D model is usually projected to a two-dimensional point cloud and then the features can be extracted. The depth map can be obtained by resample of depth value and formation for structured data while the depth value can be acquired according to the calculation of disparity in 3D data orthogonal projection or matching of corresponding views. Although the depth map can be viewed as a two-dimensional image, it is different in the way of formation. The 2D image is a projection of light reflection, while the depth map of 3D data is a projection of depth value and contains more intrinsic information of 3D [1]. Therefore, the depth map is also called the distance image which refers to the image containing some information related to object surface distance in the scene when observing from target point. In the depth map, the gray value of pixel corresponds to the depth value of the scene. Usually, gray image has various changes in scene and its texture feature is obvious and complicated, while the depth map has different characteristics, including less changing in scene, simpler in texture and clearer in outline. At the same time depth map is independent of color. So, compared with color image, the depth map is seldom affected by interference of light, shadow, and changes of environment. Feature extraction of depth map based on shape can obtain accurate descriptors. They can not only effectively describe the shape of object, but also can better express the change information of depth direction [2]. So, the features of depth map can be used as an important part of stereo vision resource features with properties of translation, rotation and scale invariant.

The rest of this paper is organized as follows: Section 2 introduces the related work and the flow chart of proposed method. Section 3 presents the extraction of shape features based on PCA-HODG. Section 4 shows the extraction of color features based on MCH. Section 5 analyzes the experimental results, and Section 6 concludes the paper.

2. Related Work and Proposed Flow Chart

Feature extraction of image is one of the most important research areas of multimedia. The studies in the area are mainly on the detection of pedestrians or faces, the matching of features and retrieval. In recent years, the feature extraction of stereo vision image based on depth map gradually attracts people's attention and becomes a hot field.

In the related research, the color, interest point, shape or gradient of depth map are mainly extracted as stereo vision image features. Cui et al. [3] proposed the method of depth map feature extraction based on color histogram. In the method, stereo vision image features can be extracted combining with RGB color feature values. In this way, the recognition and tracking of action can be realized. Zhao et al. [2] proposed a depth map feature extraction method based on color histogram of characteristic points, which constructs the stereo visual image feature when combining the monocular image color histogram. The depth map is usually expressed in the form of gray scale, and the color discrimination is small, so the accuracy of feature extraction is low.

The stereo vision image feature extraction based on depth map feature of interest points has been studied in recent years. Karpushin et al. [4] proposed the depth map feature extraction of interest points according to the detection of video frame depth map interest points. Then the stereo vision image features are extracted combining with the texture features of color image. Lu et al. [5] proposed the algorithm of Range-Sample Depth Feature (RSDF) extraction. In the algorithm, the interested points are selected according to the clear outline of depth map. The stereo vision image features can be extracted effectively based on the range sample among the interest points. Interest point feature is mostly with rotation and scale invariant, but usually only being sensitive for strong texture and two-dimensional image with constant high brightness, while the accuracy reduces greatly for brightness variation and weak texture object. So, the accuracy is usually low for these methods.

It is one of the important methods of stereo vision image feature extraction based on depth map object shape as it has distinct outline [6]. Jalal et al. [7] proposed the transformation method of depth map. The scale invariant and dimensional reduction algorithm are used to the depth map outline feature extraction. In this way, the accurate stereo vision image features can be obtained. Liu et al. [8] proposed the Geodesic Invariant Feature (GIF) extraction algorithm, in which the invariant of distant and angle measurement is considered in local depth map. So, the stereo vision image features can be effectively extracted according to the local shape characteristic of depth map. Yang et al. [9] proposed the algorithm of joint-feature guided depth map super-resolution. In the algorithm, the super-resolution of face depth map is optimized based on both depth cues and color cues to obtain the sharp and clean edges. Therefore, high quality stereo vision image shape features can be extracted for facial expression recovered accurately.

The gradient feature descriptor has received more attention in recent years. Dalal et al. [10] proposed the Histogram of Oriented Gradients (HOG) algorithm which is an important innovation of block feature extraction based on object shape. In the algorithm, the translation and rotation invariant can be realized for the quantization in position and direction of space. At the same time, the problem of illumination variation can be overcome because the feature is presented as normalized histogram in the local area. Because the depth map has the characteristics of weak texture and obvious shape of the region, the feature extraction of stereo vision image is paid more and more attention based on depth map shape and gradient detection according to the improved HOG algorithm. Spinello et al. [11] proposed the Histogram of Oriented Depths (HOD) algorithm based on HOG. In the algorithm, depth change direction is encoded locally based on the perception of depth scale space search. The feature detection can be achieved three times in processing speed. Lin et al. [12] studied the region of interest and object detection of depth map, for which low gradient pixels are removed through the filter. Then the stereo vision image features are extracted through the detection and description of high gradient region shape. Liang et al. [13] proposed the method of stereo vision image local feature extraction and representation based on the improved HOG algorithm. The accuracy of these methods is low due to the incomplete range of feature extraction.

Feature extraction of stereo vision image usually fuses hardware equipment to extract depth map. Zhang et al. [14] proposed local spatio-temporal (LST) features extraction method. In the method color-depth bag-of-features are extracted based on depth information acquired by RGB-D cameras. The depth sensor Kinect or camera are mostly used to obtain the real-time depth information. However, there are some invalid areas in the depth images produced by Kinect and related equipment, e.g., the boundaries of bodies, reflective grounds, long distances and object surfaces absorbing infrared light. These regions may induce bad effects without appropriate inpainting measurements [15]. So, it is difficult to obtain good depth information. In fact, these methods only extract features from images and are also difficult to treat weak texture image objects as the features are sensitive to brightness variation.

Feature extraction of monocular 2D image cannot acquire stereo vision image features accurately and comprehensively. It is necessary to overcome the disadvantage and some other problems in the existing algorithms, e.g., sensitive to noise, inaccurate of shape region detection and description, complexity of high dimensional descriptors, lacking of real-time, poor quality of depth map. Kang et al. [16] proposed the depth map upsampling method with low-resolution depth map and a color image, which provides reference for feature extraction of stereo vision image. According to the characteristics of binocular stereo vision

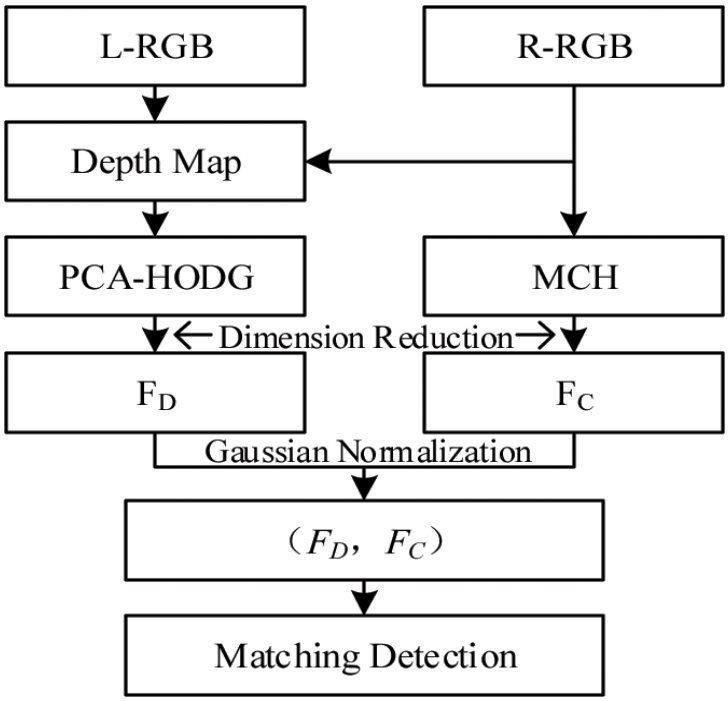

Figure 1. The flow chart of proposed method

image, the method of combining shape and color is proposed to extract the features. In the method, for depth map the Principal Component Analysis-Histogram of Oriented Depth Gradient (PCA-HODG) algorithm is proposed to extract shape features, and for left image of binocular stereo vision, the improved Main Color Histogram (MCH) algorithm is used to extract color features. The flow chart of proposed method is shown in Figure 1.

3. The Extraction of Shape Features based on PCA-HODG

3.1 Disparity calculation and depth map estimation

The algorithm of graph cut based on epipolar rectification is used for disparity calculation. According to the idea of minimum cut and maximum flow, a global energy function is constructed and optimized. The matter of disparity solving is converted into calculating the energy optimization instead [17]. At the same time spatio-temporal consistency is introduced to eliminate flicking artifacts and noise, realizing smooth in spatial and boundary maintaining. The energy function of disparity solving is defined by [18]:

$$E(f) = E_{data}(f) + E_{smooth}(f) + E_{occ}(f) \tag{1}$$

where $E_{smooth}(f)$ is the smooth item and measures the extent to which f is not piecewise smooth, $E_{data}(f)$ is the data item and measures the disagreement between f and the observed data, $E_{occ}(f)$ is the penalty function for temporal consistency.

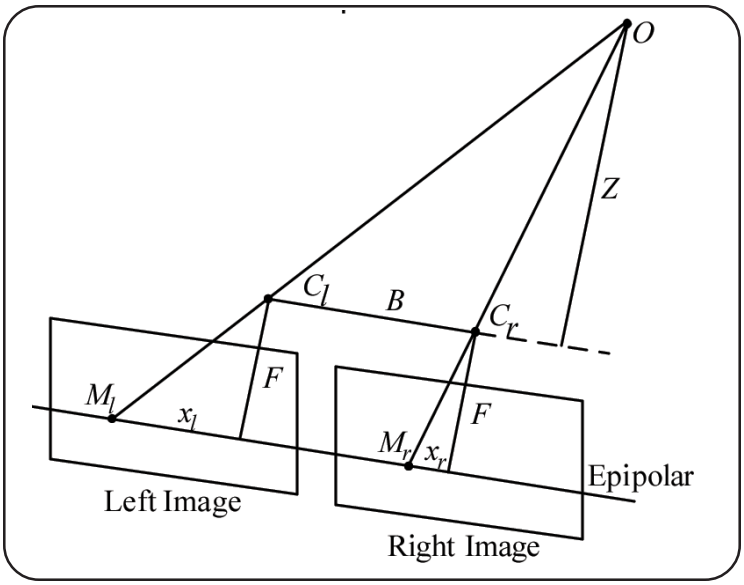

Figure 2. Relationship between depth and disparity

In the system of binocular stereo vision, disparity can be defined as vector difference of object points in each channel image associated with the focus. Binocular disparity is the difference of direction when a goal is observed from two points. The distance between the two points is called baseline. The relationship of disparity and depth is shown in Figure 2, where M_l and M_r are the matching points, and O is the target point. The depth Z can be defined by:

$$Z = \frac{BF}{x_l - x_r} = \frac{BF}{d} \tag{2}$$

where B, F and d represent the camera baseline, focal length and disparity, respectively.

For stereo vision data, the depth map often can be represented by an 8-bit greyscale image to assist rendering new views. When the depth is represented by gray value from 0 to 255 [19], the depth value can be defined as:

$$\bar{Z} = \left\lfloor 255 - \frac{255(Z - Z_{min})}{Z_{max} - Z_{min}} + 0.5 \right\rfloor \quad (3)$$

where Z_{max} and Z_{min} represent the farthest and the nearest depth value, respectively.

The examples of binocular stereo vision image and the depth map which is obtained by matching from left and right images are shown in Figure 3.

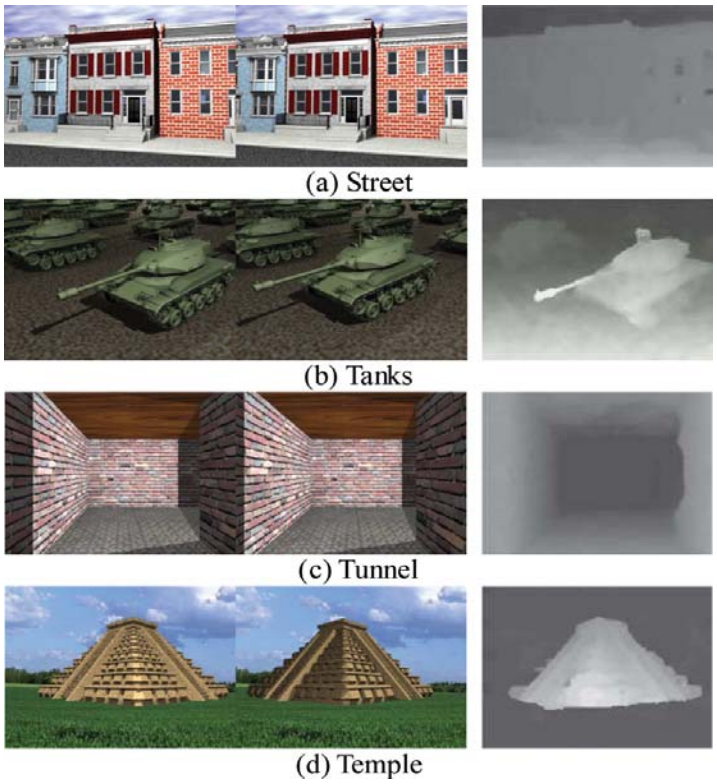

Figure 3. Examples of binocular stereo vision image and depth map

3.2 Feature Extraction and Optimization in Depth Map Window

3.2.1 Canny Edge Detection and Gradient Calculation

The method of gamma correction is used to improve the image contrast and reduce the influence of illumination and shadow in feature extraction. According to the distribution of feature point and its neighborhood pixels, the module and direction of the point can be calculated. Then the gradient information can be obtained. Canny algorithm can be used to detect the object edges as the gradient mainly exists in the object region shape edges in depth map. The module value and direction of gradient are respectively defined as:

$$G(x,y) = \sqrt{G^2{}_h(x,y) + G^2{}_v(x,y)} \quad (4)$$

$$\theta(x,y) = \arctan \frac{G_h(x,y)}{G_v(x,y)} \quad (5)$$

where $G_h(x,y)$ and $G_v(x,y)$ represent horizontal and vertical module value respectively. When the depth map is processed with [-1,0,1] non smooth gradient algorithm and convolution, the module value can be respectively defined as:

$$G_h(x,y) = d(x+1,y) - d(x-1,y) \quad (6)$$

$$G_v(x,y) = d(x,y+1) - d(x,y-1) \tag{7}$$

3.2.2 Feature Descriptors based on Histogram of Oriented Depth Gradient

According to the module values and orientations, gradient information is gathered. Gradient is divided into positive and negative direction which represents with $t(x, y)$, and is defined as:

$$t(x,y) = \begin{cases} 1, & 0 < \theta(x,y) < \pi \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

According to the principle of HOG algorithm, the selected window is divided into several blocks and each block includes many cells. At the same time, the overlap of blocks is adopted to eliminate aliasing effect [8]. If η is the number of blocks in the selected window, ξ is the number of cells of each block, β is the number of bins, then the dimensional number of feature descriptors can be represented as $n = \eta \times \xi \times \beta$. The feature descriptor of the cell k based on gradient histogram is defined:

$$H_k = \bigcup_{\delta=1}^{\beta} \sum_{x,y \in cell(k)} G(x,y)\gamma(x,y) \tag{9}$$

In the range of $[0, \pi]$, it is divided averagely into 9 directions, which generates 9-dimensional histogram bins. The size of each cell is 8× 8 pixels and each block contains 2 × 2 cells. In the window of 64 × 128, the number of blocks is 105 and the feature descriptors are 3780 dimensions [8]. Linear gradient histogram feature vectors can be expressed as:

$$H = (H_1, H_2, ..., H_n) \tag{10}$$

3.2.3 Feature descriptors optimization based on Principal Component Analysis

According to the calculation of gradient oriented histogram feature in depth map windows, region shape features can be obtained. However, the feature dimension is large and it will lead to complexity increasing of feature matching and reduction of efficiency. The depth map of stereo vision is the special gray image and the characteristic is not clear within object areas. So, dimensional reduction of feature vectors can be implemented and it has few effects on feature expression. With the PCA method, the n-dimensional feature vectors are expressed as $\eta \times \phi$ matrix, in which $\phi = \xi \times \beta$. Then it can be constructed into a covariance matrix. The mean of features and covariance matrix are respectively defined by [20]:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} H_i \tag{11}$$

$$S_T = \frac{1}{n}\sum_{i=1}^{n}(H_i - \mu)(H_i - \mu)^T = \frac{1}{n}AA^T \tag{12}$$

where A is equivalent to H, they are both feature descriptor value matrix; A^T is orthogonal transformation matrix, which constitute the new feature vector space; S_T is the diagonal matrix. The feature vector descriptor of the covariance matrix can be defined as:

$$Y(i) = \frac{1}{\sqrt{\lambda(i)}} A \cdot v(i), \quad i = 1,2,...,n \tag{13}$$

where $\lambda(i)$ is the feature value which is represented by the variance of variable value in feature vector space; $v(i)$ is the corresponding feature vector [2]. According to the PCA, feature values are ordered in descending and the first p feature vectors are selected as principal components. In the experiment, the values of p can be determined according to training test to samples in speed and accuracy of matching. The expression of dimensional reduction transformation in feature space is defined as:

$$U(i) = (H_i - \mu)Y(i), \quad i = 1,2,..., p \tag{14}$$

3.3 Feature Descriptors for Depth Map

Assume that each frame of binocular stereo vision image is $M \times N$ in resolution. According to HOG algorithm, the selected window size is 64× 128 in feature descriptor extraction of depth map and the scale transform is used for selected image to fit the size. In addition, the method of blocks overlap is used in the algorithm. However, the scale transform will lead to change of object properties as well as inaccuracy of features. In order to obtain feature blocks as much as possible and accurately, sliding window detection over a depth map is performed to extract the features [15]. At the same time, the overlap of windows is also performed to realize the full feature window detection in image region. Considering the resolution of the depth map and the size of selected window, the depth is divided into W windows and the number of W is:

$$W = \psi \times \varphi \tag{15}$$

where $\psi = \left\lceil \frac{M}{64} \right\rceil$, represents the number of windows in horizon; $\varphi = \left\lceil \frac{N}{128} \right\rceil$, represents the number of windows in vertical. In the sliding detection of window, the jump value of window in horizontal and vertical direction is $\frac{m-64}{M-1}$, $\frac{n-128}{N-1}$ in each time respectively. So, the features of W windows can represent a full depth map through division and detection of windows. In this way, W feature sequences are constructed for each depth map. For example, in the experiment, the resolution of depth map is 400× 300, so W is 21. The example of sliding window in feature detection, blocks overlap and windows overlap are demonstrated in Figures 4, 5(a) and 5(b) respectively.

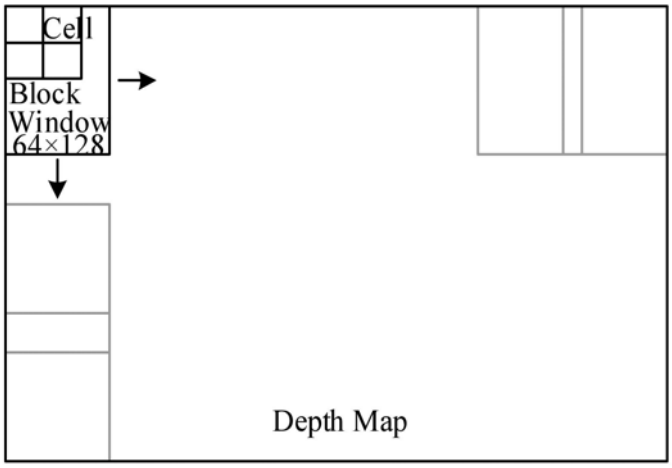

Figure 4. The example of sliding window in feature detection

For W feature sequences, each sequence has p-dimensional feature vectors and it will form $W\times p$ dimensional matrices. The method of PCA is used again to reduce the dimensions of the data. Then the first p principal components are selected as the feature values of the full depth map.

4. The Extraction of Color Features based on MCH

The left image of binocular stereo vision RGB image is used for feature extraction. Color features can be obtained without high complexity and large amount of calculation. Moreover, they are often not sensitive to rotation, scaling, fuzzy and other physical transformation. It has great advantages to measure and represent the global difference of the two images by color features based on histogram. So, color features can be selected as an important part for similarity matching and retrieval.

4.1 Calculation of Color Histogram

RGB is the most common color space in video and most of the digital images are also expressed with the RGB color space. However, the spatial structure does not satisfy human in subjective judgment of color similarity. So, it is necessary to convert it into HSV space which is the closest with the subjective perception of human eyes [21]. The conversion expressions from RGB to HSV are:

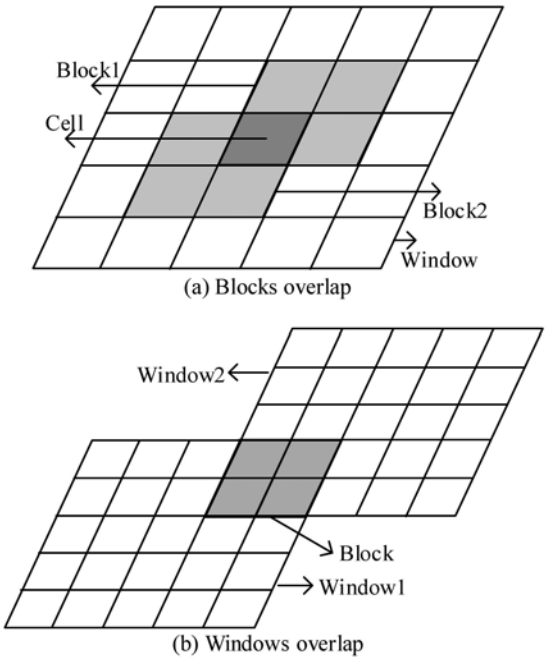

Figure 5. Overlap in feature extraction

$$H=\begin{cases}0, & \text{if } max=min\\ \dfrac{G-B}{max-min}\times 60, & \text{if } R=max\\ \dfrac{B-R}{max-min}\times 60+120, & \text{if } G=max\\ \dfrac{R-G}{max-min}\times 60+240, & \text{if } B=max\end{cases} \tag{16}$$

where $max = max(R, G, B)$, $min = \min(R, G, B)$ and if $H<0$, then $H = H + 360$.

$$S=\begin{cases}0, & \text{if } max=min\\ 1-\dfrac{min}{max}, & \text{otherwise}\end{cases} \tag{17}$$

$$V = max \tag{18}$$

HSV color space is quantified and then is synthesized into one-dimensional feature vector according to the quantization level [22]. The synthesis formula is defined as:

$$L = HQ_SQ_V + SQ_V + V \tag{19}$$

where Q_S, Q_V are quantitative series of component S and V.

Each component of HSV color space is quantified as non-equal interval for 8 segments [0,7], 3 segments [0,2] and 3 segments [0,2]. According to the quantification, $Q_S = 3$ and $Q_V = 3$. At the same time, the values of each component are $H = 7$, $S = 2$, $V = 2$. The HSV color space is quantified for 72 segments [0,71] when synthesized into one-dimensional feature vector. Then the histogram distribution is calculated and is defined by:

$$H(k) = \frac{n_k}{N} \ (k = 0,1,2,3,\ldots,K-1) \tag{20}$$

where K represents the number of colors contained in the image, n_k represents the number of pixels whose quantified color value is k, N represents the total number of pixels within the image.

4.2 Extraction of Main Color Histograms

For color histogram, pixels of high frequency are selected as the main colors, which is called the main color histogram. The pixels of low frequency can be regarded as noise. So the main color histograms can represent features of image. In order to represent the image features more comprehensively with main color, the method of cluster is used to acquire the main color histograms and the center of each cluster is considered as the main color in this paper. Based on the idea of K-means, the m dimensional feature values of main color histogram are calculated and the value of m is chosen according to the dimension of shape feature. The algorithm can be described as flows:

(1) Initialization and the number of m elements $h_1, h_2, ..., h_m$ are selected arbitrarily as the cluster centers, then m cluster spaces $K_1, K_2, ..., K_M$ are established;

(2) For a sample x in the sample set X, it can be adjusted into cluster K_J which is corresponded by h_i according to the rule of minimum distance calculation $j = \arg\min_{1 \le i \le m} \delta(x, h_i)$ that is $x \in K_J$, where $1 \le i \le m; 1 \le J \le M$;

(3) Calculate the mean of the elements in each cluster, that is $h_i = \frac{1}{n_i} \sum_{x \in K_J} x$, where n_i is the number of elements in the cluster space K_J, and then update each cluster center;

(4) If the cluster center is no longer changing or the value of E is the minimum, where $E = \sum_{i=1}^{m} \sum_{x \in K_J} \|x - h_i\|^2$, then the clustering is over, or turn to (2).

According to the algorithm above, m dimensional main color histogram features can be obtained, which are the final center h_i of each cluster, where $h_i \in \{h_1, h_2, ..., h_m\}$.

5. Experimental Results and Discussions

5.1 Experimental plan

The experiment is carried out on similarity classification simulation of binocular stereo vision images. The features are extracted from depth map and left image. Then the stereo vision images are divided into the highest similar classes respectively according to the calculation of similarity. At the same time, the robustness of feature extraction and matching is verified. Binocular stereo vision test sequences 'Street', 'Tanks', 'Temple' and 'Tunnel' provided by the University of Cambridge Computer Laboratory are used for experimental implementation [23]. The number of each sequence is 100 frames, and each frame is 400•~300 pixels in resolution with a disparity range of 64 pixels. For experimental environment, we use Windows 7 of 32-bit dual-core processor and the frequency is 3.3GHz. Matlab 7.10.0 is used for algorithm simulation.

The values of shape and color features may have considerable differences. It will be difficult for one feature to play its role and the weights of feature components may also not uniform in similarity matching. To realize the uniform of feature ratio and component weights respectively, Gaussian normalization for extracted features is implemented for the 40 dimensions. The expression of Gaussian normalization is defined by:

$$g(i,j) = \frac{1}{2}\left(\frac{f(i,j) - M_j}{3\sigma_j} + 1\right) \tag{21}$$

where $f(i)$、$g(i)$ are feature values before and after Gaussian normalization; M_i、σ_i are mean and standard deviation values in the ith dimension; if $g(i) > 1$, then $g(i) = 1$; if $g(i) < 0$, then $g(i) = 0$.

The experimental results are compared with HOD, RSDF, and GIF algorithms. The expression of similarity is:

$$D_{12} = \left[\sum_{i=1}^{p} \left((U_1(i) - U_2(i))^2 + (H_1(i) - H_2(i))^2 \right) \right]^{1/2} \tag{22}$$

where $U_1(i)$, $H_1(i)$ are shape and color feature descriptor values to be tested; $U_2(i)$, $H_2(i)$ are the values in sample library; p is the number of feature dimensions. In the expression, if the D_{12} is smaller, then the similarity is greater.

The shape and color features of binocular stereo vision image in the same sequence usually have higher similarity. So, the stereo vision images belonging to the same sequence in the database can be classified into the same category. The average accuracy of classification shows the effect of feature extraction. In the paper, the average accuracy rate of classification for each test sequence is defined by:

$$\gamma = \frac{1}{T} \sum_{q=1}^{T} \frac{m_q}{I} \tag{23}$$

where T is the number of stereo vision images in one sequence; m_q is the number of stereo vision images belonging to the sequence in first I of similarity matching results.

5.2 The selection of Feature Dimension

In this study, the feature dimension of main color histogram is determined according to the dimension of depth map. In this way, it is helpful for the normalization and matching of features. The precision of feature extraction and matching as well as the time complexity is affected by the dimension of feature. In the process of feature extraction and dimensional reduction, the proposed method in this paper is simulated to test the influence of feature dimension on classification accuracy and running time for the above four stereo vision test sequences. The experimental results of average accuracy and running time affected by feature dimension are shown in Figure 6.

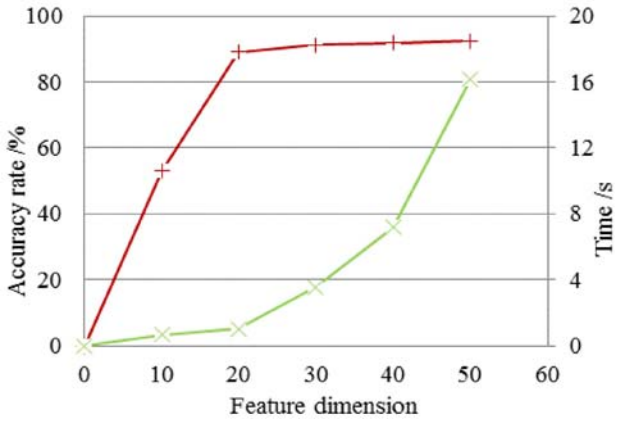

Figure 6. Average accuracy rate and running time affected by feature dimension

According to the experimental results in Figure 6, the accuracy decreases with the decrease of feature dimension, and the time complexity increases with the increase of the dimension. However, it is remarkable that the accuracy decreases rapidly with the decrease of the dimension when it is below about 20. On the contrary, the time complexity increases rapidly with the increase of the dimension when it is about more than 20. So, in this paper 20 is selected as the feature dimension of the depth map, that is p=20.

At the same time the main color feature dimension *m* is also set to 20.

5.3 Comparison of Accuracy

In the experiment, $T = 100$. The 400 binocular stereo vision images of the four sequences are stored in random order. The similarity matching of features is calculated. The accuracy rates of classification for each sequence in different algorithms are shown in Table 1.

Algorithm	Street	Tanks	Tunnel	Temple
HOD	88.12	83.47	87.72	86.03
RSDF	89.01	84.15	86.96	86.75
GIF	88.53	84.36	89.34	87.69
Proposed	90.42	85.14	92.08	88.47

Table 1. Accuracy rates of classification in different algorithms (%)

As is shown in Table 1, the proposed method in this paper has always higher average classification accuracy rates in feature similarity matching compared with the HOD, RSDF and GIF algorithms for each test sequence. For the frames of test sequence Street, the average classification accuracy rate of proposed method increased by 1.58% when compared with the RSDF algorithm, while for Tanks, Tunnel, Temple, the average classification accuracy rates increased by 0.92%, 3.07%, 0.89% respectively when compared with the GIF algorithm. So, the proposed method can achieve better feature extraction and matching classification.

5.4 Comparison of Running Time

The test experiment for running time of classification is implemented to verify the time efficiency of the proposed method and the other algorithms. In the experiment, binocular stereo vision image feature extraction and classification for each test sequence are carried out by using these algorithms and the comparative results of average running time is shown in Table 2.

Algorithm	Average running time
HOD	2.54
RSDF	3.28
GIF	3.61
Proposed	1.03

Table 2. Average running time of feature extraction and classification (s)

The experiment results in Table 2 show that the average running time of the proposed method in this paper is greatly reduced. They are decreased by 59.45%, 68.60%, 71.47% respectively compared with the HOD, RSDF, GIF algorithms. So, the proposed method can reduce the complexity effectively and improve the efficiency greatly.

5.5 Robustness Analysis of Feature Extraction

To validate the robustness of the proposed method, the rotation invariant and stability to noise of binocular stereo vision feature descriptors for each algorithm are verified. Nevertheless, the rotation and noise factors are not contained in extraction of depth maps in proposed method. The two validate experiments are: (1) rotation in the same plane for images of feature extraction. To simulate the influence of environment, zero-centred Gaussian noise with $\sigma = 15$ is added. Feature extraction is performed with the HOD, RSDF, GIF algorithms and proposed method respectively. Then calculate the accuracy rates of classification according to similarity matching. The accuracy rates of classification for the sequence 'Street' are shown in Figure 7; (2) the zero-centred

Gaussian noise with σ =15, 30, 45, 60, 75 is added into images of feature extraction respectively. Then the accuracy rates of classification for each algorithm are calculated. The accuracy rates of classification with different amount of noise for the sequence 'Street' are shown in Figure 8.

Figure 7. Accuracy rates of classification with rotation

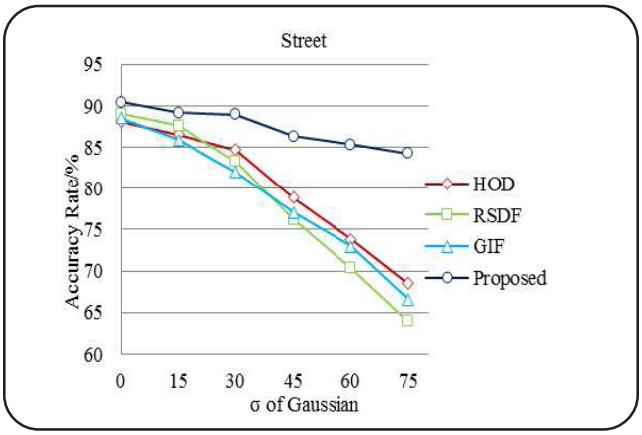

Figure 8. Accuracy rates of classification with noise

Figures 7 and 8 only list the verification results of the 'Street' sequence as the large amount of experimental data and the length limitation of the paper. From Figure 7, in the experiment of rotation invariant similarity classification, the proposed method has always higher accuracy rates in different rotation angles. The results show that the proposed method has better robustness for rotation invariant. Similarly, according to the experimental results in Figure 8, the accuracy rates of classification decrease rapidly for HOD, RSDF and GIF algorithms with the addition of noise while the proposed method is less affected. So, the proposed method has better stability to noise.

6. Conclusions

For the characteristics of binocular stereo vision image in obvious shape contour and weak texture of depth map, the PCA-HODG algorithm of depth map feature extraction is proposed. Then it is combined with the improved MCH algorithm to extract the features of binocular stereo vision image. The method can extract the features of stereo vision image comprehensively and efficiently. Moreover, the inaccuracy of feature extraction caused by poor quality depth map can be overcome. In addition, dimensional reduction can also reduce the complexity and improve the speed of similarity matching. The experimental results show that the method can better realize the feature extraction and similarity classification of images, and have better robustness. The future work of the paper will focus on construction of binocular stereo vision feature indexing and content-based retrieval.

Acknowledgment

This work was financially supported by the Science and Technology Innovation Project of Ministry of Culture of China (No. 2014KJCXXM08).

References

[1] Song, Y., Tang, J.H., Liu, F., Yan, S.C. (2014). Body surface context: A new robust feature for action recognition from depth videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 24 (6) 952-964.

[2] Zhao, Y., Liu, Z. C., Cheng, H. (2013). RGB-depth feature for 3D human activity recognition. *China Communications*, 10 (7) 93-103.

[3] Cui, W.H., Wang, W.M., Liu, H. (2012). Robust hand tracking with refined CAMshift based on combination of depth and image features. *In:* IEEE International Conference on Robotics and Biomimetics, 1355-1361. IEEE, (December).

[4] Karpushin, M., Valenzise, G., Dufaux, F. (2014). Local visual features extraction from texture+depth content based on depth image analysis. *In:* IEEE International Conference on Image Processing, 2809-2813. IEEE, (October).

[5] Lu, C.W., Jia, J.Y., Tang, C.K. (2014). Range-sample depth feature for action recognition. *In:* IEEE Conference on Computer Vision and Pattern *Recognition*, 772-779. IEEE, (June).

[6] Ma, X., Wang, H. B., Xue, B. X., Zhou, M. G., Ji, B., Li, Y. B. (2014). Depth-based human fall detection via shape features and improved extreme learning machine. *IEEE Journal of Biomedical and Health Informatics*, 18 (6) 1915-1922.

[7] Jalal, A., Uddin, M. Z., Kim, T. S. (2012). Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home. *IEEE Transactions on Consumer Electronics*, 58 (3) 863-871.

[8] Liu, Y.Z., Lasang, P., Siegel, M., Sun, Q.S. (2015). Geodesic invariant feature: A local descriptor in depth. *IEEE Transactions on Image Processing*, 24 (1) 236-248.

[9] Yang, S., Liu, J.Y., Fang, Y.M., Guo, Z.M. (2016). Joint-feature guided depth map super-resolution with face priors. *IEEE Transactions on Cybernetics*, (99) 1-13.

[10] Dalal, N., Triggs, B. (2005). Histograms of oriented gradients for human detection. *In:* IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 886-893. IEEE, (June).

[11] Spinello, L., Arras, K.O. (2011). People detection in RGB-D data. *In:* IEEE/RSJ International Conference on IntelligentRobots and Systems, 3838-3843. IEEE, Sept. 2011.

[12] Lin, Y. C., Wei, S. T., Fu, L. C. (2014). Grasping unknown objects using depth gradient feature with eye-in-hand RGB-D sensor. *In:* IEEE International Conference on Automation Science and Engineering,1258-1263. IEEE, (August).

[13] Liang, C. W., Chen, E. Q., Qi, L., Guan, L. (2016). Improving action recognition using collaborative representation of local depth map feature. *IEEE Signal Processing Letters*, 23 (9) 1241-1245.

[14] Zhang, H., Parker, L. E. (2016). CoDe4D: Color-depth local spatio-temporal features for human activity recognition from RGB-D videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 26 (3) 541-555.

[15] Wang, N. B., Gong, X. J., Liu, J. L. (2012). A new depth descriptor for pedestrian detection in RGB-D images. *In:* 21st International Conference on Pattern Recognition, 3688-3691. IEEE, (November).

[16] Kang, Y. S., Lee, S. B., Ho, Y. S. (2014). Depth map upsampling using depth local features. *Electronics Letters*, 50(3) 170-171.

[17] Boykov, Y., Veksler, O., Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11) 1222-1239.

[18] Duan, F. F. (2016). Consistent depth maps estimation from binocular stereo video sequence. *Journal of Shanghai Jiaotong University(Sience)*, 21 (2) 184-191.

[19] Lee, S. B., Ho, Y. S. (2013). Temporally consistent depth map estimation for 3D video generation and coding. *China Communications*,10 (5) 39-49.

[20] Bu, X. B., Wu, B., Jia, H. W. (2013). Research on feature extraction and classification of apples' near infrared spectra. *Computer Engineering and Applications*, 49 (2) 170-173.

[21] Zhang, X., Jiang, J., Liang, Z. H., Liu, C. L. (2010). Skin color enhancement based on favorite skin color in HSV color space. *IEEE Transactions on Consumer Electronics*, 56 (3) 1789-1793.

[22] Jiang, L. C., Shen, G. Q., Zhang, G. X. (2009). An image retrieval algorithm based on HSV color segment histograms. *Mechanical & Electrical Engineering Magazine*, 26 (11) 54-57.

[23] Richardt, C., Orr, D., Davies, I., Criminisi, A., Dodgson, N. A. (2010). Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid. *In:* Proceedings of the 11th European Conference on *Computer Vision*, 510-523. DBLP, (September).

Author Biographies

Fengfeng Duan was born in Anhui (China), during 1982. He graduated and received his Ph.D. degree in Computer Science from Communication University of China (China), in 2016.

He is currently a researcher in Hunan Normal University, China. His research interests concern communication and information system, image processing and retrieval.