Journal of Computers Vol. 29 No. 5, 2018, pp. 202-221
doi:10.3966/199115992018102905016

Human Activities Recognition Based on a Multi-level Configuration

I-Cheng Chang[1*], Jhe-Sheng Lyu[2], Chun-Man Lin[3]

[1] Department of Computer Science and Information Engineering, National Dong Hwa University,
Hualien 974, Taiwan
icchang@mail.ndhu.edu.tw

Received 7 January 2018; Revised 7 July 2018; Accepted 7 August 2018

Abstract. Human activity recognition is an important topic in computer vision due to its various promising applications, such as video indexing and retrieval, human-computer interaction and security. This study brings forward an automatic system which can recognize multi-group actions. The proposed system is based on a two-level configuration: action group classification and action type recognition. The action group classification is to classify an input action into one of the action groups. Here, we adopt the motion vectors as the action features and Support Vector Machine (SVM) as the classifier. The action type recognition is to identify the type of the action within the group which is determined by action group classification. Since each action group has its own particular features, we utilize Pictorial Structures Model (PSM) to describe the variation of the human body and extracts suitable recognition features for the input action according to its action group. The recognition process utilizes a sparse representation-based method to learn the discriminative dictionary and recognizes the action. Two experiments are performed to show the efficiency of the proposed approach. In the first experiment, we evaluate the performance of the motion vectors feature, and it performs well in the classification of action groups. In the second experiment, the performance of the two-level recognition system for multi-group actions is evaluated. The proposed system can recognize 30 different actions of three groups with natural backgrounds, and the average recognition rate achieves 91.52%.

Keywords: action recognition, ball sport action, interaction recognition, multi-group action, PSM, sparse representation, SVM

1 Introduction

1.1 Motivation

In recent years, the topic of human activity recognition has gained tremendous attention in computer vision since the information of human activities can be applied to various promising applications, such as video indexing and retrieval, human-computer interaction and security. However, it is a challenging problem because of the environmental condition or the complexity of the action, such as illumination variabilities, occlusions, or multiple interacting moving objects. Also, the camera condition also affects the results, including the viewpoint, motion, and shake of the camera. Most of previous action recognition methods focused on the recognition of the specific action group, for example, the actions of daily life or the sports-related actions. The method used in one group cannot be applied to another one because different action groups possess different action features and one set of features is not suitable for all action groups. The paper brings forward an automatic system which can classify the group of the input action and extract the suitable recognition features to recognize its action type.

[*] Corresponding Author

1.2 Related Work

Action recognition has diverse applications in computer vision, so it gains tremendous attention, and many related studies have been proposed in the past. We classify the previous studies into two categories according to the number of humans in an action video: videos with single people and videos with multiple people. We review the related literature to investigate the development of technology in the following.

Single people action recognition. The single person actions can be daily life actions or sports-related actions. Because an object's silhouette is an important feature, various shape-based methods have been proposed. Gorelick and Blank [1] regarded a human action as a 3D shape formed by a sequence of silhouettes in the space-time volume. The proposed method utilized properties of the solution to the Poisson equation to extract space-time features. The results showed that this method is fast and does not require video alignment. Guo et al. [2] used manually collected silhouettes as input to extract the feature vectors by computing the Euclidean distance measurement between the center point and the nearest silhouette boundary point. Then the feature vectors are mapped to a nonlinear Riemannian manifold which manifold is locally similar enough to a linear space to allow one to do calculus. The experiments verified the performance of the proposed method on two publicly-available datasets, where the Weizmann human action dataset includes 10 common actions: bend, jumping jack, jump forward, jump in place, run, gallop sideways, skip, walk, wave one hand, and wave both hands, and UT-tower human action dataset contains 9 actions: pointing, standing, digging, walking, carrying, running, wave1, wave2, jumping. The results of the correct classification rates were 100% and 97.2%, respectively.

However, contour-based methods suffer from one weakness; namely, they assume the availability of accurate silhouette information. Generally, silhouette information is derived from foreground extraction, so the results are susceptible to illumination changes, occlusions and camera instability. There are also many studies utilizing interest points to represent actions. Laptev and Lindeberg [3] proposed spatial-temporal interest points that can be used for the representation of video data. They observed spatial interest points to describe significant local variations for video representation in both space and time. Laptev et al. [4] built space-time grids to coarsely describe interest points for action video classification. The method can tolerate background clutter, occlusions, and scale changes. Experiments showed the method can achieve 91.8% of recognition rate when applied to KTH action dataset [5], which contains six types of human actions, namely walking, jogging, running, boxing, hand waving and hand clapping. Guha and Ward [6] detected the key points and observed the corresponding spatiotemporal cubes to describe human actions. The work is one of the few to use Random Projection (RP) in a classification framework. It successfully reduced the computational cost, and the average accuracy achieved 97.8 % on Weizmann action dataset. This approach can also be applied to facial expressions classification. Hu and Wo [7] utilized the template matching method to extract the critical points of human in the edge gradient image and used the Hidden Markov Model (HMM) to construct the online behavior recognition classifier. The database includes four types of actions: standing, sitting, bending and squatting. Ni et al. [8] proposed multi-modality fusion schemes, which combined color and depth information. It used support vector machines (SVM) with different kernels to classify different behaviors, such as making a phone call, drinking water, eating a meal, going to bed, sitting down, standing up.

Besides interest-point-based representation, some papers represented human activity using motion-based representation. Chaudhry et al. [9] proposed to represent human action by using a histogram of oriented optical flow (HOOF) and extended the Binet-Cauchy kernels into nonlinear dynamical systems (NLDS). The method is simple and does not use any preprocessing steps. However, it achieves good results which show the average recognition rate on Weizmann database is 94.4%. Vo and Ly [10] observed using only the feature of shape is not enough because some different actions own similar human shapes. The method merged the spatial pyramid histogram of edge (shape of the object) and the histogram of oriented optical flow (movement of the object) to improve the action recognition performance. Moreover, it used LDA to create discriminative features that increased the recognition rate and adopted ANN for actions classification. The results of classification rates on KTH dataset and Weizmann dataset were 96.3 % and 100%, respectively.

Sports action classification problem is also an important topic. Rodriguez et al. [18] improved traditional template-based methods that cannot generate a single template for all action sequences. The proposed method captured the general intra-class variabilities of actions to generate a single action

MACH filter for action recognition. They analyzed the response of the filter in the frequency domain to avoid the high computational cost which occurred in traditional template-based approaches. The experiments were performed on Broadcast Television action dataset, where actions in this dataset include diving, golf swinging, kicking, lifting, horseback riding, running, skating, swinging a baseball bat, and pole vaulting. The average accuracy for this dataset was 69.2%. Traditional methods utilize k-means for visual vocabulary learning. However, this kind of approaches suffers from the size of the visual vocabulary. To deal with the issue, Liu et al. [19] used diffusion maps to learn a semantic visual vocabulary. It automatically learned a semantic visual vocabulary from abundant quantized mid-level features and measured their dissimilarity using diffusion distance for recognition. The experiment was performed on a YouTube dataset contains eight categories: volleyball spiking, trampoline jumping, soccer juggling, horseback-riding, diving, swinging, golf club-swinging, and tennis racquet-swinging. The average accuracy was 76.1%. Niebles et al. [20] used a temporal structure to train models for action recognition. In the recognition process, they try to find the best matching of the model to an input action based on the learned appearances and motion segment decomposition. The average recognition rate on Olympic Sports Dataset was 72.1%. Chen et al. [21] combined the sparse descriptors (histogram of oriented gradient and histogram of optical flow) over the interest points and the dense descriptors (Local Binary Pattern) over the local patches for human action recognition. The system divided each video into several motion segments and used the non-linear Support Vector Machines (SVM) with the RBF kernel for multi-class classification. Each segment obtained a confidence value from the trained classifier. The experiments are performed on the Olympic Sports Dataset (OSD), and average accuracy was 80.0%.

Multiple people interaction recognition. We classify the recognition of multiple people interaction into two types: two-person interaction and the crowd behavior. Related to two-person interaction, an early work Park and Aggarwal [11] used a hierarchical Bayesian network (BN) for the recognition of two-person interaction. The proposed method tracked the positions of the head, arms, and legs, and used the position, distance and area between two people to represent the activities. The recognized interaction contains approaching, departing, pointing, standing hand-in-hand, shaking hands, hugging, punching, kicking, and pushing. Blunsden et al. [12] interpreted a motion trajectory as a complete shape and used a hierarchical Bayesian network for classifying interactions between two people, where the dataset includes six interactions: walking together, approach, ignore, meet, fight and split. Interactive user interfaces have become very popular in recent years, Yun et al. [13] used the Kinect to obtain body skeleton information as features and then used a Multiple Instance Learning (MIL) based classifier for action classification. The system can recognize eight action categories: approaching, departing, kicking, punching, pushing, hugging, shaking hands, and exchanging. The results showed that the MILBoost classifier outperforms SVMs if there are irrelevant actions existing in the training data. [14] provided a survey for the work on human action recognition over the past decade.

On the other hand, most behavior recognition in the crowd is used to detect the abnormal or crime-oriented behavior in the videos. Fuentes and Velastin [15] built an intelligent surveillance system which can detect potentially dangerous situations. The system was composed of foreground segmentation, a blob selection, and a tracking algorithm. They compared sequence parameters with the semantic description of the events and used the low-level description of the predefined events to describe human behavior for abnormal detection. Ryoo and Aggarwal [16] considered the spatio-temporal relationship among feature points, thereby enabling to detect ten activities. They recognized the testing video via proposed spatio-temporal relationship match kernel, which measures the similarity by constructed histograms and classified them. Particularly several pairs of interacting persons executed the activities simultaneously. Cui et al. [17] modeled group activities based on social behavior analysis. They proposed energy potentials to represent the patch around the salient points. Because uncommon energy potentials indicate the abnormal events, the approach identifies these events by determining if the energy potentials are beyond the pre-defined threshold or not.

1.3 System Overview

This paper presents a multi-group action recognition system, which includes three groups: single person action, two-person interaction and ball sport action. The former two groups cover most of the actions used in the other recognition systems, and the latter one is a new one. Our overall framework for action recognition is shown in Fig. 1. The action group classification is to classify the input action. We calculate motion vectors of the input action to extract the classification feature. Then, the Support Vector Machine (SVM) is used to classify the group of the input action. The action type recognition uses Pictorial Structures Model (PSM) to describe the variation of the human body and extracts suitable recognition features for the input action according to its action group. After obtaining suitable recognition features, we recognize the input action via sparse representation and receive the recognition results.

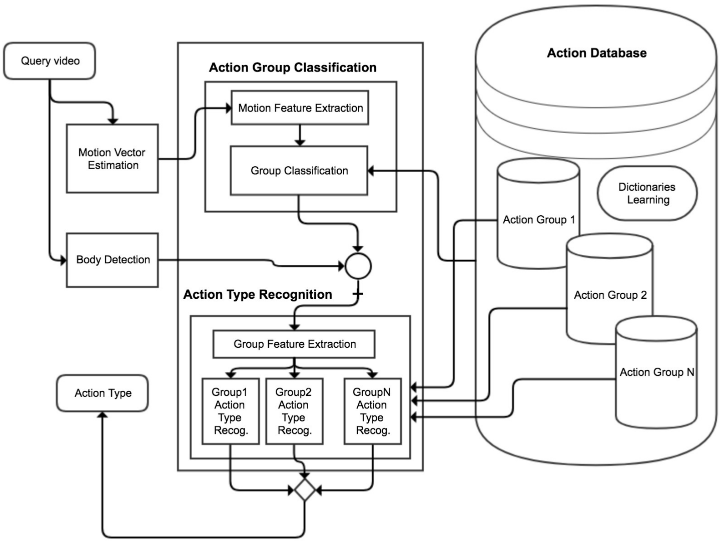


Fig. 1. System overview

In the work, we propose a multi-level action recognition system to recognize the actions of different groups. The proposed system can extract proper features of the input action for action recognition according to the action group classification results. The actions of ball activities are included in our database.

The paper is organized as follows. Section 2 describes the extraction of global classification features and group classification. Section 3 shows the action type recognition using the sparse representation. Section 4 presents experimental results for our two-level action recognition system, and conclusions are drawn in Section 5.

2 Multi-group Action Classification

Most of the existing approaches to action recognition consider recognizing the single-group actions, and these methods cannot be applied to recognize the multi-group actions. We require different features to recognize different group actions. Each group has particular features to represent actions, and those features can effectively recognize different action within the group.

After reviewing the actions of different groups, we observe that the vector variations of different action groups are significantly different. Each group has its special variation. Fig. 2 shows three action groups, where the location, magnitude, and direction of motion vectors are different, and each action has its distribution. For instance, the motion vectors of two-person interactions distribute around two different areas. However, that of single-person action is located in one area. Therefore, the motion vector field is adopted to classify the group of input action.

2.1 Feature Extraction

Since motion information is suitable to describe the activities in the video, it has been widely used for action recognition. Rohrbach et al. [41] extract histograms of oriented gradients (HOG), flow (HOF), and motion boundary histograms (MBH) as motion information to recognize cooking activities in a dense optical flow field. In the work, we adopt the optical flow field [22-23] as the motion features for our group classification task.

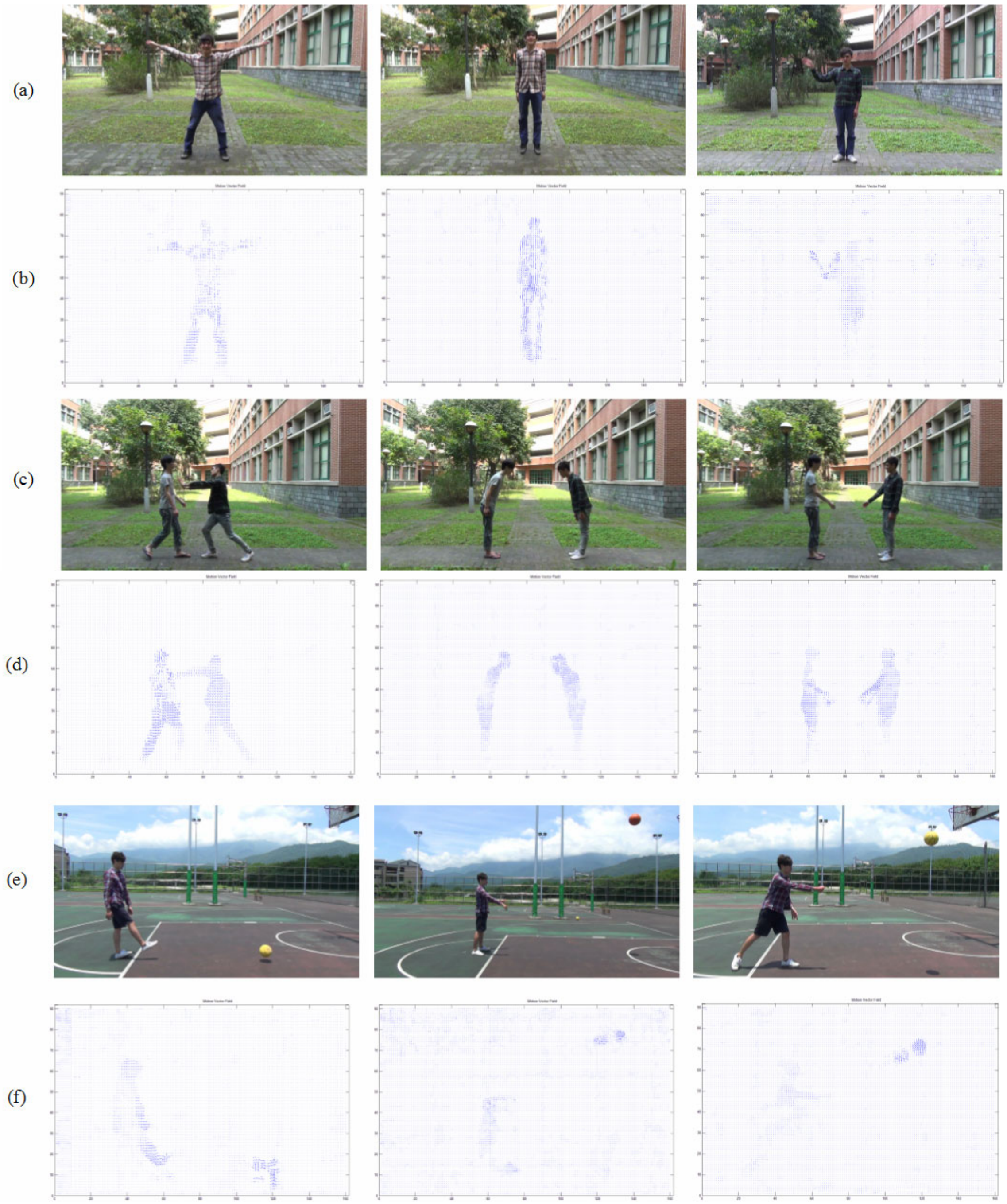
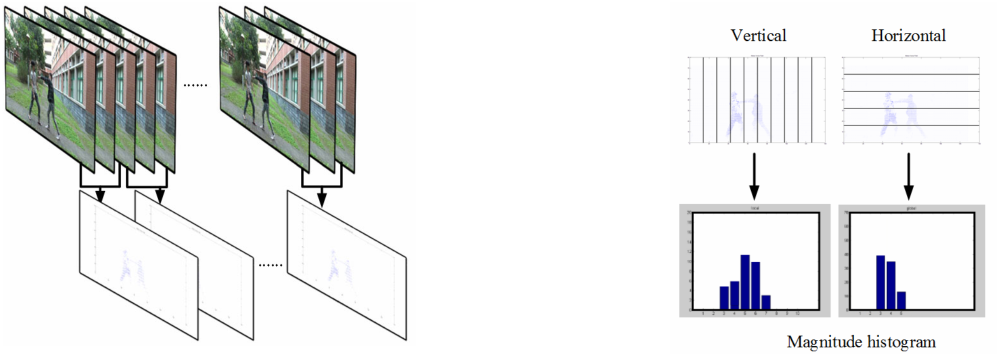


Fig. 2. Motion vectors of different action groups: the row (a), (c) and (d) show the actions of the different group. Moreover, the row (b), (d) and (f) show the corresponding motion vector field

Magnitude Histogram. Because actions of the same group occupy the nearby region, the location information of flow vectors can be used to express the action groups. We divide the optical flow field into ten and five slices to calculate a magnitude histogram for each slice along the horizontal and vertical orientation, respectively.

The motion vector histogram is defined as $h_{i,t} = \sum_{j \in S_i} F_j$, where $h_{i,t}$ is magnitude histogram of motion vectors in the i^{th} slice at time t , $i = \{1,2,3,...,15\}$, and F_j is the j^{th} flow vector, S_i is the set of motion vectors which locate in the i^{th} slice. This proposed method is used to locate and calculate the magnitude histogram of motion vectors, the magnitude histogram as shown in Fig. 3.



(a) Optical flow field

(b) The histogram of horizontal and vertical directions

Fig. 3. Magnitude histogram

Direction histogram. To represent the motion vector direction, we adopt the histogram of oriented optical flow (HOOF) proposed by Chaudhry et al. [24]. HOOF features improve the robustness of the optical flow method. When a person moves through a scene, it leads to a very particular optical flow distribution. However, optical flow distribution may be inconsistent if the activity was performed at different scale. We utilize HOOF features that represent the optical flow distribution unaffected by scale and directionality of motion. We calculate optical flow for the whole image sequence and record the angle of each flow vector. HOOF features provide a normalized histogram, which indicates the amount of flow vector in bin $1,\cdots,B$ at time t.

After computing the motion vector of all video frames, we integrate two histograms to form a 45-dimensional feature. The feature format is shown in Fig. 4.

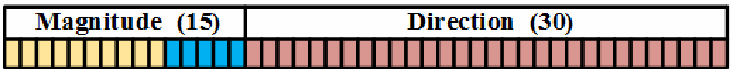


Fig. 4. Action group classification feature: there are 15-dimensional of magnitude histogram, yellow and blue parts are magnitude histogram of motion vectors along the horizontal and vertical orientations in the slice, respectively; the pink part is 30-dimensional for direction histogram

2.2 Classification Based on SVM

After representing action videos by motion feature, we are going to assign the video to its action group. To solve this problem, we adopt support vector machine (SVM) as the classifier. Support vector machine (SVM) classifiers are commonly used in the field of machine learning [25-26]. It trains a classifier by finding an optimal separating hyperplane for classification. SVM was introduced first for binary classification and extended later for multi-class classification. The original input space is mapped into a higher dimensional feature space for classification by the kernel function. Fig. 5 illustrates a flowchart of our action group classification.

3 Human Action Type Recognition

In the section, we present the feature of video by using Pictorial Structures Model (PSM). The data are combined to form a feature vector for action representation and integrated into the training set. The training set utilizes a sparse representation-based method to learn the discriminative dictionary and optimizes it for recognition.

3.1 Action Representation

The feature representation of human action plays an important role in the recognition performance. We take advantage of the PSM [33] method to overcome unstable factors for the recognition system caused by various circumstances. PSM detects the location of human body parts and combines that information

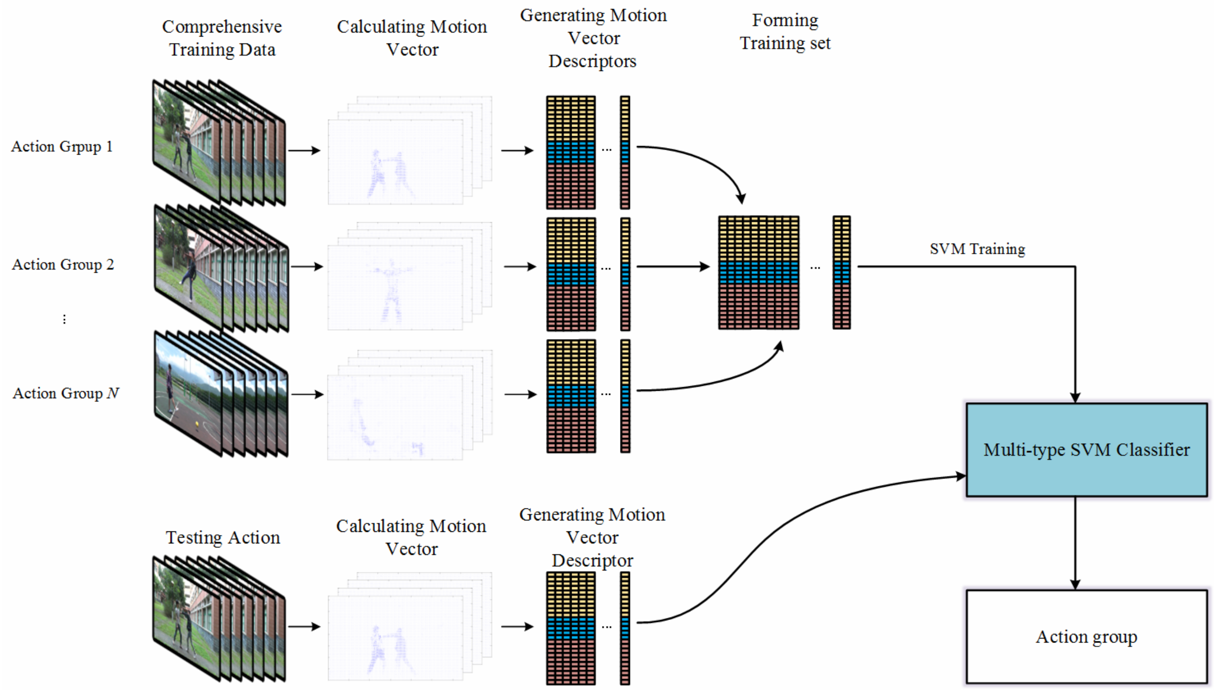

Comprehensive Training Data
Calculating Motion Vector
Generating Motion Vector Descriptors
Forming Training set
Action Grpup 1
Action Group 2
Action Group N
SVM Training
Multi-type SVM Classifier
Testing Action
Calculating Motion Vector
Generating Motion Vector Descriptor
Action group

Fig. 5. Flowchart of the action group classification: in the training stage, we convert the action video into image sequences and calculate motion vectors. These motion vectors form a training set which is used to learn a model for an SVM classifier

into feature vectors to express action sequence. In addition, we further utilize human detection results of PSM to detect ball position.

Body parts position detection. It is not easy to locate human posture from videos because the human parts have a different appearance when moving in the scene. Besides, the conditions include not only illumination variabilities and viewpoints but also different clothes wearing, posture, non-rigid deformation and other visual properties. The factors will increase the uncertainty of detection and localization. The PSM detection system provides a flexible model to describe an object using the composition of multiple parts. The appearance of body parts is quite suitable to describe human postures in static images. In this paper, PSM is used for body parts detection.

PSM is a flexible human model composed of several blocks. It uses the multiscale mixture deformable model to describe changeable human motion and is composed of two models: the appearance model and the deformation model. The appearance model is a description of human appearance. We utilize the Histogram of Oriented Gradient (HOG) feature to model the description of the appearance model. The deformation model represents the relationship between the joints of the skeleton, where a kinematic tree is used to model it. However, PSM has a better detection accuracy on the head and torso parts than on the hands and feet. The appearance of the hands and feet are useful clues for recognition, but they are difficult to detect. For this reason, we utilize the method proposed by Zhan [42] which incorporates the temporal information to limit the searching range of PSM. Given different weights over time, this approach improves the detection accuracy on the hands and feet.

Model parameters estimation. PSM model is a tree structure which each node inside represents a body part, and an associated appearance model is used to describe the part's exterior. The relationships between nodes are described by the deformation model. We utilize the method proposed by Felzenszwalb and Huttenlocher [35] to estimate appearance and deformation models. Given the model parameter $\theta=(u,E,c)$, where $u=\{u_1,...,u_m\}$ are the appearance parameters of each body part, where m is the number of body parts. E is a set of edges and $c=\{c_{ij}\mid(i,j)\in E\}$ indicates the connected edge.

Let training set be, where n is the number of images, and $L=\{l_1,l_2,...,l_n\}$ is the human configuration labelled manually. The parameter estimation can be written as:

$$p(I_1,...,I_n,L_1,...,L_n \mid \theta)=\prod_{k=1}^{n} p(I_k,L_k \mid \theta) \tag{1}$$

We use maximum likelihood (ML) to estimate the parameter θ, and assume the condition probability term of the right-hand side is generated independently. Since $P(L \mid I)=P(I \mid L)P(L)$, the ML estimation can be expressed as:

$$\theta^{*}=\arg\max_{\theta}\prod_{k=1}^{n} p(I_k \mid L_k,\theta)\prod_{k=1}^{n} p(L_k \mid \theta) \tag{2}$$

The first term in this equation is the appearance model of the body parts, while the second term is the deformation model. The parameter θ is the set of connection parameters. Both of them can be estimated independently.

Finally we get:

$$E^{*}=\arg\max_{E}\prod_{(i,j)} q(u_i,u_j)=\arg\min_{E}\sum_{(i,j)\in E} -\log q(u_i,u_j) \tag{3}$$

A result of the PSM human pose estimation is shown in Fig. 6. There are 26 parts of the human body which include two parts of the head, four parts of the torso and twenty of human limbs, respectively.

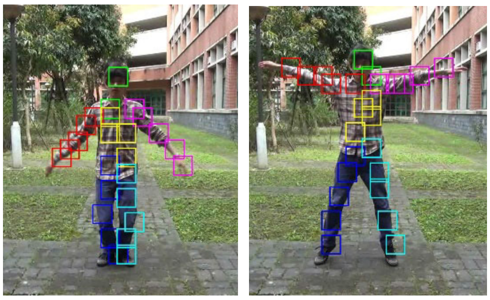
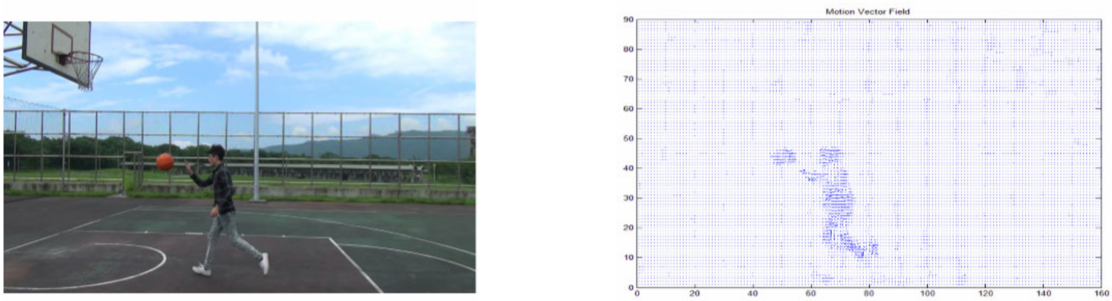


Fig. 6. Results of pose estimation: PSM model detection results contains 26 body parts in different cases

Ball position detection. In order to extract features for ball sport action representations, we detect not only body position but also ball location in the video. We combine the body position and the ball location information for ball sport actions representation. The ball position detection is based on the PSM pose estimation results, and further determines the position of the ball. We can observe that the motion vector is caused by human activity including ball motions. In the action sequence, the ball is not always close to the people. In some frames, the ball is far away from the people. Therefore, we can use this information to get the ball. As illustrated in Fig. 7.

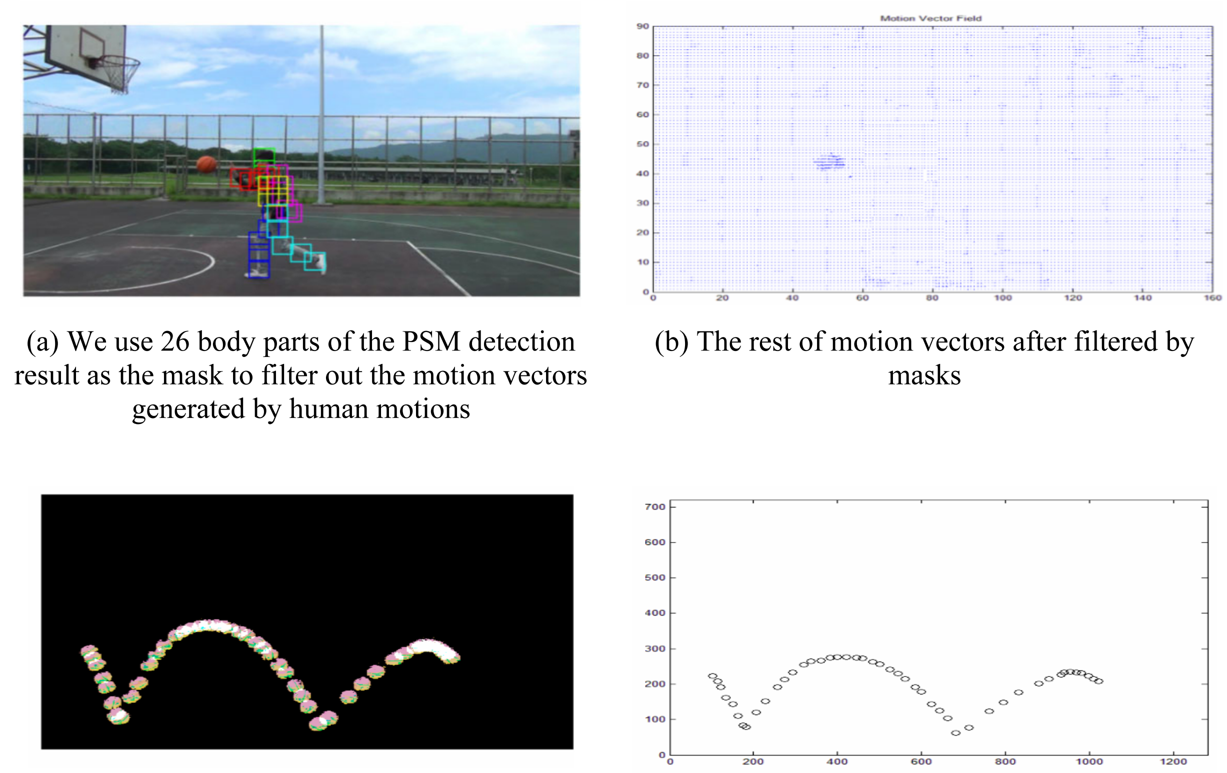


(a) The original image in action sequence

(b) The displacement between two frames generates motion vector

Fig. 7. Motion distribution of scene

When the human and the ball are separated, we use the body parts of the PSM detection result as the masks, and those masks, after expansion, are used to filter out the motion vectors generated by human motions. Then, the rest of the motion vectors are related to ball displacement. The key idea underlying this approach is that we track the position of color information in this region to represent the ball continuous trajectory in the action sequence. This approach can locate coordinate positions for different color balls which is not limited to a par-ticular color. The result of ball position detection is illustrated in Fig. 8.



(a) We use 26 body parts of the PSM detection result as the mask to filter out the motion vectors generated by human motions

(b) The rest of motion vectors after filtered by masks

(c) Ball coordinate position in the HSV color space

(d) Continuous ball's trajectory in the action sequence

Fig. 8. Result of ball position detection

Action feature description. After obtaining the body parts and ball position information via feature extraction, we can compute feature vectors based on the action group to represent human motions. First, we define operator $\Psi_{T_i} = \left\{ F_{T_i}{}^{p}, F_{T_i}{}^{d}, F_{T_i}{}^{v}, F_{T_i}{}^{a} \right\}$, where T_i is the action group, $T_i = 1,2,3$, and they are single person, two person interactions, and ball sport actions, respectively. The operator Ψ_{T_i} selects corresponding T_i according to the result of the action group classification to describe human actions. The operator extracts the feature set based on the results of the PSM pose estimation, and ball position detection for action type recognition. Our feature set contains four different forms: position, distance, velocity and area. We explain each operator Ψ_{T_i} as follows.

Let $p_{i,t} \in \Re^2$ and $v_{i,t} \in \Re^2$ be the location and velocity of body part i at time t in image sequence. Let T be all frames within the sequence. The action feature is a single vector as a combination of all computed features $F_{T_i}(\cdot\,;t)$, where $t \in T$.

Operator Ψ_1: It describe the human features of posture changes in the action video within the single person action group. Feature $F_{T_i}(\cdot\,;t)$ is shown as follows:

(a) Body part position: The body part position feature F_1^p is defined as the relative position of body part at time t. It is defined as:

$$F_1^p(i;t) = p_{i,t} \tag{4}$$

where *i* depicts the torso and four limbs, $t \in T$.

(b) Body part distance: The body parts distance feature F_1^d is defined as the Euclidean distance between two body parts. It represents the relative distance between two parts at time t. It is defined as:

$$F_1^d(i,j;t) = \left\| p_{i,t} - p_{j,t} \right\| \tag{5}$$

where $i\ and j$ are two body parts from the torso and limbs, $t \in T$.

(c) Body part velocity F_1^v: The velocity feature is the velocity of one body part. It is defined as:

$$F_1^v(i;t_1,t_2) = \frac{p_{i,\mathrm{t}_2} - p_{i,\mathrm{t}_1}}{t_2 - t_1} \tag{6}$$

where p_{i,t_1} and p_{i,t_2} are the position at time t_1 and t_2, i indicates the different body parts of head and limbs.

(d) Area spanned by body parts: The body parts area feature F_1^a represents the area of the area spanned by the three body parts. It is defined as:

$$F_1^a(i,j,k;t) = \left\langle p_{i,t}, p_{j,t}, p_{k,t} \right\rangle \tag{7}$$

where $\left\langle p_{i,t}, p_{j,t}, p_{k,t} \right\rangle$ indicate the area spanned by body parts p_i, p_j, p_k at time t, $t \in T$.

Operator Ψ_2: PSM detects body parts of two people in the second group action, and we use multiple features to describe the interactions between two persons. Let x be the one person, and y be the other one. The relationship features between two persons, $F_{T_i}(\cdot\ ;t)$ is defined as:

(a) Body part position: The body parts position feature F_2^l is defined as the relative coordinates of body part at time *t*. It is defined as:

$$F_2^p(x,y,i;t) = \left(p_{x^i,t}, p_{y^i,t} \right) \tag{8}$$

where p_{x^i}, p_{y^i} are the location of torso and four limbs of x and y, $t \in T$.

(b) Body part distance: The body parts distance features F_2^d is defined as the Euclidean distance of body parts between two persons. It represents the relative distance between two parts at time *t* and is defined as:

$$F_2^d(x,y,i;t) = \left\| p_{x^i,t} - p_{y^i,t} \right\| \tag{9}$$

where p_{x^i}, p_{y^i} are two parts of the body of x and y, and *i* indicates the part of the torso or limbs, $t \in T$.

(c) Body part velocity: The feature F_2^v represents the velocity of body parts at time *t* and it is defined as:

$$F_2^v(x,y,i;t) = \left(v_{x^i,t}, v_{y^i,t} \right) \tag{10}$$

where $v_{x^i,t}\ and\ v_{y^i,t}$ are the velocity of x and y at time t, where *i* indicates the different body parts of the head and limbs.

(d) Body part area: The feature F_2^a represents the area spanned by three body parts, that is, torso, left and right foot. It is defined as:

$$F_2^a(x,y,i,j,k;t) = \left(\left\langle p_{x^i,t}, p_{x^j,t}, p_{x^k,t} \right\rangle, \left\langle p_{y^i,t}, p_{y^j,t}, p_{y^k,t} \right\rangle \right) \tag{11}$$

where $\left\langle p_{x^i,t}, p_{x^j,t}, p_{x^k,t} \right\rangle, \left\langle p_{y^i,t}, p_{y^j,t}, p_{y^k,t} \right\rangle$ indicate the area spanned by body parts p_i, p_j, p_k of x and y at time t, $t \in T$.

Operator Ψ_3 : The relationship between the human and ball has the significant discrimination in the ball sport action group. We extract not only the location of body parts but also the ball position to represent this action group. Let $\left\{ F_{T_i}{}^{p}, F_{T_i}{}^{d}, F_{T_i}{}^{v}, F_{T_i}{}^{a} \right\}$ be the feature set, and the feature $F_{T_i}(\cdot\ ;t)$ between the body parts and ball are defined as:

(a) Position of body parts and object: The body parts and object position feature F_3^p is defined as the relative coordinates of body part and ball at time t. It is defined as:

$$F_3^p(i;t) = p_{i,t} \tag{12}$$

where p_i is the location of the torso, four limbs, and the ball at time *t*.

(b) Distance between body parts and the object: The distance F_3^d is defined as the Euclidean distance between two points. It represents the relative distance between parts and ball at time t. It is defined as:

$$F_3^d(i,o;t) = \left\| p_{i,t} - p_{o,t} \right\| \tag{13}$$

where $p_{i,t}$ is the location of body parts of the torso and the limbs, and $p_{o,t}$ is the ball location at time t.

(c) Velocity of body parts and the object: The feature F_3^v represents the velocity for the body parts and ball, and it is defined as:

$$F_3^v(i,o;t) = \left(v_{i,t}, v_{o,t} \right) \tag{14}$$

where $v_{i,t}$ *and* $v_{o,t}$ are the velocity of the part i and ball o at time t.

(d) Area spanned by body parts: The feature F_3^a represents the area spanned by three bodies. It is defined as:

$$F_3^a(i,j,k;t) = \left\langle p_{i,t}, p_{j,t}, p_{k,t} \right\rangle \tag{15}$$

where $\left\langle p_{i,t}, p_{j,t}, p_{k,t} \right\rangle$ indicate the area spanned by body parts p_i, p_j, p_k at time t.

Covariance descriptor of motion features. Our system uses the covariance matrix of motion feature ([36-37]) as a basis for sparse representation. A set of basis vectors after learning obtained using learned dictionaries for comparison and recognition. Then the feature-covariance matrix, which captures the position, distance, velocity and area information of body parts and ball to form the feature vector, provides a good discriminative representation for the particular action group. The motion feature covariance matrix representation C_S is expressed as:

$$C_S = \frac{1}{|S|} \sum_{s \in S} (\mathrm{f}(s) - \mu_F)(\mathrm{f}(s) - \mu_F)^T \tag{16}$$

where $\mu_F = \sum_{s \in S} \frac{1}{|S|} \mathrm{f(s)}$ is the mean feature vector of all the training data, the dimension of the covariance matrix is related to the dimension of the feature vectors. If f(s) is d-dimensional, the C_S only has $(d^2 + d)/2$ independent numbers and d is usually much smaller than, thus C_S usually lies in a much lower-dimensional space. Regardless of the number of samples, the covariance matrix has the ability to provide lower-dimensional feature representation. The input action sequence selects the most suitable operator Ψ_{T_i} based on the action group classification result to calculate the covariance matrix as depicted in Fig. 9.

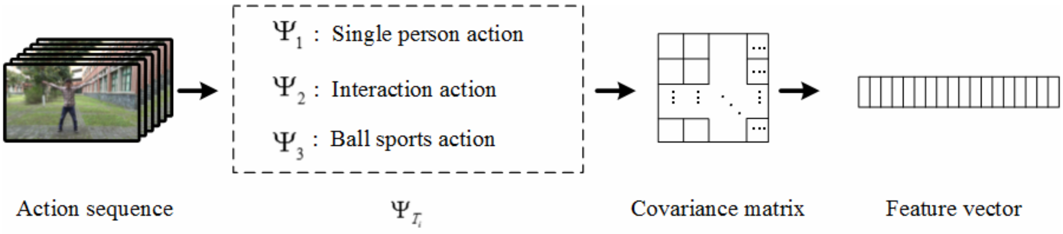

Ψ_1 : Single person action
Ψ_2 : Interaction action
Ψ_3 : Ball sports action
Action sequence
Ψ_{T_i}
Covariance matrix
Feature vector

Fig. 9. Operator Ψ_{T_i} of different groups for action representation: testing sample selects operator Ψ_{T_i} according to its action group to calculate covariance matrix for action recognition

3.2 Recognition via Sparse Representation

Sparse representation has been widely applied to many computer vision tasks in recent years. The main idea underlying sparse representation is that the testing sample is expressed as a linear combination of the fewest possible training vectors. John et al. [27] developed an effective framework based on sparse representation for face recognition. After that, sparse representation was used in classification problems, such as face recognition [28], image super-resolution [29-30], and action recognition [31-32]. We build an action recognition system based on sparse representation because it has sufficient ability to discriminate classes.

In our system, each group of actions is expressed through operator Ψ and all the training samples are collected to form a training set $P_i = [\mathrm{p}_{i,1}, \mathrm{p}_{i,2}, \cdots, \mathrm{p}_{i,n_i}]$, where i indicates the i^{th} action, n_i is the total number of training samples in action *i*. We build a matrix $P = [P_1, P_2, \cdots, P_M] \in \Re^{K \times M}$ from training data, where K is the dimension of the feature vector, and M is the number of actions. A query sample can be expressed as:

$$\mathrm{p}_{query} = P\alpha \in \Re^{K} \tag{17}$$

where $\alpha \in \Re^{K}$ is the vector of coefficients. Since $N >> M$ ($N = \sum_{j=1}^{M} n_j$), the condition is underdetermined and the solution α is not unique. Ideally, the only non-zero coefficients in α are those which correspond to the category of the testing sample. Solving the optimization problem is to seek a sparse solution:

$$\alpha^{*} = \arg\min \|\alpha\|_0, \text{ s.t. } \mathrm{p}_{query} = P\alpha \tag{18}$$

where $\|\cdot\|_0$ denotes the L0-norm which counts the number of non-zero entries in coefficients α. However, it is an NP-hard optimization problem when seeking a sparse solution in an underdetermined system. If the optimal solution α^{*} is sufficiently sparse, solving the L0-norm minimization problem is equivalent to solving the L1-norm minimization problem ([34]):

$$\alpha^{*} = \arg\min \|\alpha\|_1, \text{ s.t. } \mathrm{p}_{query} = P\alpha \tag{19}$$

Now the problem is a convex optimization problem, and it can be solved in polynomial time. In practice, the action video samples may not express the testing sample exactly due to noises and it can be changed to the following problem:

$$\alpha^{*} = \arg\min \|\alpha\|_1, \text{ s.t. } \|P\alpha - \mathrm{p}_{query}\|_2 \le \varepsilon \tag{20}$$

Since the noise and modeling error produce non-zero coefficients spreading across more than one action category, we utilize the reconstruction residual error (RRE) measure ([28]) to decide the label. Let $\alpha^{*} := [\alpha^{*}_{i,1} \ \alpha^{*}_{i,2} \ \cdots \ \alpha^{*}_{i,n_i}]^{T}$ indicate the coefficients corresponding to training samples from the category i. The RRE measure of the category P_i is defined as follows:

$$R_i(\mathrm{p}_{query,} \, \alpha^{*}) = \| \ \mathrm{p}_{query} - P_i\alpha^{*}_i \ \|_2 \tag{21}$$

Therefore, the category label of the testing sample p_{query} is as follows:

$$\text{label}(\mathrm{p}_{query,}) = \arg\min_{i} R_i(\mathrm{p}_{query}, \alpha^*) \qquad \textbf{(22)}$$

K-SVD algorithm is used to generate the dictionary in the learning process.

Here we summarize the action type recognition based on the sparse representation in Fig. 10. In the training stage, we select the operator Ψ_{T_i} to extract features for training data based on its action group and then use the covariance matrix of features to form the action descriptor. K-SVD algorithm is adopted to learn an overcomplete dictionary. In the testing stage, given a query sample $p_{query} \in \Re^K$, we select operator Ψ_{T_i} based on the action group to extract features and get α^* by solving L1-minimization problem. Then, RRE is computed for each category and the label is obtained.

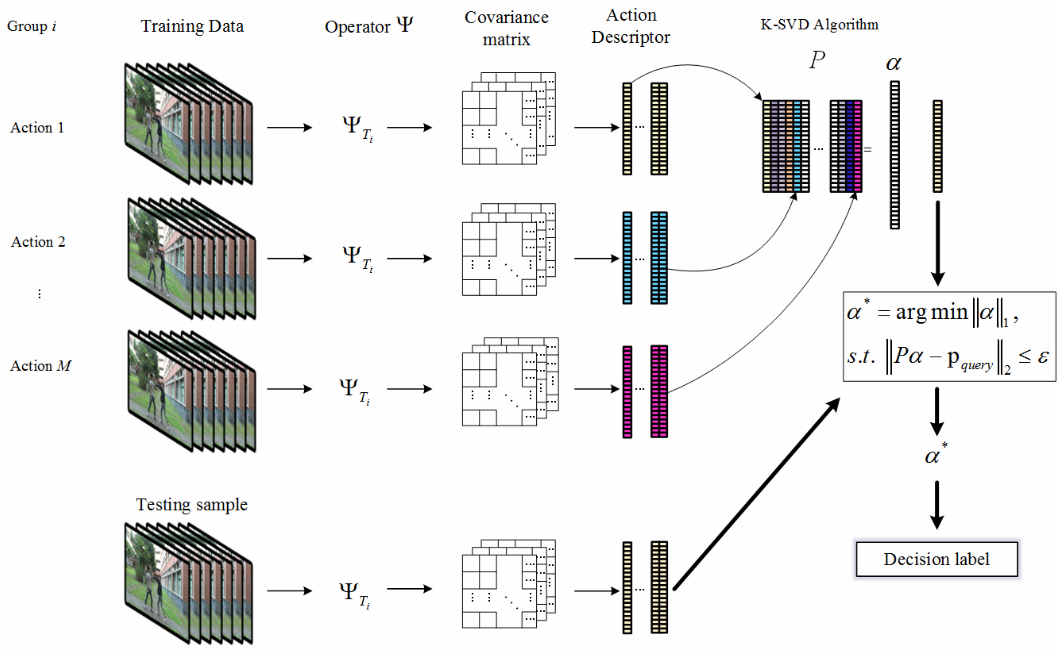


Fig. 10. The flowchart of action type recognition within a group *i*: we select the operator Ψ to extract features according to its action group, and perform dictionary training

4 Experimental Results

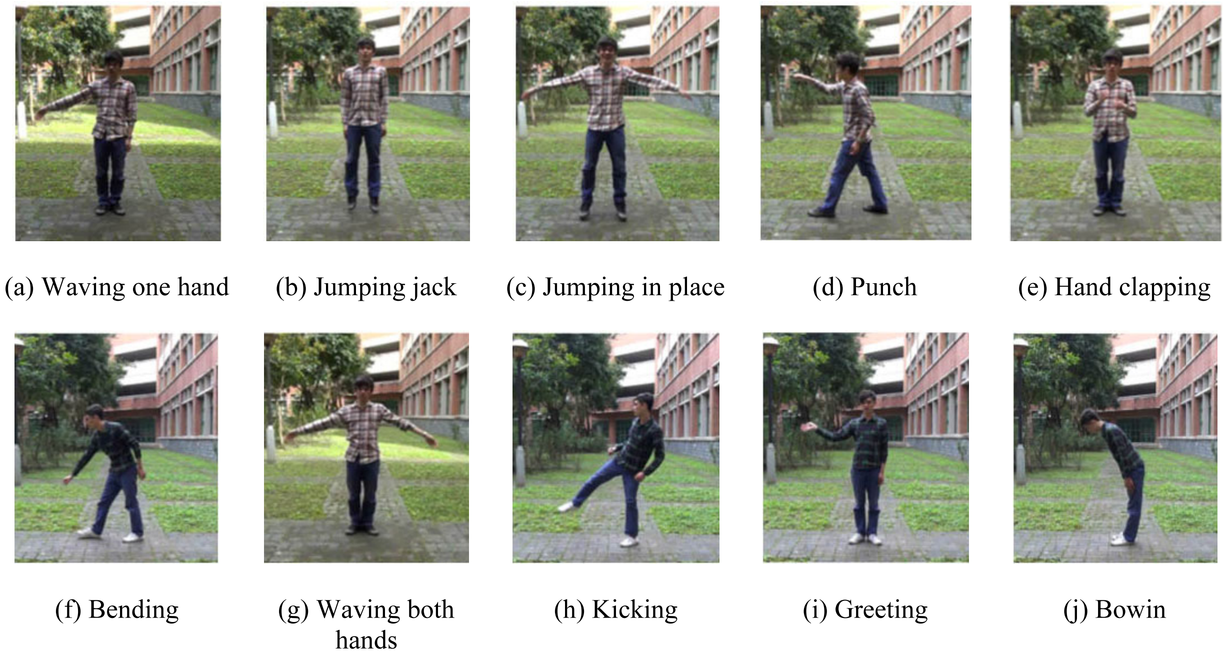
The proposed system is composed of two levels, action group classification and action type recognition. We are going to demonstrate the effectiveness of the proposed recognition system and evaluate the performance of the motion feature.

4.1 Database of Multiple Group

We conduct a series of experiments on multi-group action database which is built by ourselves. We use the SONY HDR-PJ430V camera to record the human action with the specification, the frame rate of 30 frames per second (fps) and resolution of 1280*720. The environment is outdoor scenes and the database contains 30 different actions that includes the daily life common actions, the interaction between two persons and the ball sport actions. The scenes contain natural backgrounds with different illumination conditions.

According to the property of the action, we partition the action database into three groups: single person action, two-person interaction and ball sport action, respectively. Each group has 10 different actions, and each action is performed 9 times. Therefore, it includes 90 actions in one group and three groups have a total of 270 action videos. The duration of each action is approximately 1-3 seconds.

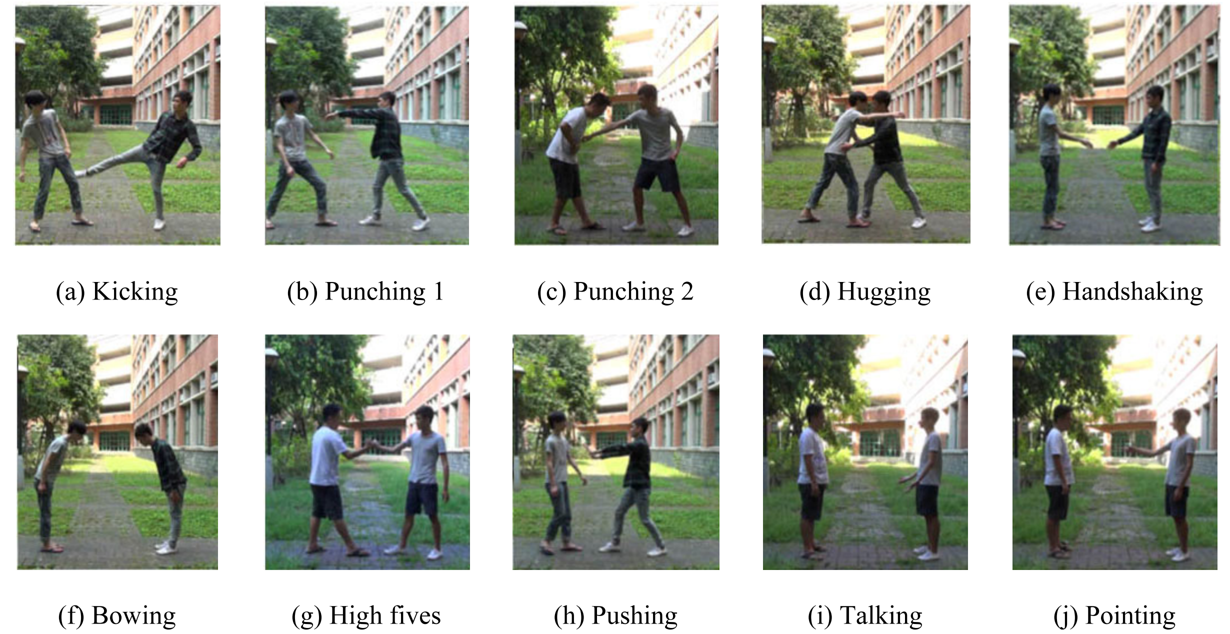
To begin with, the single action group included daily life common actions. There is only one person on the scene and the action is relatively simple. Compared to Weizmann dataset and KTH dataset, the proposed single person database contains 10 action types and the videos have complex background with larger illumination variation. Our single person action group is shown in Fig. 11.


(a) Waving one hand (b) Jumping jack (c) Jumping in place (d) Punch (e) Hand clapping

(f) Bending (g) Waving both hands (h) Kicking (i) Greeting (j) Bowin

Fig. 11. Single person action group

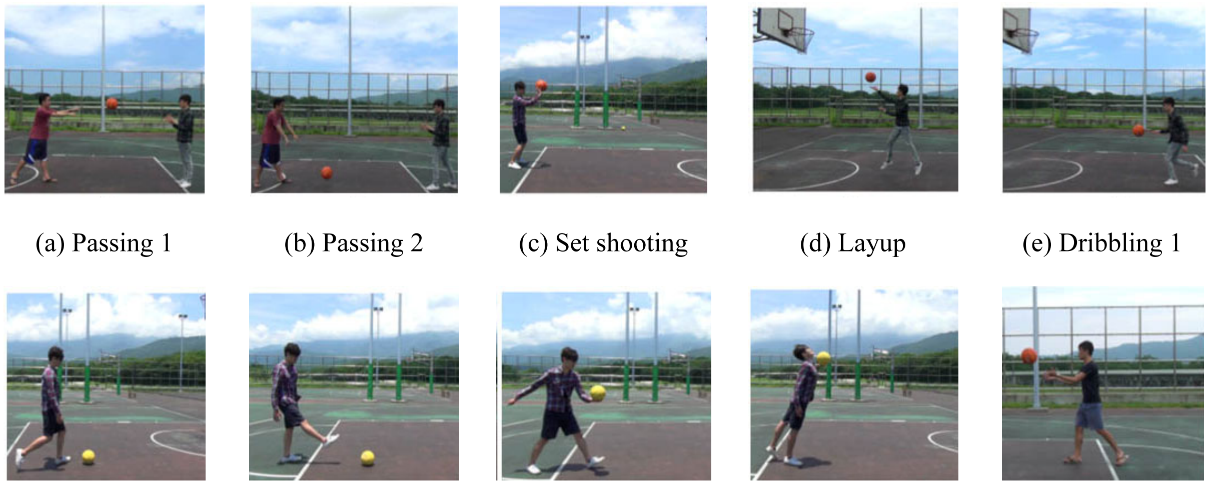
In the interaction group, all actions involve two people in the videos. We refer to the actions of [13] which contains 8 interactions in the indoor scene and then expand our database to 10 two-person interactions. Some actions are different interactions, but they have similar body motions, such as “high fives” and “handshaking”, being hands outstretched. These actions increase the recognition difficulty and accordingly are more challenging. The interaction group is shown in Fig. 12.


(a) Kicking (b) Punching 1 (c) Punching 2 (d) Hugging (e) Handshaking

(f) Bowing (g) High fives (h) Pushing (i) Talking (j) Pointing

Fig. 12. Interaction group

There are three typical actions in the ball sport action and Fig. 13 shows the 10 ball sport action.

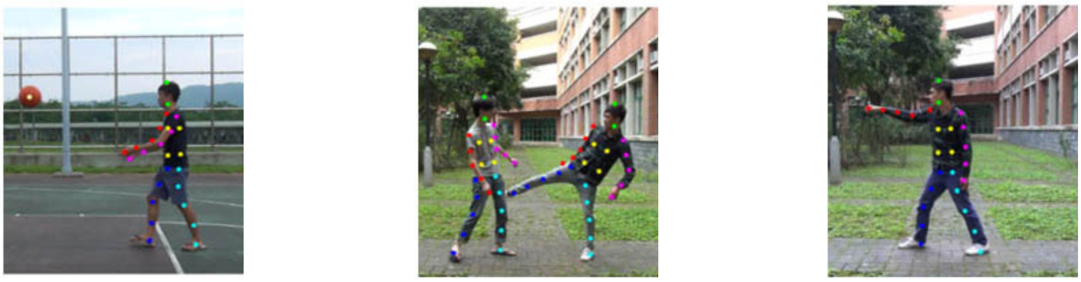


(a) Passing 1 (b) Passing 2 (c) Set shooting (d) Layup (e) Dribbling 1

(f) Dribbling 2 (g) Kicking (h) Serve (i) Chesting (j) Catching

Fig. 13. Ball sport action group

We utilize a multi-group action database as the training set for sparse representation. We manually label the interesting points for each body part in a video sequence, and the training features are computed from these labelled points. We use different colors to label body parts and the object for each action group: head for green, torso for yellow, left hand for red, right hand for magenta, left leg for blue, right leg for cyan, and beige for ball. The body parts of the single person action group are labeled as 26 parts. And we label a total of 52 body parts for two persons in the two person interaction group. Then, in the ball sport action group, we mark a total of 27 positions for the moving person and the ball. Finally, we extract the basis vectors according to its action group for dictionary learning. The labeled examples of each group are shown in Fig. 14.



(a) Ball sports action (b) Two people's interaction (c) Single person action

Fig. 14. The label examples of an action frame: different examples of manually labeled for human body parts and ball position

4.2 Experimental Setting

The recognition system performance evaluation is based on leave-one-out cross validation (LOOCV). In all experiments, we select one as a testing sample and use the remaining samples as the training set. Then, we identify the action of the testing sample. We repeat the procedure for all samples in the action database and calculate the correct classification rate (CCR).

The PAESE images [38] and Buffy dataset [39] are used for PSM model training. The PAESE images are mainly composed of pictures of humans and the scene resembles a realistic environment. Each image has its own illumination condition, and the poses of the humans are natural. Buffy database is a TV show of which the illumination of images is darker than that of PAESE images. Since the clothing color and body scale are different among the images, it is a more challenging database. We apply an image labeled approach of dictionary learning to train the PSM. After obtaining the trained model, we can detect body parts for testing samples.

4.3 Features Selection

In order to obtain the best recognition results, we have to find out the most suitable features for each action group. We evaluate the performance of different feature combinations by observing their average recognition rates. The basic features for the feature combinations consist of position, distance, velocity, and area.

Table 1 is the recognition results which uses different feature combinations for three groups. Group 1 to Group 3 are single person action, two-person interaction and ball sport action, respectively. The overall results are summarized as follows.

Table 1. Multi-group action recognition rates using different feature combinations

Feature	Group 1	Group 2	Group 3
Position	84.31%	82.62%	81.20%
Distance	54.70%	76.10%	73.64%
Velocity	41.87%	39.74%	45.29%
Area	31.15%	39.66%	41.15%
Position + Distance	84.32%	86.27%	83.54%
Position + Velocity	80.06%	76.63%	79.20%
Position + Area	76.32%	83.42%	87.24%
Distance + Velocity	53.42%	60.91%	64.26%
Distance + Area	54.92%	63.74%	58.64%
Velocity + Area	31.63%	45.69%	44.82%
Position + Distance + Velocity	81.48%	79.61%	86.24%
Position + Distance + Area	84.21%	90.67%	85.97%
Distance + Velocity + Area	76.87%	81.20%	76.91%
Position + Velocity + Area	84.42%	79.64%	89.49%
Position +Distance	91%	90.33%	92.89%

According to the results of Group 1, we observe that we attain unsatisfactory results when we only use distance or area features. The reason is that the changes in single person actions are small. However, we obtain better recognition results when combined with all the features. Therefore, we choose the combination of all four features for Group 1. Table 1 shows of the recognition rate reach the highest for Group 2 when the feature combination contains position, distance, and area. The recognition rate reaches 92.89% for Group 3 when the feature combination contains all four features. It is higher than the other two groups.

4.4 Recognition Results

In the section, we demonstrate the performance of the two-level action recognition system for our multi-group action database.

Performance of action classification. We evaluate the ability of the feature of motion vectors while in action group classification. The confusion matrix is shown in Table 2. Because actions of the same group occupy the nearby region, the location information of flow vectors can be used to express the action groups. We divide the optical flow field into ten and five slices to calculate a magnitude histogram for each slice along the horizontal and vertical orientation, respectively. This idea indicates that motion vectors continuously appear when an object moves in optical flow field. In contrast, the background has no significant changes. The classification results indicate that motion vectors have excellent discrimination with the average classification accuracy rate of 99.59%.

Table 2. Confusion matrix of different action group

Action Group	Single Person Action	Interaction Action	Ball Sports Action
Single Person Action	99.41%	0.59%	0
Interaction Action	0.05%	99.95%	0
Ball Sports Action	0.20%	0.37%	99.43%

Multi-group action recognition results. In the experiment, we evaluate the proposed action recognition system. Nine samples of each action are selected as the testing videos, so there are 270 testing videos. These testing videos are converted to the feature vectors according to its corresponding action group since each group has its own specific moving characteristics. Moreover, different dictionaries are learned for different groups to recognize the action type of the testing video. Table 3 to Table 5 show the confusion matrices of single person action, two-person interaction and ball sports action, respectively.

Table 3. Confusion matrix of single person action group

Action	Action Type									
Type	a	b	c	d	e	f	g	h	i	j
a	87.78%	0.00%	0.00%	1.11%	0.00%	0.00%	1.11%	1.11%	8.89%	0.00%
b	0.00%	91.11%	0.00%	1.11%	1.11%	0.00%	5.56%	0.00%	0.00%	1.11%
c	0.00%	3.33%	94.44%	0.00%	0.00%	0.00%	0.00%	1.11%	0.00%	1.11%
d	0.00%	2.22%	0.00%	90.00%	1.11%	0.00%	1.11%	5.56%	0.00%	0.00%
e	0.00%	2.22%	0.00%	0.00%	90.00%	1.11%	0.00%	4.44%	0.00%	0.00%
f	0.00%	1.11%	0.00%	0.00%	0.00%	93.33%	0.00%	0.00%	2.22%	3.33%
g	0.00%	5.56%	0.00%	0.00%	3.33%	1.11%	88.89%	0.00%	1.11%	0.00%
h	0.00%	0.00%	0.00%	0.00%	2.22%	1.11%	2.22%	93.33%	1.11%	0.00%
i	0.00%	0.00%	0.00%	0.00%	1.11%	0.00%	0.00%	2.22%	88.89%	1.11%
j	0.00%	1.11%	2.22%	1.11%	2.22%	0.00%	0.00%	1.11%	0.00%	92.22%

Table 4. Confusion matrix of interaction action group

Action	Action Type									
Type	a	b	c	d	e	f	g	h	i	j
a	91.11%	2.22%	1.11%	0.00%	0.00%	0.00%	0.00%	3.33%	0.00%	2.22%
b	0.00%	87.78%	7.78%	0.00%	1.11%	1.11%	2.22%	0.00%	0.00%	0.00%
c	0.00%	8.89%	86.67%	0.00%	0.00%	1.11%	0.00%	1.11%	1.11%	1.11%
d	0.00%	0.00%	1.11%	88.89%	1.11%	3.33%	0.00%	2.22%	0.00%	3.33%
e	0.00%	0.00%	0.00%	1.11%	94.44%	0.00%	0.00%	3.33%	1.11%	0.00%
f	1.11%	1.11%	0.00%	2.22%	1.11%	93.33%	0.00%	1.11%	0.00%	0.00%
g	0.00%	0.00%	1.11%	2.22%	1.11%	0.00%	92.22%	2.22%	1.11%	0.00%
h	0.00%	7.78%	1.11%	1.11%	0.00%	1.11%	0.00%	88.89%	0.00%	0.00%
i	0.00%	0.00%	0.00%	0.00%	4.44%	0.00%	1.11%	0.00%	91.11%	3.33%
j	0.00%	1.11%	0.00%	0.00%	1.11%	0.00%	0.00%	1.11%	4.44%	92.22%

Table 5. Confusion matrix of ball sports action group

Action	Action Type									
Type	a	b	c	d	e	f	g	h	i	j
a	98.89%	0.00%	0.00%	0.00%	1.11%	0.00%	0.00%	0.00%	0.00%	0.00%
b	0.00%	93.33%	0.00%	0.00%	0.00%	5.56%	0.00%	1.11%	0.00%	0.00%
c	1.11%	0.00%	98.89%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
d	0.00%	1.11%	0.00%	91.11%	6.67%	0.00%	0.00%	0.00%	0.00%	1.11%
e	0.00%	1.11%	0.00%	0.00%	90.00%	0.00%	444.00%	222.00%	1.11%	1.11%
f	0.00%	6.67%	0.00%	0.00%	0.00%	92.22%	0.00%	0.00%	1.11%	0.00%
g	0.00%	0.00%	0.00%	1.11%	6.67%	0.00%	86.67%	0.00%	0.00%	0.00%
h	0.00%	0.00%	0.00%	2.22%	0.00%	0.00%	1.11%	96.67%	0.00%	0.00%
i	0.00%	1.11%	1.11%	0.00%	0.00%	7.78%	0.00%	1.11%	88.89%	0.00%
j	0.00%	0.00%	0.00%	4.44%	2.22%	0.00%	1.11%	0.00%	0.00%	92.23%

Table 6 shows the recognition results of actions for each group. The average recognition rates of each group are 91%, 90.67% and 92.89%, respectively. Here the average correct recognition rate is 91.52% for three groups.

Table 6. Average classification rates of each action

Action	Action Type									
Type	1	2	3	4	5	6	7	8	9	10
1	87.78%	91.11%	94.44%	90.00%	90.00%	93.33%	88.89%	93.33%	88.89%	92.22%
2	91.11%	87.78%	86.67%	88.89%	94.44%	93.33%	92.22%	88.89%	91.11%	92.22%
3	98.89%	93.33%	98.89%	91.11%	90.00%	92.22%	86.67%	96.67%	88.89%	92.23%

5 Conclusions

The paper presents an action recognition system based on a multi-level configuration, which can recognize 30 different actions of three groups in different backgrounds. The proposed system is consist of two levels: action group classification and action type recognition. Given a query action video, the corresponding action group is determined in the first level. In the second level, the suitable features generated from the PSM models within the action group are extracted and the action type is recognized through sparse representation. Our experimental results show that the average correct recognition rate of the proposed method can reach a good performance. In the future, we plan to expand the database to include more actions; moreover, it is also a critical work to explore more effective features.

Acknowledgements

This work was supported by Ministry of Science and Technology, Taiwan, ROC, under Grant No. MOST 106-2221-E-259-020-MY2 and MOST 107-2218-E-259-002.

References

[1] L. Gorelick, M. Blank, Actions as space-time shapes, IEEE Transactions on Pattern Analysis and Machine Intelligence 29(2007) 2247-2253.

[2] K. Guo, P. Ishwar, J. Konrad, Action recognition using sparse representation on covariance manifolds of optical flow, in: Proc. 2010 IEEE International Conference on Advanced Video and Signal Based Surveillance, 2010.

[3] I. Laptev, T. Lindeberg, Space-time interest points, in: Proc. 2003 IEEE International Conference on Computer Vision, 2003.

[4] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Proc. 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[5] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Proc. 2004 International Conference on Pattern Recognition, 2004.

[6] T. Guha, R.K. Ward, Learning sparse representations for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 34(2012) 1576-1588.

[7] C.-H. Hu, S.-L. Wo, An efficient method of human behavior recognition in smart environments, in: Proc. 2010 International Conference on Computer Application and System Modeling, 2010.

[8] B. Ni, G. Wang, P. Moulin, RGBD-HuDaAct: a color-depth video database for human daily activity recognition, in: Proc. 2011 IEEE International Conference on Computer Vision Workshops, 2011.

[9] R. Chaudhry, A. Ravichandran, G. Hager, R. Vidal, Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions, in: Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[10] V. Vo, N. Ly, An effective approach for human actions recognition based on optical flow and edge features, in: Proc. 2012 International Conference on Control, Automation and Information Sciences, 2012.

[11] S. Park, J.K. Aggarwal, Recognition of two-person interactions using a hierarchical Bayesian network, in: Proc. 2003 IWVS First ACM SIGMM international workshop on Video surveillance, 2003.

[12] S. Blunsden, E. Andrade, R. Fisher, Nonparametric classification of human interaction, in: Proc. 2007 Iberian Conference on Pattern Recognition and Image Analysis, 2007.

[13] K. Yun, J. Honorio, D. Chattopadhyay, T.L. Berg, D. Samaras, Two-person interaction detection using body-pose features and multiple instance learning, in: Proc. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012.

[14] S. Herath, M. Harandi, F. Porikli, Going deeper into action recognition: A survey, Image and Vision Computing 60(2017) 4-21.

[15] L.M. Fuentes, S.A. Velastin, Tracking-based event detection for CCTV systems, Pattern Analysis and Applications 7(2004) 356-364.

[16] M.S. Ryoo, J.K. Aggarwal, Spatio-temporal relationship match: video structure comparison for recognition of complex human activities, in: Proc. 2009 IEEE International Conference on Computer Vision, 2009.

[17] X. Cui, Q. Liu, M. Gao, D.N. Metaxas, Abnormal detection using interaction energy potentials, in: Proc. 2011 IEEE Conference on Computer Vision and Pattern Recognition, 2011.

[18] M.D. Rodriguez, J. Ahmed, M. Shah, Action MACH a spatiotemporal maximum average correlation height filter for action recognition, in: Proc. 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[19] J. Liu, Y. Yang, M. Shah, Learning semantic visual vocabularies using diffusion distance, in: Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[20] J.C. Niebles, C.-W. Chen, F.-F. Li, Modeling temporal structure of decomposable motion segments for activity classification, in: Proc. 2010 European Conference Computer Vision, 2010.

[21] J. Chen, G. Zhao, V.P. Kellokumpu, M. Pietikäinen, Combining sparse and dense descriptors with temporal semantic structures for robust human action recognition, in: Proc. 2011 IEEE International Conference on Computer Vision Workshops, 2011.

[22] B. Horn, B. Schunck, Determining optical flow, Artificial Intelligence 17(1981) 185-203.

[23] B. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: Proc. 1981 International Joint Conferences on Artificial Intelligence, 1981.

[24] R. Chaudhry, A. Ravichandran, G. Hager, R. Vidal, Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions, in: Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[25] C.-W. Hsu, C.-J. Lin, A comparison of methods for multi-class support vector machines, IEEE Transactions on Neural Networks 13(2)(2002) 415-425.

[26] N. Cristianini, J. Shawe-Taylor, An introduction to Support Vector Machines, Cambridge University Press, Cambridge, 2000.

[27] J. Wright, A. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 31(2009) 210-227.

[28] E. G. Ortiz, A. Wright, M. Shah, Face recognition in movie trailers via mean sequence sparse representation-based classification, in: Proc. 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[29] J. Yang, J. Wright, T.S. Huang, Y. Ma, Image super-resolution via sparse representation, IEEE Transactions on Image

Processing 19(2010) 2861-2873.

[30] M.-C. Yang, C.-T. Chu, Y.-C. F. Wang, Learning sparse image representation with support vector regression for single-image super-resolution, in: Proc. 2010 17th IEEE International Conference on Image Processing (ICIP), 2010.

[31] T. Guha, R.K. Ward, Learning sparse representations for human action recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 34(2012) 1576-1588.

[32] K. Guo, P. Ishwar, J. Konrad, Action recognition in video by sparse representation on covariance manifolds of silhouette tunnels, in: Proc. 2010 ICPR Recognizing Patterns in Signals, Speech, Images and Videos, 2010.

[33] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part based models, IEEE Transactions on Pattern Analysis and Machine Intelligence 32(2010) 1627-1645.

[34] D.L. Donoho, For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution, Communications on Pure and Applied Mathematics 59(2006) 797-829.

[35] P.F. Felzenszwalb, D.P. Huttenlocher, Pictorial structures for object recognition, International Journal of Computer Vision 61(2005) 55-79.

[36] O. Tuzel, F. Porikli, P. Meer, Region covariance: a fast descriptor for detection and classification, in: Proc. 2006 European Conference on Computer Vision, 2006.

[37] O. Tuzel, F. Porikli, P. Meer, Pedestrian detection via classification on riemannian manifolds, IEEE Transactions on Pattern Analysis and Machine Intelligence 30(2008) 1713-1727.

[38] M. Everingham, L.V. Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL visual object classes challenge, Int J Comput Vis 88(2010) 303-338. doi:10.1007/s11263-009-0275-4

[39] M. Eichner, M. Marin-Jimenez, A. Zisserman, V. Ferrari, 2D Articulated human pose estimation and retrieval in (almost) unconstrained still images, International Journal of Computer Vision 99(2012) 190-214.

[40] E. Horowitz, S. Sahni, S. Anderson-Freed, Fundamentals of Data Structures in C, 2nd ed., Silicon Valley PR, San Jose, CA, 2002.

[41] M. Rohrbach, S. Amin, M. Andriluka, B. Schiele, A database for fine grained activity detection of cooking activities, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[42]S.-Z. Zhan, I.-C. Chang, Pictorial structures model-based human interaction recognition, in: Proc. 2014 International Conference on Machine Learning and Cybernetics, 2014.