Jiang *et al. SpringerPlus (2016) 5:1629*
DOI 10.1186/s40064-016-3279-x

SpringerPlus

RESEARCH Open Access

Human motion segmentation and recognition using machine vision for mechanical assembly operation

Qiannan Jiang*, Mingzhou Liu, Xiaoqiao Wang, Maogen Ge and Ling Lin

*Correspondence:
1191758741@qq.com
School of Mechanical and Automotive Engineering, Hefei University of Technology, 193 Tunxi Road, Hefei 23009, Anhui, China

Abstract

The observation, decomposition and record of motion are usually accomplished through artificial means during the process of motion analysis. This method not only has a heavy workload, its efficiency is also very low. To solve this problem, this paper proposes a novel method to segment and recognize continuous human motion automatically based on machine vision for mechanical assembly operation. First, the content-based dynamic key frame extraction technology was utilized to extract key frames from video stream, and then automatic segmentation of action was implemented. Further, the SIFT feature points of the region of interest (ROIs) were extracted, on the basis of which the characteristic vector of the key frame was derived. The feature vector can be used not only to represent the characteristic of motion, but also to describe the connection between motion and environment. Finally, the classifier is constructed based on support vector machine (SVM) to classify feature vectors, and the type of therblig is identified according to the classification results. Our approach enables robust therblig recognition in challenging situations (such as changing of light intensity, dynamic backgrounds) and allows automatic segmentation of motion sequences. Experimental results demonstrate that our approach achieves recognition rates of 96.00 % on sample video which captured on the assembly line.

Keywords: Motion recognition, Mechanical assembly operation, Key frame extraction, SIFT feature points, Support vector machine

Background

Gilbreth (1917) said that the world's largest waste is the waste of motion. Therefore, we should find the issues of action and improve workers' movement through motion analysis, thereby eliminating the waste of time, alleviating fatigue and improving work efficiency (Salvendy 2001; Florea et al. 2003). The first steps of motion analysis are to decompose, identify and record motion sequence, which are performed through the repeated manual observation of operations, in a general way. This is the main reason why motion analysts have the problem of heavy workload and inefficiency. Therefore, in order to reduce the workload and improve the efficiency of motion analysis, a novel method is needed, which could automatically accomplish the motion segmentation, recognition and record by machine.

With the development of image acquisition technology and image processing technology, human motion recognition based on machine vision has become an active area with applications in several domains such as visual surveillance (Lao et al. 2009), video retrieval, patient monitoring (Jalal et al. 2012) and human–computer interaction (Poppe 2010; Turaga et al. 2008). However, segmentation and recognition of continuous human motion for mechanical assembly operation is both one of the most common and most difficult problems in the area of machine vision. Campbell et al. (1996) utilized 3D data gathered in real-time from stereo video cameras and HMMs to learn and recognize gestures. Davis et al. (1997) proposed a view-based approach for the representation and recognition of action and used 18 aerobics exercises to test it. This technique was also incorporated into the Kids Room: an interactive, narrative play-space for children. Carlsson and Sullivan (2001) presented a matching algorithm which matches shape information extracted from individual frames to store prototypes representing key frames of the action to recognize specific actions of tennis players in long video sequence. Stauffer and Grimson (2000) developed a visual monitoring system to track people in indoor environments and outdoor environments using multiple cameras. Laptev (2005) applied space–time features to detect walking people, along with occlusions and dynamic cluttered backgrounds. Ellis et al. (2007) proposed a novel dynamic context model to classify the behavior of group interactions in smart meeting room environment. Niebles et al. (2008) proposed a novel unsupervised learning algorithm using latent topic models to categorize and localize the human actions, and this algorithm was tested on challenging datasets: the KTH human motion dataset, the Weizmann human action dataset, and figure skating actions dataset. Liu et al. (2008) proposed a human actions recognition method using multiple features which was tested on publicly available data sets. Chen et al. (2009) applied MoSIFT algorithm to recognize human actions in surveillance videos. Recognition of human actions in surveillance videos is a part of the TRECVID Event Detection task. Shi et al. (2011) proposed a discriminative semi-Markov model approach to solve the inference problem of simultaneous segmentation and recognition, and this model was verified on KTH dataset. Zhang et al. (2012) proposed a human action recognition approach based on an improved BOW model and latent topic model. Cui et al. (2012) proposed a Matrix-based approach for unsupervised human action categorization. The above two approaches were tested on two datasets, the KTH datasets and WEIZMANN datasets. Jiang et al. (2012a) proposed a shape-motion prototype-based approach, and this approach achieved recognition rates of 92.86 % on a large gesture dataset with dynamic backgrounds. Bousmalis et al. (2013) presented the infinite HCRF (iHCRF), which is capable of automatically learning the optimal number of hidden states for a classification task. The UNBC-McMaster Shoulder Pain Expression database was used for experimental verification. Ellis et al. (2013) presented algorithms for reducing latency of recognizing actions in designing interactive, and evaluated these algorithms on two existing datasets—the MSR Action 3D dataset and the MSRC-12 Kinect Gesture dataset. Reddy and Shah (2013) proposed an action recognition method using the scene context information and motion features to solve the action recognition problem on a HMDB51 dataset. Park and Trivedi (2005) developed a driver activity analysis system using a rule-based decision tree to recognize driver activity. Kim and Medioni (2008) proposed a key visual functionality to recognize six kinds of human actions (walking,

sitting, raising hand, lying, falling, and standing up) in the Intelligent Home environment. Cisek et al. (2014) used foreground-weighted histogram decomposition to recognize human action on three datasets:UCF50, HMDB51, and Olympic sports. Slama et al. (2014) developed an accurate action-recognition system using learning on the Grassmann manifold to recognize human action and activity on three public 3D action datasets: MSR-action3D, UT-kinect and UCF-kinect datasets. Yu and Lee (2015) proposed a dynamic classification model called supervised MTRNN for human action classification and the inference of mental states. 16 samples corresponding to each action were used to test this model. Guo and Chen (2015) applied regularized multi-task learning base on spatial–temporal feature to recognize human actions on the TJU dataset. However, the issue of segmentation and recognition of continuous human motion for mechanical assembly operation has not been previously discussed in the literature.

The segmentation and recognition of continuous human motion for mechanical assembly operation is a problem that has challenge sex in the field of machine vision.

1. A continuous video sequence contains a series of motions, and there is no obvious boundary between motions. In addition, the speed of motion affect the motion time. Therefore, unsupervised motion segmentation is quite difficult.
2. Imaging conditions such as light intensity and image background are constantly changing in a realistic assembly environment.
3. There is close connection between human motion and the reality environment, so the types of human motion not only depend on the characteristics of motions, but also is directly related to objects in the environment.

The above three points affect the robustness of the segmentation and recognition algorithm of continuous human motion in the assembly environment. In this paper, we proposed a novel automatic segmentation and recognition method for assembly operations. In order to solve the first problem, the proposed method can segment motion in continuous video using dynamic key frame extraction technology based on content. In order to solve the second problem, we use SIFT feature points matching to find the feature points of ROIs, and the SVM is used to build the classifier to classify the feature vector, which make the recognition of motion possess good robustness. The method also extracts the feature points of the human hand and the assembly work piece, and the displacement vector between the feature points is used to represent the relationship between the human and the environment, and then the third problem is solved.

This paper presents an automated segmentation and recognition method that effectively accomplishes the observation, decomposition and record of human motion using SVM and machine vision in assembly environment. The remainder of this paper is organized as follows. Second section describes key frame extraction and feature points extraction of ROIs. Third section presents the proposed motion recognition algorithm. Fourth section demonstrates the experimental validation process, and conclusions are drawn in the last section.

Key frames extraction and preprocessing

Key frame extraction

In order to reduce the number of the images to be processed and ensure the action recognition algorithm is timely, key frame extraction from the video stream is required before image analysis and recognition. The criterion of key frame extraction is to calculate the dissimilarity between image frames (Chatzigiorgaki and Skodras 2009). Content-based key frame extraction is based on the change of visual information such as color and texture of the image (Lew et al. 2006). If the visual information of a frame varies significantly, the frame is the key frame and is extracted. A novel content-based dynamic key frame extraction algorithm is implemented as follow:

Assume that the video stream contains S images, and every frame image has $P_1 \times P_2$ pixels. The gray value of each pixel is represented by $H(a,\ b)$ $(a = 0,\ 1,\ 2, \ldots, P_1 - 1;\ b = 0,\ 1,\ 2,\ \ldots, P_2 - 1)$. Each image is divided into $K_1 \times K_2$ sub-blocks with same size, and every sub-block contains $\bar{P}$ pixels.

$$\bar{P} = \frac{P_1 \times P_2}{K_1 \times K_2} \tag{1}$$

The average gray value of each sub-block is expressed as follows:

$$E[H_t] = E\left[H_t(i,j)\right] = \frac{1}{\bar{P}} \sum_{k=1}^{\bar{P}} H(a,b) \tag{2}$$

$$\left(i = 0,\ 1,\ 2, \ldots, K_1 - 1; j = 0,\ 1,\ 2, \ldots, K_2 - 1; t = 0,\ 1,\ 2, \ldots, N - 1\right)$$

where i and j represent the row of sub-block and the column of sub-block respectively. $E\left[H_t(i,j)\right]$ is the average gray value of the sub-block which is in the ith row and jth column. Therefore, the average gray value $E(H)$ of each image and the dispersion degree $\sigma^2\left(\mathrm{E}\left[H_t(i,j)\right]\right)$ of gray value of every sub-block can be obtained.

$$E(H) = \frac{1}{K_1 \times K_2} \sum_{i=0}^{K_1-1} \sum_{j=0}^{K_2-1} E\left[H_t(i,j)\right] \tag{3}$$

$$\sigma^2(E[H_t]) = \left\{E\left[H_t(i,j)\right] - E(H)\right\}^2 \tag{4}$$

The feature vector of any image S is expressed as follows:

$$F_s = \left\{E[H_1]^s, \sigma^2(E[H_1])^s, \ldots, E[H_N]^s, \sigma^2(E[H_N])^s\right\} \tag{5}$$

The feature vector of the pth image is represented as F_p. The feature vector of the qth image is represented as F_q. Then the Euclidean distance of the two feature vectors is calculated by the following equation:

$$Dis\left(F_p, F_q\right) = sqrt\left\{\sum_{k=1}^{N} \left(E[H_t]^p - E[H_t]^q\right)^2 + \left(\sigma^2(E[H_t])^p - \sigma^2(E[H_t])^q\right)^2\right\} \tag{6}$$

The first frame is set as key frame, which is extracted from video stream, and the value of threshold T is calculated. The threshold T is a parameter used for determining whether the image is a key frame or not. The Euclidean distance of the two neighboring frame of images is obtained by Inter-frame Difference Algorithm and the set D of all Euclidean distances is composed. The set D of all Euclidean distances is expressed as follows:

$$D = \{Dis(F_1, F_2), Dis(F_2, F_3), \cdots, Dis(F_m, F_{m+1}), \cdots, Dis(F_{S-1}, F_S)\} \tag{7}$$

where the set D contains $S-1$ elements. And using the formula (8), the average value $\overline{Dis}$ of all elements is calculated. The value of $\overline{Dis}$ is taken as the value of T. The key frame is extracted as shown in Fig. 1.

$$\overline{Dis} = \frac{\sum_{m=1}^{S-1} Dis(F_m, F_{m+1})}{S-1} \tag{8}$$

Firstly, calculate the Euclidean distance between the feature vectors of the first key frame and following frames by function (6). If the distance is over T, the subsequent

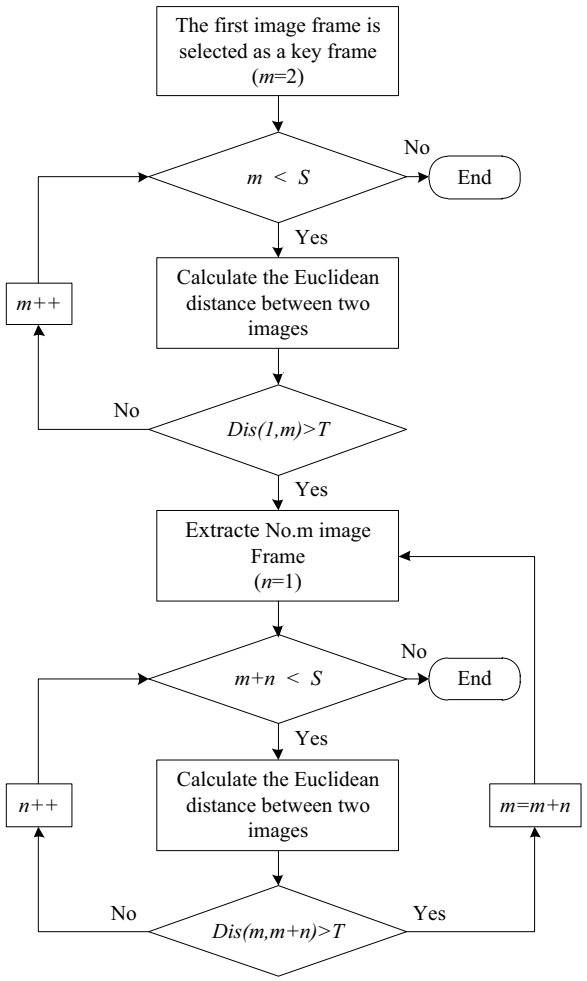

Fig. 1 Key frame extraction process

m frame is the key frame, which is extracted from the video stream. Extract key frames from video stream as described in the previous steps until the last frame is calculated. Following the above steps, we can get a key frame sequence of the video stream finally.

ROIs extraction of hand

Complexion is one of the most significant features of hand. Complexion has a favorable invariance for the image scaling, translation and rotation. It also has certain robustness to the change of image dimensional, capturing direction and angle. (Van den Bergh and Van Gool 2011; Kurakin et al. 2012). Therefore, by use of complexion information is the most effective and direct way to detect human hand. Range of skin color is much more compact and hand extraction is less susceptible to the effects of light and other objects in YCbCr and HSV space than in RGB space. But the color space transformation between RGB and HSV is more complicated than that between RGB and YCbCr, so human hand complexion is modeled and segmented in YCbCr space. First of all, the RGB color space is transformed into the YCbCr color space, threshold segmentation is made to the Cr component. Then, we segment the skin color regions with complex backgrounds based on skin color model. After segmentation is finished, multiple connected domains are obtained. The empty of the connected domains are filled by closing operation of mathematical morphology. Finally, the hand objects can be identified according to the hand shape feature. This process of hand detection is shown in Fig. 2.

ROIs extraction of workpiece

Template matching algorithm is widely used in projects as a classic recognition algorithm. Input template and target templates are compared to decide whether the two templates are the same. Using the results of template matching to determine whether an image contains target templates and find the locations of target templates in the image (Jain and Zongker 1997). The advantage of template matching is lesser amount of calculation, and the disadvantage is that the robustness of template matching is poor and recognition result is directly dependent on template construction (Pereira and Pun 2000). Because the shape characteristic of workpiece is really good, the shape-based template matching algorithm is adopted to detect the ROIs of workpiece. When serious block, confusion and nonlinear illumination changes happen on image, the algorithm still has a very high recognition rate (Mohan et al. 2001; Breuer et al. 2007). The algorithm always uses the inner product of the gradient vector of pixel as similarity measure. The best matching location is searched by calculating the minimum inner product. At the same

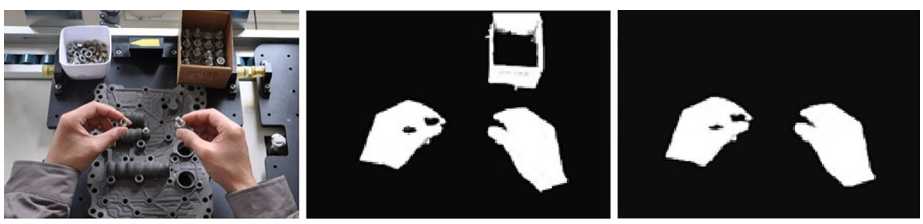

Fig. 2 The process of hand detection. **a** Original hand image. **b** Skin color regions. **c** Hand regions

time, In order to speed up the searching process, the pyramid of the multi-level images was built. Pyramid image-matching strategy is adopted to realize fast ROI extracting (Tanimoto 1981).

Feature points extraction of ROI

Lowe (2004) proposed an optimized SIFT characteristic operator. The operator is invariant to luminance, perspective, translation, rotation and scale change. It also maintains a better matching result, despite the external factors such as shape and background change, environmental noise, and occlusion. Thus, the matching algorithm based on SIFT descriptor has been successfully applied in object recognition, robot location, fingerprint and face recognition and other fields (Mortensen et al. 2005; Li et al. 2010; Mikolajczyk and Schmid 2005).

The SIFT feature points extraction and matching algorithm includes three steps as follow:

- Detect SIFT feature points of ROIs from the sample image and key frame image
  SIFT feature point detection is essentially local extreme point detection in different Gaussian (DOG) scale spaces. Each pixel point of ROI is compared with its 26 neighborhood points which contain eight adjacent points in the same scale space and eighteen points in vertically adjacent scale spaces, if a point is a maximum or minimum, the point is the feature point of ROI in the scale (Lowe 2004). Because the DOG value is sensitive to noise and edge, local extreme points need to be further checked for determination of feature points (Aprovitola and Gallo 2014).
- Generate SIFT feature point descriptor
  Firstly, samples were collected from the neighborhood window of feature points and were subjected to gradient histogram calculating. The peak position of the histogram is the main direction of the gradient of the feature points. The feature point's main direction makes the feature points possess rotational invariance (May et al. 2012). In order to improve the stability of matching, 128-dimensional vector is use to describe SIFT feature point descriptor.
- Obtain SIFT feature points set of key frame using feature points matching
  The Euclidean distance between SIFT feature points of sample image and SIFT feature points of key frame is calculated, and its value is used as the similarity measure between feature points. Through the above mentioned calculation, nearest neighbor and second nearest neighbor of the SIFT feature point of sample image can be found from all SIFT feature points in the key frame. The ratio of closest distance A and second closest distance B is $d(a,b)$. If $d(a,b)$ is less than $T_{a,b}$, the SIFT feature point of template image and the SIFT feature point of key frame are matched (Jiang et al. 2012b). By use of the above method, the SIFT feature points set of ROI in key frame is obtained through feature points matching.

Proposed motion recognition algorithm

First of all, the multiple displacement vector sets can be obtained by calculating the displacement between feature points of different ROIs, and the displacement vector sets

are the feature vectors of key frame. The feature vector classifier should be obtained by training and testing before motion recognition. Eventually, input the feature vectors of key frames into the classifier to identify the kind of feature vector, and identify the type of therblig based on the time-sequence characteristics of key frames and the judgment rules of particular scenario. Motion recognition algorithm flow is illustrated as shown in Fig. 3.

Get feature vectors from key frame

The ROIs of image consist of human hands and two workpieces in mechanical product assembly process. The assembling process needs to use both hands to accomplish a task. Left hand's feature points set M_1 (contains m_1 feature points), right hand's feature points set M_2 (contains m_2 feature points), workpiece 1 feature points set N_1 (contains n_1 feature points) and workpiece two feature points set N_2 (contains n_2 feature points) can be obtained through the SIFT feature point matching. If the displacement vectors set between feature points of points set A and feature points of points set B is denoted by (A, B), we can get six displacement vectors sets (M_1, M_2), (M_1, N_1), (M_1, N_2), (M_2, N_1), (M_2, N_2) and (N_1, N_2) by any combination of all four feature points sets, and feature points of hands are the starting points of the displacement vector. The above six displacement vectors sets constitute the feature vectors of key frame, and each kind of displacement vectors set includes R_i ($i = 1, 2, 3, 4, 5, 6$) displacement vectors. R_i satisfies the following equation:

$$R_i = A^j \times B^k \tag{9}$$

where A^j is the number of feature points in points set A and B^k is the number of feature points in points set B.

Construct classifier based on SVM

SVM is a kind of novel machine learning method which developed on the basis of statistical learning theory. Compared with traditional learning method, SVM employs structural risk minimization criterion to minimize the learning error and simultaneously decrease the generalization error. It has been developed for solving the classification and

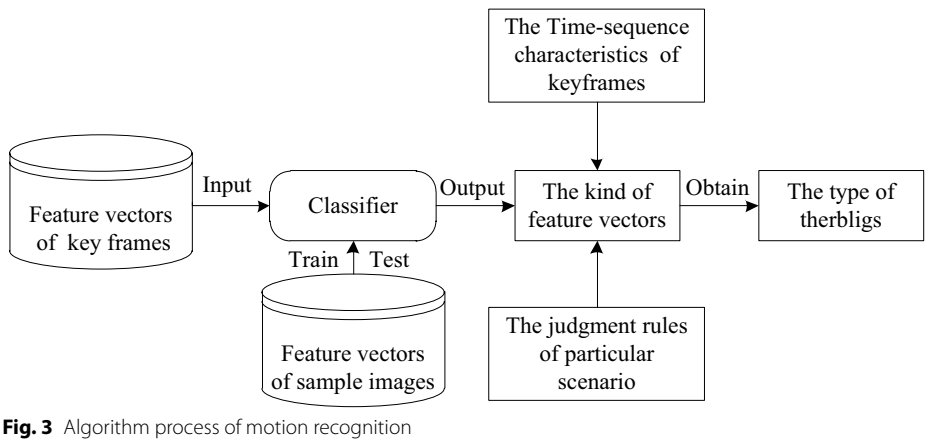

Fig. 3 Algorithm process of motion recognition

regression problems, and it also is a major achievement in machine learning research in recent years (Vapnik 2000). There are many unique advantages of SVM in solving small samples, nonlinear and high-dimensional pattern recognition problems (Brezak et al. 2012). It can not only find the global optimal solution from the limited sample information, but also describe the train sample accurately and identify any test samples without error (Vapnik 2000). For the method based on back propagation (BP) neural network or Radial basis function (RBF) neural network, when the dimension of input vector is more, it may cause the network scale is too large, which leads to difficulty in training and other issues. While the computations load of the SVM method is almost independent of input vector dimension, therefore, it is suitable to deal with the problem of large input dimension. In recent years, SVM-based method had been widely applied in texture classification, time series forecast and face recognition because it need much less training time and carry out an efficient calculation (Benkedjouh et al. 2015; Kao et al. 2014; He and Li 2014). Therefore, SVM will be used as a classification algorithm of motion recognition method in this paper.

The process of classifier construction based on SVM is as follows:

- Sample preparation

  Sample images were screened out from video stream, and the pictures are handled as the above- described way, then the displacement vectors set between ROIs can be obtained, which is a feature vector of image. The type of feature vector is obtained through expert evaluating method before sample training. If two ROIs are close to each other, the type of feature vector is labeled as $z = \{1\}$; If two ROIs are not close to each other, the type of feature vector is labeled as $z = \{0\}$. A feature vector is expressed as $\boldsymbol{T} = (A, B) = \{(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k), \ldots, (x_K, y_K)\}$ $(k = 1, 2, \ldots, K)$, where (x_k, y_k) represents the coordinate of each displacement vector and K is the number of displacement vectors. Therefore, $(\boldsymbol{T}, z)$ is regarded as input sample for the training of SVM classifier.
- Classifier construction

  On the basis of the distribution of displacement vectors in a two-dimensional coordinate system, it is shown that the problem is linearly inseparable. To solve the nonlinear problem, the calculation is completed in the low dimensional space firstly. Then SVM maps data from sampling space to higher dimensional characteristic space by kernel functions that satisfy Mercer condition. Eventually, the optimal separating hyperplane is constructed in the high dimensional feature space. Minimize the distance between all sample points and hyperplane, so that the nonlinear problem is converted into linear divisible problem to get optimum relation (Boser et al. 1996; Cristianini and Shawe-Taylor 2000). We map the training samples into a high-dimensional feature space via a nonlinear mapping determined by a kernel function. The mapping process is shown in Fig. 4.

If $\boldsymbol{X} = \boldsymbol{T} = (A, B) = \{(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k), \ldots, (x_K, y_K)\}$, $(k = 1, 2, \ldots, K)$, $Y_k = z$, samples set is expressed as $\{(\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2), \ldots, (\boldsymbol{X}_k, Y_k), \ldots, (\boldsymbol{X}_K, Y_K)\}$ containing K samples, where $\boldsymbol{X}_k$ is the input samples and $\boldsymbol{X}_k \in R^2$, and Y_k ($Y_k = 0$ or $Y_k = 1$) is the expect output of the sample data. Map the sample points in two-dimension space to the

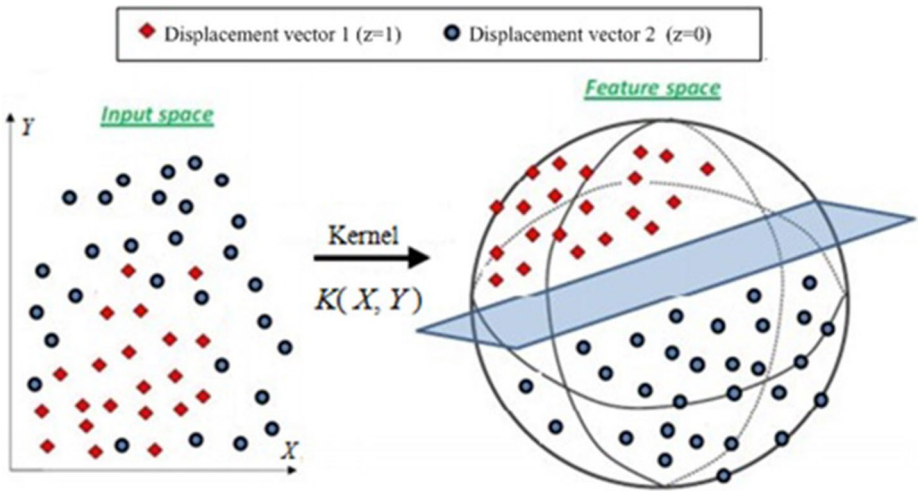

Fig. 4 The mapping process *Displacement vector 1* indicates that two ROIs are close to each other; *Displacement vector 2* indicates that two ROIs are not close to each other; Input space is the low dimensional space; Feature space is the high dimensional space which determined by kernel functions

feature space, and then the optimal classification discriminant function $F(X)$ is obtained according to the SVM learning algorithm. $F(X)$ is the feature vector classifier.

$$F(X) = \operatorname{sgn}\left(\sum_{k=1}^{K} \alpha_k^* Y_k \cdot K(\boldsymbol{X}, \boldsymbol{X}_k) + b^*\right) \tag{10}$$

Motion recognition

Motion recognition is essential to identify the type of therblig in mechanical product assembly process. A simple assembly operation normally includes four types of therbligs: Reach, Grasp, Move and Assemble. Displacement vector can be divided into three kinds for simple mechanical product assembly process: displacement vector (M, N) of hands relative to workpieces, displacement vector (M, M) between hands and displacement vector (N, N) between workpieces. Three types of classifiers $F_1(X)$, $F_2(X)$ and $F_3(X)$ are trained based on the above displacement vectors. Then, the type of therblig in key frames can be recognized by the process of Fig. 5.

First of all, input different kinds of displacement vectors into different classifiers to identify the kind of the displacement vector. Then, identify the type of therblig based on the time-sequence characteristics of key frames and the judgment rules of particular scenario.

- If the key frame contains class II displacement vector, class IV displacement vector and class VI displacement vector, the type of therblig is Reach.
- If the key frame contains class I displacement vector, class IV displacement vector and t class VI displacement vector, the type of therblig is Grasp or Move.
- If the key frame contains class I displacement vector, class III displacement vector and class V displacement vector, the type of therblig is Assemble.

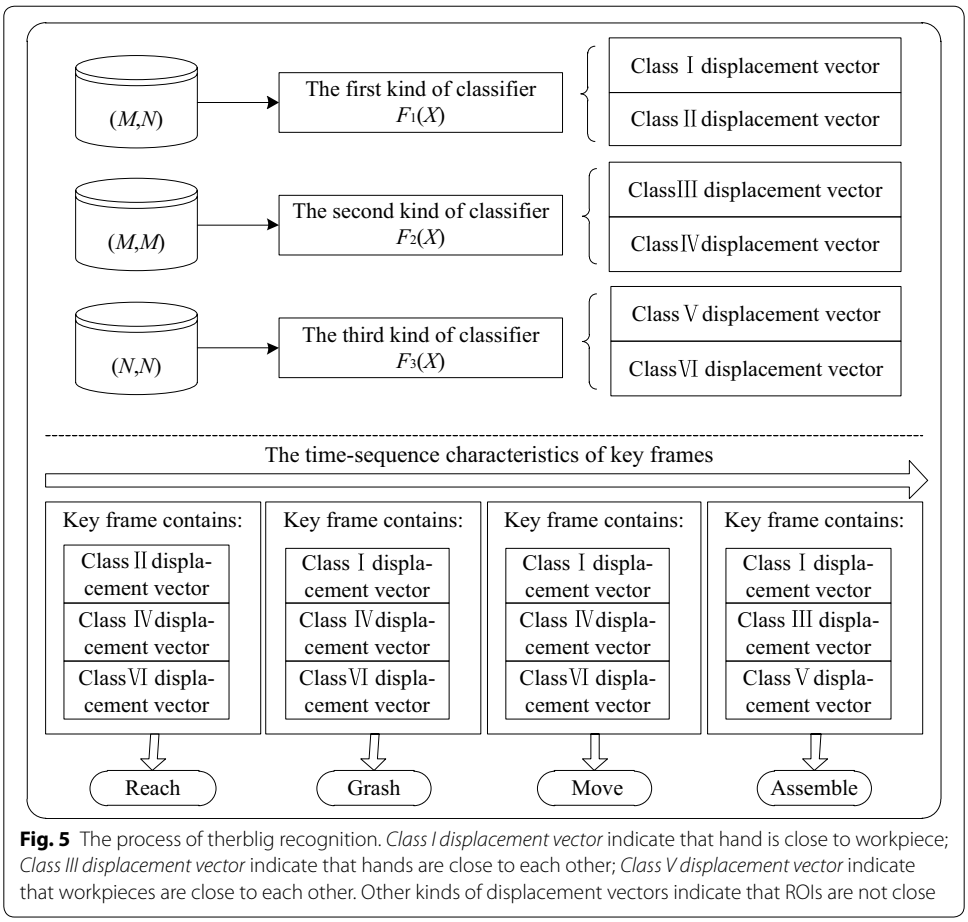

Fig. 5 The process of therblig recognition. *Class I displacement vector* indicate that hand is close to workpiece; *Class III displacement vector* indicate that hands are close to each other; *Class V displacement vector* indicate that workpieces are close to each other. Other kinds of displacement vectors indicate that ROIs are not close

There is much different visual information between the image of containing Grasp and the image of containing Move, so both of these images are key frames and are extracted from video streaming. Since the images of video stream are arranged in chronological order, the key frame of containing Grasp is before the key frame which contains Move. Therefore, Grasp and Move can be separated in accordance with the time-sequence characteristics of key frames.

Implementation

In this section, an experiment simulated the real bolt assembly operation of mechanical product. It was provided to verify the feasibility and robustness of the vision-based motion recognition method. The experimental framework consists of an operator, workpieces and an image acquisition system. The workpieces contain bolts (M8 × 15) and hex nuts (M8 × 1.5). The image acquisition system contains the light source, a CCD color camera, a computer and a prototype system in the mechanical product assembly line, as shown in Fig. 6. The light source is a kind of artificial daylight (DH, LER2-90SW2). An industrial CCD progressive scan RGB color camera (DH, SV2000GM/C 1/1.8″ 1628 × 1236, active pixels) mounted with a 16 mm lens (DH, model M3Z1228C-MP) was applied to capture the motion images of operator. The motion images were transmitted from the camera to computer through the Gigabit Ethernet. The HALCON

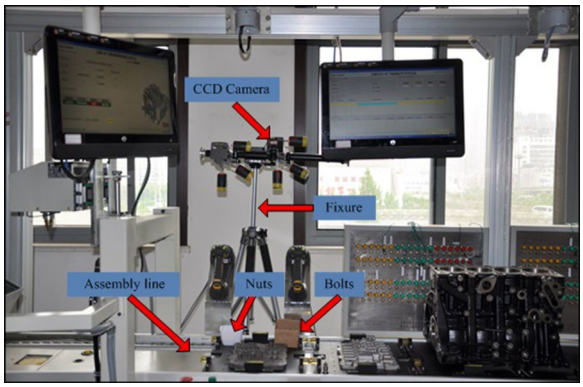

Fig. 6 Prototype system in the mechanical product assembly line

software was used for picture processing and the MATLAB software was used for model analysis in the experiment, and the configuration of the computer is listed as follows: OS: Windows 7(32 bit); CPU: Intel Core i5-3337U; RAM: 6 GB; MATLAB version: 2011b; HALCON version: 11.0. In the actual environment, the background and light intensity are both changed.

Experimental procedure:

- A sample image of left hand was captured by using machine vision system, and the left hand was in a state of grabbing bolt (This example picture is named L-G-P);
- A sample image of right hand was captured by using machine vision system, and the right hand was in a state of grabbing (This example picture is named R-G-P);
- A sample image of hands and the hands was captured by using machine vision system, and the hands were in a state of mounting bolts (This example picture is named D-A-P);
- The video of bolt assembly operations was captured by the machine vision system, and the operator worked with both hands. The video stream contains 500 motion cycles, and each cycle contains four types of therbligs: Reach, Grasp, Move and Assemble.

Image preprocessing

Sample images and the image in video stream need a series of preprocessing before the feature extraction. Firstly, the impacts of noise and light were eliminated by use of Gaussian filter (Zhang and Parker 2011; Wu et al. 2012), and the RGB color space was converted to the YCbCr color space. Then the color threshold range was set from 140 to 160 in the Cr component, multiple connected regions were obtained. Then the empty areas of connected regions were filled by closing operation. Eventually, the connected regions larger than 35,000 pixels were extracted. The connected regions were the ROI of hand. After obtaining the ROIs of hand, images were matched to bolt template and nut template; the ROIs of workpiece were obtained.

Motion recognition

The ROIs of sample images were obtained after preprocessing, then the SIFT feature points of ROIs were detected and SIFT feature point descriptor was generated. Key frames were extracted by the content-based dynamic key frame extraction algorithm. Four examples of key frames were shown as Fig. 7. The as seen in Fig. 7, the technology of video key-frame extraction not only realized key frame extraction but also segmented motion sequences automatically. The feature points of ROIs in key frames were obtained based on the SIFT feature points of sample images, and the feature vectors of key frames were acquired by calculating the displacement vector between feature point sets. Hands were used as starting points to calculate the displacement vector between hand and workpiece; left-hand was taken as starting point to calculate the displacement vector between hands; hex bolt was used as starting point to calculate the displacement vector between workpieces. According to the coordinates of feature points and the calculation rules in the above, the feature vectors of key frames were obtained. The feature vector of each key frame contains six displacement vector sets. The six displacement vector sets were respectively expressed as (M_1, M_2), (M_1, N_1), (M_1, N_2), (M_1, N_1), (M_2, N_2) and (N_1, N_2). The set (M_1, M_2) contains 3364 displacement vectors. The set (N_1, N_2) contains 441 displacement vectors. Each of the four remaining sets of feature vector contains 1218 displacement vectors.

When using the content-based dynamic key frame extraction algorithm for key frame extraction, parameters including K_1, K_2 and T must be determined in advance. In this study, the optimal values of the three parameters were determined as $K_1 = 16$, $K_2 = 12$, $T = 14.14$. 2524 key frames were extracted from the video stream which contains 500

Fig. 7 Four key frames

motion cycles. Before classifier training and data testing, the experts identified the therblig type of each key frame. In this paper, the SVM model was constructed by using Gauss kernel function (Cherkassky and Ma 2004).

$$K(X, X_k) = \exp\left(-\frac{|X - X_k|^2}{2\delta^2}\right) = \exp(-\gamma |X - X_k|^2) \tag{11}$$

We used cross validation method to select the most accurate parameters (C, γ) as the parameter of classifier model. Among these key frames, 1516 key frames (containing 300 motion cycles) were used for training and 1008 key frames (containing 200 motion cycles) were used for the testing. Each key frame of the above contained three classes of displacement vectors: (M, N), (M, M) and (N, N) which were used to select the parameters of F_1 (X), F_2 (X) and F_3 (X) respectively. The optimal parameters of classifier models were obtained by cross validation: $C_1 = 2.0$, $\gamma_1 = 0.03125$; $C_2 = 3.5$, $\gamma_2 = 0.05556$; $C_3 = 5.0$, $\gamma_3 = 0.125$. Finally, all the 2524 samples were used for final validation. The result confusion matrix is shown in Fig. 8, with an average performance of 96 %.

According to the confusion matrix shown in Fig. 8, we can see that Grasp and Move are more easily to be confused, which are consistent with our observations. Actually some Grasp and Move instances are even difficult to differentiate for people. Grasp and Move do have similar feature vector, but the spatial–temporal characteristic is slightly different. Unfortunately, the spatial–temporal characteristics are difficult to be reflected from the pixel-wise. The experimental result on motion recognition demonstrates the robustness of the proposed method.

In order to demonstrate the advantage of the proposed methods, we compare the results with other reported ones, which are shown in Table 1. The five compared state-of-the-art methods require a pre-segmentation of the continuous motion sequence into

Fig. 8 Confusion matrix of motion categories, with an average performance of 96 %. *Rows* represent actual motion categories; *Columns* represent recognition results of our method

Table 1 Comparison of different methods on the video of bolt assembly operations

Methods	Recognition objects	Accuracy (%)
Our method	Continuous motion sequence	96.00
Schuldt et al. (2004)	Motion segments	71.75
Niebles et al. (2008)	Motion segments	83.30
Jiang et al. (2012a)	Motion segments	84.35
Reddy and Shah (2013)	Motion segments	93.40
Lu et al. (2015)	Motion segments	95.05

elementary segments, a tedious manual operation. We can see that the result of our method is in parallel with the best supervised method (Lu et al. 2015), which recognized human motions by two-level Beta process hidden Markov model, and gain 2.6 % improvement from Reddy and Shah (2013), which used Sphere/Rectangle-tree to construct a novel framework for motion categorization.

Record of motion

The long video has 500 motion cycles,and each cycle contains four types of therbligs. So there is 2000 therbligs in the video stream. The above study has demonstrated that 2524 key frames were extracted from the video stream. It is clear that the number of key frames is larger than the number of therbligs. Therefore, the adjacent key frames in the video may belong to the same type of therbligs. In order to avoid repeated recording of motion, when the motion types of adjacent key frames are different, the all motions are recorded, else, we record one motion. The results of therblig recognition were compared with actual therbligs, and the comparison results of the first four cycles are as shown in Fig. 9.

Figure 9 shows that the vision-based motion recognition method can accurately identify the type of therbligs, but there are minor differences between the motion times obtained by using the proposed method and observed values. The cause of this phenomenon is that the identification object of the vision-based motion recognition method proposed in this paper is key frame, but not each frame of video stream. If two key frames A and B contain different therbligs, the time interval between them is the time of

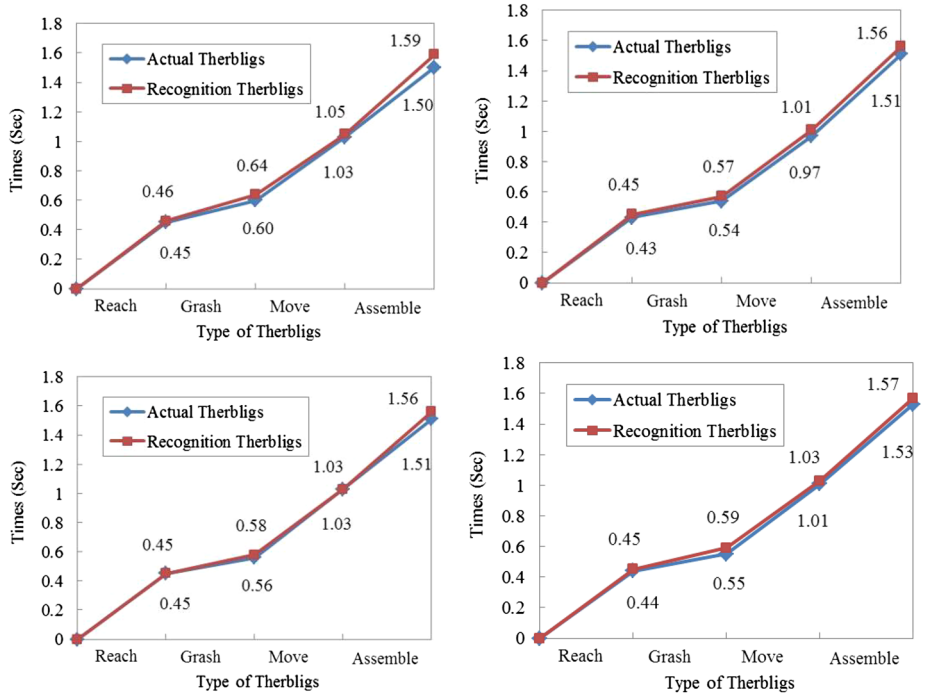

Fig. 9 The comparison results of the first four cycles. The actual therbligs were obtained by the experts through the observation, decomposition and recording of motion. The recognition therbligs were the results obtained with such method in this paper

the therblig in key frame A. When each frame of video stream is used to recognize, the time difference in Fig. 8 would not exist. However, identifying all the frames one by one will greatly reduce timeliness and generate a large amount of redundant data during the recording. Therefore, when the time accuracy of motion analysis is not high, it is reasonable to use the key frame as the object of motion recognition.

Conclusions

This study proposed a novel machine-vision-based motion segmentation and recognition method for mechanical product assembly operation. The experiment results have demonstrated that the proposed method can segment motion automatically, identify the type of the motion accurately and record each motion and its time. The study makes the following major contributions. (1) The relationship between motion and objects were established by using the displacement vector between SIFT points in different ROIs; (2) In order to improve the timeliness of the proposed method, the key frame extraction technology was applied to reduce the number of images to be processed, and the image processing technique was applied to reduce the number of pixels to be processed; (3) The proposed motion recognition algorithm was constructed based on SVM to classify feature vectors; (4) This method not only accomplishes the motion segmentation, recognition and record automatically but also reduces the workload of motion analysts and improves the efficiency of motion analysis. In addition, the proposed algorithm is general, which can be straightforwardly extended to other motion recognition fields.

Abbreviations

SVM: support vector machine; ROIs: regions of interest; SIFT: scale invariant feature transform; DOG: difference of gaussian; BP: back propagation; RBF: radial basis function.

Authors' contributions

QJ carried out the studies, participated in the building of continuous human action recognition algorithm for mechanical assembly operation and drafted the manuscript. XW participated in the design of content-based dynamic key frame extraction algorithm. ML participated in the extraction of feature vectors and constructed SVM-based classifier. MG carried out the design of the study and analyzed the experimental data. LL conceived of the study, and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work is supported by the National Science Foundation of China (No. 51375134).

Competing interests

The authors declare that they have no competing interests.

Informed consent

Informed consent was obtained from all individual participants included in the study.

Received: 4 April 2016 Accepted: 9 September 2016
Published online: 21 September 2016

References

Aprovitola A, Gallo L (2014) Edge and junction detection improvement using the Canny algorithm with a fourth order accurate derivative filter. In: Tenth international conference on signal-image technology and internet-based systems (SITIS), 2014. IEEE, New York, pp 104–111

Benkedjouh T, Medjaher K, Zerhouni N, Rechak S (2015) Health assessment and life prediction of cutting tools based on support vector regression. J Intell Manuf 26(2):213–223

Boser BE, Guyon IM, Vapnik VN (1996) A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory. vol 5. ACM, New York, NY, USA, pp 144–152

Bousmalis K, Zafeiriou S, Morency L, Pantic M (2013) Infinite hidden conditional random fields for human behavior analysis. IEEE Trans Neural Netw Learn Syst 24(1):170–177

Breuer P, Eckes C, Müller S (2007) Hand gesture recognition with a novel IR time-of-flight range camera—a pilot study. In: Gagalowicz A, Philips W (eds) Computer vision/computer graphics collaboration techniques. Springer, Berlin, pp 247–260
Brezak D, Majetic D, Udiljak T, Kasac J (2012) Tool wear estimation using an analytic fuzzy classifier and support vector machines. J Intell Manuf 23(3):797–809
Campbell LW, Becker DA, Azarbayejani A, Bobick AF, Pentland A (1996) Invariant features for 3-D gesture recognition. In: Proceedings of the second international conference on automatic face and gesture recognition, 1996. IEEE, New York, pp 157–162
Carlsson S, Sullivan J (2001) Action recognition by shape matching to key frames. In: Workshop on models versus exemplars in computer vision, vol 1, p 18
Chatzigiorgaki M, Skodras AN (2009) Real-time key frame extraction towards video content identification. In: 16th international conference on digital signal processing, 2009. IEEE, New York, pp 1–6
Chen MY, Hauptmann A, Chen MY, Hauptmann A (2009) Mosift: recognizing human actions in surveillance videos. Ann Pharmacother 39(1):150–152
Cherkassky V, Ma Y (2004) Practical selection of SVM parameters and noise estimation for SVM regression. Neural Netw 17(1):113–126
Cisek A, Sch Fer W, Szczurek A (2014) Human action recognition across datasets by foreground-weighted histogram decomposition. In: 2014 IEEE conference on computer vision and pattern recognition (CVPR), vol 690. IEEE Computer Society, Washington DC, pp 764–771
Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge
Cui P, Wang F, Sun LF, Zhang JW, Yang SQ (2012) A matrix-based approach to unsupervised human action categorization. IEEE Trans Multimed 14(1):102–110
Davis JW, Bobick AF (1997) The representation and recognition of action using temporal templates. In: IEEE conference on computer vision and pattern recognition. pp 928–934
Ellis DPW, Poliner GE (2007) Identifying 'Cover Songs' with chroma features and dynamic programming beat tracking. In: IEEE international conference on acoustics, speech and signal processing, 2007. ICASSP 2007. vol 4, pp IV-1429–IV-1432
Ellis C, Masood SZ, Tappen MF, Laviola JJ, Sukthankar R (2013) Exploring the trade-off between accuracy and observational latency in action recognition. Int J Comput Vision 101(3):420–436
Florea NR, Kotapati S, Kuti JL, Geissler EC, Nightingale CH, Nicolau DP (2003) Cost analysis of continuous versus intermittent infusion of piperacillin-tazobactam: a time-motion study. Am J Health Syst Pharm 60(22):2321–2327
Gilbreth FB (1917) Fatigue study: the elimination of humanity's greatest unnecessary waste: a first step in motion study. Nature 99(2471):23
Guo W, Chen G (2015) Human action recognition via multi-task learning base on spatial–temporal feature. Inf Sci 320:418–428
He K, Li X (2014) A quantitative estimation technique for welding quality using local mean decomposition and support vector machine. J Intell Manuf 1–9. doi: 10.1007/s10845-014-0885-8
Jain AK, Zongker D (1997) Representation and recognition of handwritten digits using deformable templates. IEEE Trans Pattern Anal Mach Intell 19(12):1386–1390
Jalal A, Uddin MZ, Kim TS (2012) Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home. IEEE Trans Consum Electron 58(3):863–871
Jiang S, Pang G, Wu M, Kuang L (2012a) An improved K-nearest-neighbor algorithm for text categorization. Expert Syst Appl 39(1):1503–1509
Jiang Z, Lin Z, Davis LS (2012b) Recognizing human actions by learning and matching shape-motion prototype trees. IEEE Trans Softw Eng 34(3):533–547
Kao LJ, Lee TS, Lu CJ (2014) A multi-stage control chart pattern recognition scheme based on independent component analysis and support vector machine. J Intell Manuf 1–12. doi: 10.1007/s10845-014-0903-x
Kim K, Medioni GG (2008) Distributed visual processing for a home visual sensor network. In: Proceedings of the 2008 IEEE workshop on applications of computer vision, vol 00. IEEE Computer Society, Washington DC, pp 1–6)
Kurakin A, Zhang Z, Liu Z (2012) A real time system for dynamic hand gesture recognition with a depth sensor. In: signal processing conference (EUSIPCO), 2012 proceedings of the 20th European. IEEE, New York, pp 1975–1979
Lao W, Han J, De With PHN (2009) Automatic video-based human motion analyzer for consumer surveillance system. IEEE Trans Consum Electron 55(2):591–598
Laptev I (2005) On space-time interest points. Int J Comput Vision 64(2–3):432–439
Lew MS, Sebe N, Djeraba C, Jain R (2006) Content-based multimedia information retrieval: State of the art and challenges. ACM Trans Multimed Comput Commun 2(1):1–19
Li Y, Snavely N, Huttenlocher DP (2010) Location recognition using prioritized feature matching. In: Daniilidis K, Maragos P, Paragios K (eds) Computer vision–ECCV 2010. 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II, Springer, Berlin, pp 791–804
Liu J, Ali S, Shah M (2008) Recognizing human actions using multiple features, CVPR. In: IEEE Conference on computer vision and pattern recognition, 2008, IEEE, New York, pp 1–8
Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vision 60(2):91–110
Lu L, Yi-Ju Z, Qing J, Qing-ling C (2015) Recognizing human actions by two-level Beta process hidden Markov model. Multimedia Syst. doi:10.1007/s00530-015-0474-5
May M, Turner M, Morris T (2012) Analysing false positives and 3D structure to create intelligent thresholding and weighting functions for SIFT features. In: Ho YS (ed) Advances in image and video technology. Springer, Berlin, pp 190–201
Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. IEEE Trans Pattern Anal Mach Intell 27(10):1615–1630
Mohan A, Papageorgiou C, Poggio T (2001) Example-based object detection in images by components. IEEE Trans Pattern Anal Mach Intell 23(4):349–361

Mortensen EN, Deng H, Shapiro L (2005) A SIFT descriptor with global context. In: IEEE computer society conference on computer vision and pattern recognition, 2005. vol 1. IEEE, New York, pp 184–190
Niebles JC, Wang H, Li FF (2008) Unsupervised learning of human action categories using spatial-temporal words. Int J Comput Vision 79(3):299–318
Park S, Trivedi M (2005) Driver activity analysis for intelligent vehicles: issues and development framework. In: IEEE proceedings. Intelligent vehicles symposium, 2005, vol 1. IEEE, pp 644–649
Pereira S, Pun T (2000) Robust template matching for affine resistant image watermarks. IEEE Trans Image Process 9(6):1123–1129
Poppe R (2010) A survey on vision-based human action recognition. Image Vis Comput 28(6):976–990
Reddy KK, Shah M (2013) Recognizing 50 human action categories of web videos. Mach Vis Appl 24(5):971–981
Salvendy G (2001) Handbook of industrial engineering: technology and operations management. Wiley, New York
Schuldt C, Laptev I, Caputo, B (2004) Recognizing human actions: a local SVM approach. In: Proceedings of the 17th international conference on pattern recognition, 2004. IEEE, ICPR 2004. vol 3, pp 32–36
Shi Q, Cheng L, Wang L et al (2011) Human action segmentation and recognition using discriminative Semi-Markov Models[J]. Int J Comput Vis 93(1):22–32
Slama R, Wannous H, Daoudi M, Srivastava A (2014) Accurate 3D action recognition using learning on the grassmann manifold. Pattern Recogn 48(2):556–567
Stauffer C, Grimson WEL (2000) Learning patterns of activity using real-time tracking. IEEE Trans Pattern Anal Mach Intell 22(8):747–757
Tanimoto SL (1981) Template matching in pyramids. Comput Graphics Image Process 16(4):356–369
Turaga P, Chellappa R, Subrahmanian VS, Udrea O (2008) Machine recognition of human activities: a survey. IEEE Trans Circuits Syst Video Technol 18(11):1473–1488
Van den Bergh M, Van Gool L (2011) Combining RGB and ToF cameras for real-time 3D hand gesture interaction. In: IEEE workshop on applications of computer vision (WACV), 2011. IEEE, New York, pp 66–72
Vapnik V (2000) The nature of statistical learning theory. Springer Science & Business Media, Berlin
Wu D, Zhu F, Shao L (2012) One shot learning gesture recognition from rgbd images. In: IEEE Computer Society Conference on computer vision and pattern recognition workshops (CVPRW), 2012. IEEE, New York, pp 7–12
Yu Z, Lee M (2015) Real-time human action classification using a dynamic neural model. Neural Netw 69:29–43
Zhang H, Parker LE (2011) 4-dimensional local spatio-temporal features for human activity recognition. In : IEEE/RSJ International Conference on intelligent robots and systems (IROS), 2011. IEEE, New York, pp 2044–2049
Zhang X, Miao Z, Wan L (2012) Human action categories using motion descriptors. IEEE International Conference on image processing. IEEE, New York, pp 1381–1384