Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 387464, 11 pages
http://dx.doi.org/10.1155/2015/387464

Research Article

Human Action Recognition Based on Fusion Features Extraction of Adaptive Background Subtraction and Optical Flow Model

Shaoping Zhu[1] and Limin Xia[2]

[1]*Department of Information Management, Hunan University of Finance and Economics, Changsha 410205, China*
[2]*School of Information Science and Engineering, Central South University, Changsha 410083, China*

Correspondence should be addressed to Limin Xia; xlm@mail.csu.edu.cn

Received 27 October 2014; Revised 15 March 2015; Accepted 2 April 2015

Academic Editor: Hassan Askari

Copyright © 2015 S. Zhu and L. Xia. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A novel method based on hybrid feature is proposed for human action recognition in video image sequences, which includes two stages of feature extraction and action recognition. Firstly, we use adaptive background subtraction algorithm to extract global silhouette feature and optical flow model to extract local optical flow feature. Then we combine global silhouette feature vector and local optical flow feature vector to form a hybrid feature vector. Secondly, in order to improve the recognition accuracy, we use an optimized Multiple Instance Learning algorithm to recognize human actions, in which an Iterative Querying Heuristic (IQH) optimization algorithm is used to train the Multiple Instance Learning model. We demonstrate that our hybrid feature-based action representation can effectively classify novel actions on two different data sets. Experiments show that our results are comparable to, and significantly better than, the results of two state-of-the-art approaches on these data sets, which meets the requirements of stable, reliable, high precision, and anti-interference ability and so forth.

1. Introduction

Human actions recognition based on computer vision is of great scientific and practical importance. It has numerous significant theoretic values and wide potential applications, such as real-time video surveillance system, interpretation of sport events, and human computer interactions [1–3]. In recent years, human actions recognition has become a research hot spot. Tremendous amount of researches has been carried out in the field of human actions recognition from video sequence. There are many human actions recognition methods already presented. Common method to recognize human actions aims at both still images and video sequences [4, 5]. In paper [6], Madabhushi and Aggarwal proposed a Bayesian approach for human activity recognition. Park and Aggarwal [7] used a hierarchical Bayesian network (BN) to recognize two-person interactions. Duong et al. [8] proposed activity recognition and abnormality detection using the switching hidden semi-Markov model. Bayesian approach and Hidden Markov Models (HMM) have been extensively used to detect simple and complex events that occur in the scenarios.

Human action evolves dynamically. Due to a large number of parameters which need to be set, learning those models is also hard. Thus these approaches may limit their practical use. Recently, one popular approach for human action recognition is applied, in which a representation known as "temporal templates" captures both motion and shape to recognize human action. Bobick and Davis [9] applied temporal templates to recognize human movement. In papers [2, 10], Niebles et al. and Savarese et al. presented an unsupervised learning method for human action categories using spatial temporal words. Iosifidis et al. [11] used Linear Discriminant Analysis (LDA) to achieve human body postures in the dyneme space and produce a time invariant multiview movement representation, which obtained high human movement classification accuracy. In paper [12], Qian et al. aggregated multiple SVMs to accomplish the recognition of actions and achieved the best classification performance. Schüldt et al. [13] performed human action recognition by training SVM classifiers. However, these approaches ignore the contextual information provided by different frames in a video; the modeling and learning frameworks are rather

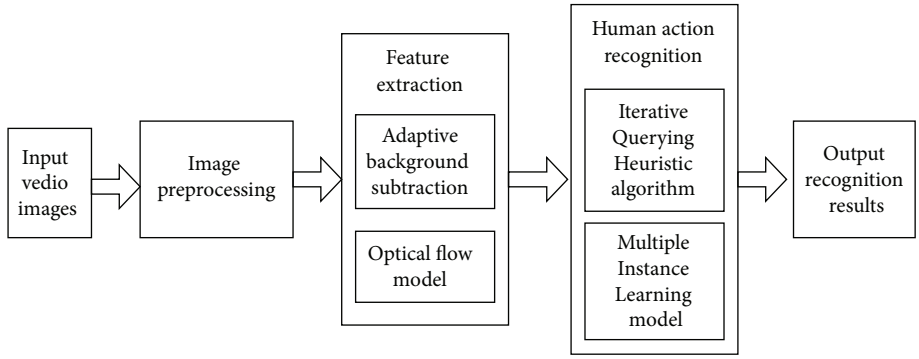


Figure 1: Flowchart of human action recognition based on fusion features extraction.

simple. Danafar and Gheissari [14] firstly constructed motion descriptors with optic flow and then classified human action with SVM. In papers [15, 16], the proposed approaches used two latent topic models, PLSA and LDA, to learn and recognize human action categories, which can save a lot of computation cost. Shotton et al. [17] proposed a new method to quickly and accurately predict human pose by designing an intermediate representation in terms of body parts, which obtained high accuracy on both synthetic and real test sets. Xia et al. [18] presented a novel approach for human action recognition with histograms of 3D joint locations (HOJ3D), which was real-time and achieved superior results on the challenging 3D action dataset. Jalal et al. [19] employed the radon transformation on depth silhouettes for human home action recognition. The result was interesting but at the same time not as good as one would expect. Ni et al. [20] combined color and depth information for human action recognition using spatiotemporal interest points (STIPs) and motion history images (MHIs). These results were interesting but at the same time not as good as one would expect.

Even great efforts have been made for decades, but these approaches have been fraught with difficulty because they are often inconsistent with other evidence of human actions [21]. The human actions recognition is still an immature technology and a challenging problem in computer vision to achieve robust human actions recognition from image sequences because of occlusion, human postures, illumination conditions, changing backgrounds, camera movements. With moving target and cameras, nonstationary background, few vision algorithms can categorize and recognize human actions well. It is essential for intelligent and natural human computer interaction to recognize human actions automatically.

In this paper, we present a novel method for human action recognition from video sequences. This approach includes two steps of feature extraction and human action recognition. In the extracting feature, global silhouette features of human are extracted by using adaptive background subtraction algorithm; local optical flow features are extracted by using optical flow model. Then fusing global silhouette feature vector and local optical flow feature vector form a hybrid feature vector. Finally an optimized Multiple Instance Learning algorithm is used for human action recognition. In addition, in order to improve the recognition accuracy, an Iterative Querying Heuristic (IQH) optimization algorithm is used to train the Multiple Instance Learning model.

Given unlabeled human actions video sequence, it is our goal to automatically learn different human actions categories in the data and apply the Multiple Instance Learning model for human actions categorization and recognition in the new video sequences. Our approach is illustrated in Figure 1.

The rest of this paper is organized as follows. In Section 2, we describe human action features representation based on adaptive background subtraction and optical flow models. In Section 3, we give details of algorithm for human action recognition using an optimized Multiple Instance Learning algorithm. Section 4 shows experiment result, also comparing our approach with two state-of-the-art methods, and the conclusions are given in the final section.

2. Human Actions Hybrid Features Representation

Recognizing human actions from video sequences is both a challenging problem and an interesting research area. Generally, two important questions are involved in human action recognition. One is how to represent efficiently human actions, which is the key for a good human actions recognition method, and the other is how to model reference movements, which efficiently deal with variations at spatial and temporal scales within similar motion categories.

Extracting useful motion information from raw video data is crucial for human actions recognition. The choice of the motion features affects the result of the human action recognition method directly. Many factors often influence the single feature differently, such as appearance of human body, environment, and video camera. So the accuracy of action recognition is limited. After fully considering the advantages and disadvantages of different features, we propose a hybrid feature method on the basis of studying the representation and recognition of human actions, which is fusing global silhouette feature and local optical flow feature.

2.1. Global Silhouette Feature Representation. Human silhouettes in single frame image can be used to describe the information of which overall shape of a human body movement changes [22, 23].

Adaptive background subtraction [24] is a kind of very good method for global silhouette feature extraction. We use it to determine the general area of the movement and the human body shadow.

Assume that all the actions are performed before the static background. We use adaptive background subtraction to determine the motion area and extract the human body silhouette. The step for adaptive background subtraction algorithm is as follows.

Step 1 (initialize the background model). First Taking T background image continuously, then through the image, a single gaussian distribution is created to set the initial background statistical model $H(u_i, v_i^2)$. Forming the Gaussian distribution needs the mean and standard deviation, which can be calculated using

$$u_i = \frac{1}{T}\sum_{t=1}^{T} u_{it},$$
$$v_i^2 = \frac{1}{T}\sum_{t=1}^{T} (u_{it} - u_i)^2, \tag{1}$$

where u_{ij} is color value of the point i in the tth image.

Step 2 (extracting foreground area). Assuming that color value of the point i is r_i in current image, the image binarization is expressed as

$$B_i = \begin{cases} 1, & \text{if } (r_i - u_i > 3\sigma_i) \\ 0, & \text{else,} \end{cases} \tag{2}$$

where points of all signs "1" are foreground area and points of all signs "0" are background region.

Step 3 (update background model). Assume that $u_i(t)$ is color expectations of the point i at time t, $v_i^2(t)$ is color variance of the point i at time t, and $r_i(t)$ is color value of the point i in time t which collect images. At time $t + 1$, there is

$$u_i(t+1) = \begin{cases} \alpha u_i(t) + (1-\alpha) r_i(t), & (\text{flag}_i = 0) \\ u_i(t), & (\text{flag}_i = 1), \end{cases}$$
$$v_i^2(t+1) = \begin{cases} \alpha v_i^2(t) + (1-\alpha)(r_i(t) - u_i(t))^2 & (\text{flag}_i = 0) \\ v_i^2(t) & (\text{flag}_i = 1), \end{cases} \tag{3}$$

where α is used to control the background update rate, $0 < \alpha < 1$. Each point in the binary image is marked by flag_i, if value of the flag_i is "1," which is foreground, and flag_i is "0," which is the background. Thus, the area marked as "0" constitutes the background region and the area marked as "1" constitutes the foreground area.

So we can get the foreground of the binary image. Figure 2 shows an example of adaptive background subtraction to extract the moving human body silhouette.

Silhouette features have the following advantages.

Silhouette features can be used to describe the shape of human movement information simply and visually [25].

Silhouette features are easy to be extracted.

Binary silhouette figure is not sensitive to texture and color of the foreground image.

Suppose that a video sequence V contains T frame image I—namely, $V = [I_1, I_2, \ldots, I_T]$—and S is the sequence of motion silhouette corresponding to video V—namely, $S = [s_1, s_2, \ldots, s_T]$. We use the contour vector to describe the overall human silhouette and shape information. The step is as follows.

Step 1. Use Canny operator for edge profile of each frame silhouette and calculate the coordinates of the edge profile. Such human body contour can be set with n points, namely, $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.

Step 2. Calculate the center of mass about human body contour, which is expressed as

$$(x_c, y_c) = \left(\frac{1}{n_t}\sum_{i=1}^{n_t} x_i, \frac{1}{n_t}\sum_{i=1}^{n_t} y_i\right), \tag{4}$$

where (x_c, y_c) is the center of mass and (x_i, y_i) is edge points of the contour. n_t is the number of edge points in the tth image.

Step 3. Calculate the distance from the center to the edge points, which is expressed as

$$d_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}, \quad i = 1, 2, \ldots, n_t, \tag{5}$$

where d_i is the edge distance from the center of mass to edge points, which the ith edge points correspond to.

Step 4 (normalize contour vector). We use 2-norm for contour vector normalization processing [26]. 2-norm is the most popular general approaches for computing approximations of the condition number. It is the incremental condition estimation (ICE) of a triangular matrix and can be viewed as a special case of the framework. The strategy is naturally connected to adaptive techniques and contains clearly visible links to matrix decompositions. After the normalization processing of contour vector, we perform interval such as resampling. We obtain the fixed point N, where N is set as 180.

2.2. Local Optical Flow Feature Representation. Silhouette image extraction of incorrect information may result in that contour vector characteristics cannot accurately express movement characteristic. Optical flow characteristics are effective and accurate motion information representation in video sequences [27].

Optical flow-based human actions representation has attracted much attention. The human action is a dynamic

Figure 2: The processing pipeline of the adaptive background subtraction: (a) the original images, (b) the background images, and (c) the body silhouettes.

event; it can be represented by the motion information of the human. The local optical flow features (optical flow vector) are estimated by optical flow model.

Firstly, we compute the optical flow vector of human actions, namely $\mu = (\mu_L, \mu_H)$ at each frame using optical flow equation, which is expressed as

$$I_L \mu_L + I_H \mu_H + I_t = 0, \tag{6}$$

where $I_L = \partial I/\partial x$, $I_H = \partial I/\partial y$, $I_t = \partial I/\partial t$, $\mu_L = dx/dt$, and $\mu_H = dy/dt$, where (x, y, t) is the image in pixel (x, y) at time t, where $I(x, y, t)$ is the intensity in pixel (x, y) at time t, and μ_L, μ_H are the horizontal and vertical velocities in pixel (x, y).

We can obtain $\mu = (\mu_L, \mu_H)$ by minimizing the objective function:

$$L = \int_D \left[\lambda^2 \left\|\nabla\mu\right\|^2 + \left(\nabla I \cdot \mu + I_t\right)^2\right] dx\, dy. \tag{7}$$

We use the iterative algorithm [28] to compute the optical flow; optical flow is decomposed into longitudinal and transverse two components, namely, the longitudinal optical flow and transverse optical flow, which is expressed as

$$\begin{aligned} \mu_L^{k+1} &= \overline{\mu}_L^k - \frac{I_L\left[I_L\overline{\mu}_L^k + I_H\overline{\mu}_H^k + I_t\right]}{\lambda + I_L^2 + I_H^2}, \\ \mu_H^{k+1} &= \overline{\mu}_H^k - \frac{I_H\left[I_L\overline{\mu}_L^k + I_H\overline{\mu}_H^k + I_t\right]}{\lambda + I_L^2 + I_H^2}, \end{aligned} \tag{8}$$

where k is the number of iterations, $\overline{\mu}_L^k$, $\overline{\mu}_H^k$ are the average velocity of the neighborhood of point (x, y), and $\mu_L^0 = \mu_H^0 = 0$ is initial value of velocity.

After standardization, optical flow diagrams are divided into 2×2 subframe, where they are set S_1, S_2, S_3, S_4, respectively. The size of subframe is set 60×60, and the centers of subframe are set G_i as shown in Figure 3, where $i = 1, 2, 3, 4$. Then each subframe is divided into 18 subareas in the centers of subframe and degree of each central angle is 20° as shown in Figure 4, where they are set S_i^j, and where $j = 1, 2, \ldots, 18$.

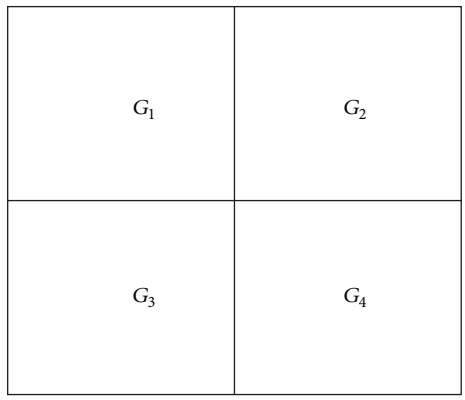


Figure 3: 2 × 2 subframe figure of optical flow diagrams.

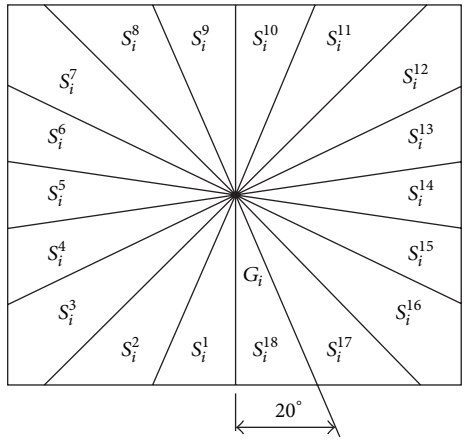


Figure 4: 18 subareas of each subframe in the centers of subframe.

Thus each frame image of optical flow graph is divided into 72 subareas.

We calculate the sum of the longitudinal optical flow $O_{L_{i,j}}$ and the sum of the transverse optical flow $O_{H_{i,j}}$ at subarea:

$$O_{L_{i,j}} = \sum_{k=1}^{K} \mu_L^{k+1} \in S_i^j,$$
$$O_{H_{i,j}} = \sum_{k=1}^{K} \mu_H^{k+1} \in S_i^j, \tag{9}$$

where S_i^j is the subarea, $i = 1, 2, 3, 4$ and $j = 1, 2, \ldots, 18$, and k is the number of the longitudinal optical flows or the transverse optical flows.

Optical flow information of the image can be represented by the sum of longitudinal optical flow and transverse optical flow in 72 subareas, which is expressed as

$$O_L = \left[O_{L_{1,1}}, \ldots, O_{L_{1,18}}, \ldots, O_{L_{4,1}}, \ldots, O_{L_{4,18}}\right],$$
$$O_H = \left[O_{H_{1,1}}, \ldots, O_{H_{1,18}}, \ldots, O_{H_{4,1}}, \ldots, O_{H_{4,18}}\right], \tag{10}$$

where O_L is the sum of longitudinal optical flow in 72 subareas and O_H is the sum of transverse optical flow in 72 subareas.

We calculate local optical flow vector O_T:

$$O_T = \left[O_L, O_H\right], \tag{11}$$

where we use 2-norm for O_T normalization processing and get the local optical flow vector of the current frame image.

2.3. Human Action Hybrid Feature Representation. In order to improve the accuracy of human action recognition, fusing the contour vector and local optical flow vector form a hybrid feature vector, which is expressed as

$$F_T = \left[O_T, D_T\right], \tag{12}$$

where F_T is the hybrid feature vector of each frame image, O_T is local optical flow vector, and D_T is contour vector.

Human actions are represented by the hybrid feature vectors of contour vector and local optical flow vector. Because human actions can be regard as motion, the local optical flow features can describe human actions effectively. In addition, silhouette features can describe the shape of human movement information simply and visually. Thus, the hybrid features have been shown to perform reliably with noisy image sequences and have been applied in various tasks, such as action classification and action recognition.

3. Human Actions Recognition

After characterizing human actions, there are many methods to recognize human actions. Because human actions recognition is innately a Multiple Instance Learning problem, we use the Multiple Instance Learning algorithm to learn and recognize human actions. The Multiple Instance Learning model has been applied to various computer visions, such as object recognition, action recognition, and human detection [29, 30]. In order to increase recognition efficiency without compromising accuracy, an Iterative Querying Heuristic (IQH) optimization algorithm is used to train the Multiple Instance Learning model, and then the improved Multiple Instance Learning framework is used to learn a unified classifier instead of individual classifiers for all categories.

3.1. The Multiple Instance Learning Algorithm Analysis. The Multiple Instance Learning is one of the most efficient machine learning algorithms at present. Its idea was originally proposed by Keeler et al. [31] in 1990. It was called Integrated Segmentation and Recognition (ISR), and its key idea is to provide a different way in constituting training samples. Training samples are in "bags," they are not singletons, and all of the samples in a bag share a label. Each bag contains a large number of instances. Samples are organized into positive bags or negative bags of instances, where at least one instance is positive in a positive bag, but all instances are negative in a negative bag. In the Multiple Instance Learning, it must simultaneously learn that samples in the positive bags are positive along with the parameters of the classifier [32, 33]. In this paper, the Multiple Instance Learning is used for human actions with nonaligned training samples. The step for recognizing human actions based on the Multiple Instance Learning is as follows.

Step 1. Given dataset $\{X_i, C_i\}_{i=1}^{N}$, where $X_i = \{x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{iN}\}$ is on behalf of training bags, which contains at least one positive sample in a positive bag, $C_i = \max_j(C_{ij})$ represents the score of the sample, and $C_i \in \{0, 1\}$, where $C_i = \{C_{i1}, C_{i2}, \ldots, C_{ij}, \ldots, C_{iN}\}$. N is the number of all weak classifiers.

Step 2. Update all weak classifiers k with data $\{x_{ij}, C_i\}$.

Step 3. Initialize all strong classifiers: $H_{ij} = 0$ for all i, j, where strong classifiers are composed of all the weak classifier.

Step 4. For $k = 1$ to K, for $m = 1$ to N.

In the ith bag, when the jth sample is positive, the probability is calculated as follows:

$$P_{ij}^{m} = \sigma\left(H_{ij} + h_m\left(x_{ij}\right)\right), \tag{13}$$

where $P_{ij}^{m} = p(C_i \mid x_{ij}) = 1/(1 + \exp(-C_{ij}))$.

In the positive bag, the probability is calculated as follows:

$$P_{i}^{m} = 1 - \prod_{j}\left(1 - p_{ij}^{m}\right), \tag{14}$$

where $P_{i}^{m} = p(C_i \mid X_i)$.

The likelihood, which is assigned to a set of training bags, is expressed as

$$L^{m} = \sum_{i}\left(C_i \log\left(p_i^{m}\right) + \left(1 - C_i\right)\log\left(1 - p_i^{m}\right)\right). \tag{15}$$

End for.

Find the maximum m^* from N and obtain the current optimal weak classifier as follows:

$$m^{*} = \arg\max_{m}\left(L^{m}\right). \tag{16}$$

The strong classifier is produced by the m^* as follows:

$$\begin{aligned} h_k(x) &\longleftarrow h_{m^*}(x), \\ H_{ij} &= H_{ij} + h_k(x). \end{aligned} \tag{17}$$

End for.

Step 5. k weak classifiers constitute a strong classifier, which is expressed as

$$H(x) = \sum_{k} h_k(x), \tag{18}$$

where h_k is a weak classifier, which can make binary predictions by using $\operatorname{sign}(H_K(x))$.

In the Multiple Instance Learning, samples come into positive bags and negative bags of instances. Each instance x_{ij} is indexed by two indices, where i stands for the bag and j stands for the instance in the bag. All instances in a bag share a bag label C_i. The weight of each sample consists of the weight of bags and the weight of samples in the bag, where the number of the samples can be interpreted as a likelihood ratio. P_{ij}^{m} is the probability of positive instances in the bags, so the weight of samples is P_{ij}^{m}. We calculate $w_{ij} = \partial \log L^{m}/\partial y_{ij}$ and obtain the weight of the bags w_{ij} [34].

3.2. Optimizing the Multiple Instance Learning Algorithm for Human Actions Recognition. The Multiple Instance Learning algorithm provides a general paradigm for a more relaxed form of supervised learning. In the Multiple Instance Learning, the learner gets unordered sets of instances or bags instead of receiving example or label pairs, and labels are provided for each bag rather than for each instance, where a positive bag contains at least one positive instance. In the initial training stages, training and evaluating have a direct effect on both the features and the appropriate thresholds selected, and it is the key to a fast and effective classifier. The samples have high score in positive bags. The final classifier labels these samples to be positive. The rest of the samples have a low score in the positive bags, which are assigned a low weight. The final classifier classifies these samples as negative samples. Setting the detection threshold and training a complete classifier, we obtain the desired false positive rates and false negative rates. Retrain the initial weak classifier to get a zero false negative rate on the positive samples. Repeat the process to train the second classifier and yield a zero false negative rate on the remaining samples.

During the inference stage, given a testing image, we can treat each aspect in the Multiple Instance Learning model as one class of human actions. Human actions recognition needs large amount of training data, so it will result in long training time. In this paper, we adopt an Iterative Querying Heuristic (IQH) algorithm to train the Multiple Instance Learning model [35]. The main step is as follows.

Input. Training bags $\{x_1, x_2, \ldots, x_M\}$, labels $\{y_1, \ldots, y_m\}$, and parameters τ, w, and p, where τ is iterations times, p is instances per iteration, and w controls how many new instances, are considered in each iteration.

Step 1. Initialize $H_{ij} = 0$, h_k being any classifier in $H(x)$.

Step 2. For $\tau = 1, \ldots, \beta$.

Step 3. Query w new candidate instances per bag: $Z_i^{\tau} = I_i^{\tau-1} \cup \{p_1^i, \ldots, p_w^i\}$, where $p_i^j \in x_i, \forall i$.

Step 4. Keep p the highest scoring instance by using h_r, where h_r is the form of classifier and can label the instances of a positive bag x as varying with the latent parameter.

$I_i^{\tau} \subset Z_i^{\tau} s \cdot \tau \cdot |I_i^{\tau}| = \widehat{\rho}$ and $h_r^{\tau-1}(p) \geq h_r^{\tau-1}(p')$ for all $p \in I_i^{\tau}$, $p' \in Z_i^{\tau} \setminus I_i^{\tau}$.

Step 5. Train $\overline{h}_r^{\tau}$ with the selected instances: $\overline{h}_r^{\tau} \leftarrow \wedge(\{I_1^{\tau}, \ldots, p_m^{\tau}\}, \{y_1, \ldots, y_m\})$.

Step 6. End for.

Step 7. Return h_r^{β} and the corresponding $\overline{h}_r^{\beta}$.

In Iterative Querying Heuristic [36], for positive bags, we use the current estimate of the classifier to select the most positive instances, which ensure that at least one of the queried instances is positive. For negative bags, all instances are negative. In this case, we select the closest instances to

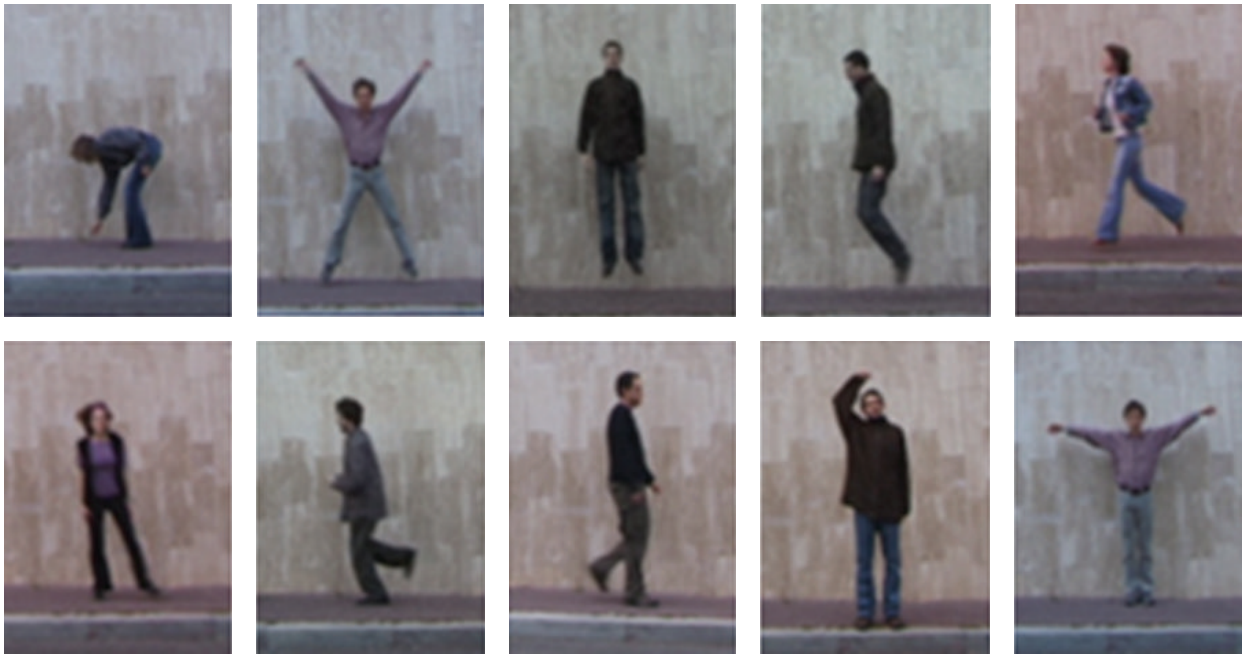


Figure 5: Key frames for Weizmann human action data set.

the decision boundary of our current classifier, which correspond to the most difficult negative instances. Then these selected instances are used to find a better classifier. Thus, Iterative Querying Heuristic is expected to take advantage of a large number of instances per bag and need not actually train with all of them at the same time. Each image has class label information in the training images, which is important for the classification task. This kind of class label information is used to learn the Multiple Instance Learning model. So each image directly corresponds to certain human actions on training sets. Iterative Querying Heuristic algorithm makes the training more efficient and improves the overall recognition accuracy significantly.

4. Experimental Results and Analysis

In this section we presented some experiments to validate our proposed approach. The effectiveness of the proposed algorithm was verified by using C++ and Matlab hybrid implementation on a PC with Pentium 3.2 GHz processor and 4G RAM. We tested our algorithm using Weizmann human action data set and the ballet data set. The reason of selecting these two types of videos is that they are challenging databases.

4.1. Experiment on Weizmann Data Set. Weizmann human action data set [37] is the largest available video sequence dataset of human actions, whose sample images are shown in Figure 5. Weizmann database contains 83 video sequences and nine groups of images, where each performs nine different actions.

They are "running," "walking," "jumping-jack" ("jack"), "jumping-forward-on-two-legs" ("jump"), "jumping-in-place-on-two-legs" ("pjump"), "galloping-sideways" ("sideways"), "waving-two-hands" ("wave2"), "waving-one-hand" ("wave1"), and "bending," respectively. We track and stabilize the figures using the adaptive background subtraction masks that come with this data set. After an automatic preprocessing step, the video sequences were tracked and stabilized, and all the figures appeared in the center of the field of view. In this experiment, we studied recognition results of nine kinds of human actions from Weizmann human motion dataset. We investigated the performance of the proposed approach method for human action recognition. On the Weizmann dataset, the results of human action recognition are shown in Table 1, where the correct recognition rate is defined as follows:

$$\text{The correct recognition rate} = \frac{\text{the number of the times of correct recognition}}{\text{total number of the samples}} \times 100\%. \tag{19}$$

In Table 1 we can see that our method can correctly recognize most of human actions. The recognition rate is as high as 100% for "bending," "jumping-jack," "jumping-in-place-on-two-legs," "galloping-sideways," "waving-two-hands," and "waving-one-hand". Our method achieves 98.5% average recognition rate. A few mistakes were confusions between "jumping-forward-on-two-legs" and "jumping-in-place-on-two-legs" because these two kinds of actions were similar to each other.

Table 1: The results of human action recognition on Weizmann database.

The action categories	Total number of samples	The number of the times of correct recognition	The correct recognition rate (%)
Bending	200	200	100.0
Running	200	190	95.0
Walking	200	199	99.5
Jack	196	196	100.0
Jump	195	180	92.3
pjump	192	192	100.0
Sideways	186	186	100.0
Wave2	192	192	100.0
Wave1	180	180	100.0
The average recognition rate			98.5

4.2. Experiment on Ballet Data Set. The ballet data set [38] contains several minutes of digitized instructional ballet from an instructional ballet DVD. These images were preprocessed by aligning and scaling so that each human figure could be tracked and stabilized for all images. Finally, we obtained 50 tracks. All the frames in these video sequences were manually labeled with one of eight action labels (e.g., "left-to-right hand opening," "right-to-left hand opening," "standing hand opening," "leg swinging," "jumping," "turning," "hopping," and "standing still."). Some sample frames are shown in Figure 6.

In order to examine the accuracy of our proposed approach, we researched recognition accuracy of eight kinds of human actions from ballet human motion data set by using our method, which is the most challenging of the database. Recognition results were presented in the confusion matrices as shown in Figure 7, where Action1 to Action8 indicate "left-to-right hand opening," "right-to-left hand opening," "standing hand opening," "leg swinging," "jumping," "turning," "hopping," and "standing still," respectively. Each cell in the confusion matrix is the average result of every human action, respectively.

As Figure 7 shows, our algorithm can correctly classify most actions. The only exception is "standing still." This is intuitively reasonable since it is difficult to reliably obtain optical flow features for "standing still" action.

In order to testify the effectiveness of our proposed approach, our method is compared with SVM and LDA on the same data set and the same experimental settings, which were two state-of-the-art approaches for human action recognition. We used 260 different human action images for this experiment. For each method, we used the single and hybrid feature, respectively, in the experiment. The average recognition accuracy was observed, which is displayed in Table 2.

We can see that our method improves the recognition accuracy with either a single or hybrid feature from the above experiments. When we did experiments with hybrid characteristics, our method achieves 98.5% average recognition rate, whereas "SVM" obtains a result of 91.2% and "LDA" gets a result of 93.7%. In the experiment of single feature, our method gets a result of 94.6%, whereas "SVM" and "LDA" only achieve 85.4% and 89.3% average recognition rate, respectively. The reason is that we improve the recognition accuracy in the two stages of human action feature extraction and human action recognition. In the stage of human action feature extraction, we use a hybrid feature vector that combines global silhouette feature and local optical flow feature, which are reliable with noisy image sequences and describe human action effectively. In the stage of human action recognition, we use the improved Multiple Instance Learning algorithm to classify human action. Our method performs significantly better.

Table 2: Comparison of recognition accuracy for three methods.

Recognition method	Recognition rate (%) using single feature	Recognition rate (%) using hybrid feature
SVM	85.4	91.2
LDA	89.3	93.7
Our method	94.6	98.5

5. Conclusion

In this paper, we present a new method for human action recognition in video sequences, which focuses on two key problems extracting more useful and discriminating feature and improving the accuracy of classifier. The main contribution can conclude as follows.

In feature extraction and representation, we extracted global silhouette feature using adaptive background subtraction method, and optical flow model was used for extracting motion features. Then fusing two kinds of feature formed a hybrid feature vector.

In action modeling and recognition, we proposed the improved Multiple Instance Learning algorithm for human action recognition using Iterative Querying Heuristic (IQH) algorithm, so that the recognition efficiency can be increased without compromising accuracy.

Experiments were performed on Weizmann human action data set and ballet data set. Experiments evaluated the proposed method. Experimental results revealed that our proposed method performed better than previous ones. Our algorithm can also recognize multiple actions in complex motion sequences.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by Research Foundation for Science & Technology Office of Hunan Province under Grant (no. 2014FJ3057), by Hunan Provincial Education Science and

Figure 6: Key frames for Weizmann human action data set.

	Action 1	Action 2	Action 3	Action 4	Action 5	Action 6	Action 7	Action 8
Action 1	0.99	0.01	0	0	0	0	0	0
Action 2	0.01	0.98	0.01	0	0	0	0	0
Action 3	0.04	0.03	0.89	0.03	0.01	0	0	0
Action 4	0	0	0.01	0.99	0	0	0	0
Action 5	0	0	0	0	1.00	0	0	0
Action 6	0	0	0	0	0	0.97	0	0.03
Action 7	0	0	0	0	0	0	1.00	0
Action 8	0	0	0.07	0.02	0.05	0.02	0.06	0.78

Figure 7: Confusion matrix for human action recognition on the ballet data set.

"Twelve Five" planning issues (no. XJK012CGD022), by the Teaching Reform Foundation of Hunan Province Ordinary College under Grant (no. 2012401544), by the Foundation for Key Constructive Discipline of Hunan Province, and by the Foundation for Key Laboratory of Information Technology and Information Security in Hunan Province.

References

[1] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification via pLSA," in *Computer Vision—ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7–13, 2006, Proceedings, Part IV*, vol. 3954 of *Lecture Notes in Computer Science*, pp. 517–530, Springer, Berlin, Germany, 2006.

[2] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[3] A. Bissacco, M. H. Yang, and S. Soatto, "Detecting humans via their pose," in *Advances in Neural Information Processing Systems*, vol. 19, pp. 169–176, MIT Press, Boston, Mass, USA, 2007.

[4] V. Delaitre, I. Laptev, and J. Sivic, "Recognizing human actions in still images: a study of bag-of-features and art-based representations," in *Proceedings of the 21st British Machine Vision Conference (BMVC '10)*, vol. 2, pp. 7–15, Aberystwyth, UK, August 2010.

[5] W. Yang, Y. Wang, and G. Mori, "Recognizing human actions from still images with latent poses," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 2030–2037, IEEE, June 2010.

[6] A. Madabhushi and J. K. Aggarwal, "A bayesian approach to human activity recognition," in *Proceedings of the 2nd IEEE Workshop on Visual Surveillance (VS '99)*, pp. 25–32, IEEE, 1999.

[7] S. Park and J. K. Aggarwal, "A hierarchical Bayesian network for event recognition of human actions and interactions," *Multimedia Systems*, vol. 10, no. 2, pp. 164–179, 2004.

[8] T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh, "Activity recognition and abnormality detection with the switching hidden semi-Markov model," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 838–845, June 2005.

[9] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[10] S. Savarese, A. DelPozo, J. C. Niebles, and L. Fei-Fei, "Spatial-temporal correlatons for unsupervised action classification," in *Proceedings of the IEEE Workshop on Motion and Video Computing (WMVC '08)*, pp. 1–8, Copper Mountain, Colo, USA, January 2008.

[11] A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Multi-view human movement recognition based on fuzzy distances and linear discriminant analysis," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 347–360, 2012.

[12] H. Qian, Y. Mao, W. Xiang, and Z. Wang, "Recognition of human activities using SVM multi-class classifier," *Pattern Recognition Letters*, vol. 31, no. 2, pp. 100–111, 2010.

[13] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR '04)*, vol. 3, pp. 32–36, IEEE, August 2004.

[14] S. Danafar and N. Gheissari, "Action recognition for surveillance applications using optic flow and SVM," in *Computer Vision—ACCV 2007*, vol. 4844, pp. 457–466, Springer, Berlin, Germany, 2007.

[15] Z. Lu, Y. Peng, and H. H. S. Ip, "Image categorization via robust pLSA," *Pattern Recognition Letters*, vol. 31, no. 1, pp. 36–43, 2010.

[16] M. B. David, Y. N. Andrew, and I. J. Michael, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[17] J. Shotton, T. Sharp, A. Kipman et al., "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[18] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '12)*, pp. 20–27, Providence, RI, USA, June 2012.

[19] A. Jalal, M. Z. Uddin, J. T. Kim, and T.-S. Kim, "Recognition of human home activities via depth silhouettes and $\mathfrak{R}$ transformation for smart homes," *Indoor and Built Environment*, vol. 21, no. 1, pp. 184–190, 2012.

[20] B. Ni, G. Wang, and P. Moulin, "RGBD-HuDaAct: a color-depth video database for human daily activity recognition," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV' 2011)*, pp. 1147–1153, Springer, November 2011.

[21] T. H. Thi, L. Cheng, J. Zhang, L. Wang, and S. Satoh, "Structured learning of local features for human action classification and localization," *Image and Vision Computing*, vol. 30, no. 1, pp. 1–14, 2012.

[22] L. Ren, G. Shakhnarovich, J. K. Hodgins, H. Pfister, and P. Viola, "Learning silhouette features for control of human motion," *ACM Transactions on Graphics*, vol. 24, no. 4, pp. 1303–1331, 2005.

[23] M. Yamada, K. Ebihara, and J. Ohya, "A new robust real-time method for extracting human silhouettes from color images," in *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 528–533, IEEE, Nara, Japan, April 1998.

[24] J. M. McHugh, J. Konrad, V. Saligrama, and P.-M. Jodoin, "Foreground-adaptive background subtraction," *IEEE Signal Processing Letters*, vol. 16, no. 5, pp. 390–393, 2009.

[25] A. Mokhber, C. Achard, and M. Milgram, "Recognition of human behavior by space-time silhouette characterization," *Pattern Recognition Letters*, vol. 29, no. 1, pp. 81–89, 2008.

[26] S. Drgas and A. Dabrowski, "Kernel alignment maximization for speaker recognition based on high-level features," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH' 11)*, pp. 489–492, August 2011.

[27] M. J. Black and A. D. Jepson, "Estimating optical flow in segmented images using variable-order parametric models with local deformations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 10, pp. 972–986, 1996.

[28] C.-I. Chang and A. Plaza, "A fast iterative algorithm for implementation of pixel purity index," *IEEE Geoscience and Remote Sensing Letters*, vol. 3, no. 1, pp. 63–67, 2006.

[29] J. F. Murray, G. F. Hughes, and K. Kreutz-Delgado, "Machine learning methods for predicting failures in hard drives: a multiple-instance application," *Journal of Machine Learning Research*, vol. 6, pp. 783–816, 2005.

[30] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 288–303, 2010.

[31] J. D. Keeler, D. E. Rumelhart, and W. K. Leow, "Integrated segmentation and recognition of hand-printed numerals," in *Proceedings of the Conference on Advances in Neural Information Processing Systems 3 (NIPS '90)*, pp. 557–563, Morgan Kaufmann, San Francisco, Calif, USA, 1990.

[32] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Proceedings of the 11th Annual Conference on Neural Information Processing Systems (NIPS '98)*, pp. 570–576, December 1998.

[33] R. C. Bunescu and R. J. Mooney, "Multiple instance learning for sparse positive bags," in *Proceedings of the 24th International Conference on Machine Learning (ICML '07)*, pp. 105–112, ACM, Corvallis, Ore, USA, June 2007.

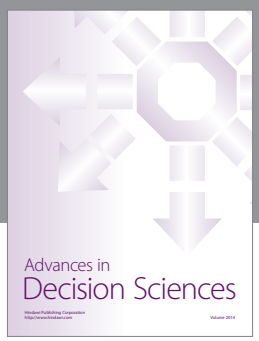
[34] S. Zhu, "Facial expression recognition based on MILBoost," *Journal of Software*, vol. 9, no. 9, pp. 2435–2442, 2014.

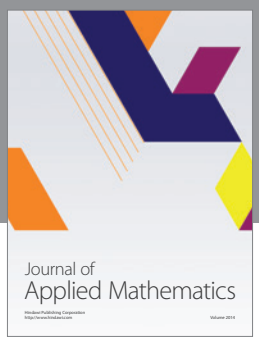
[35] S. P. Zhu, "Human action recognition based on improved MIL," *Applied Mechanics and Materials*, vol. 713–715, pp. 2152–2155, 2015.

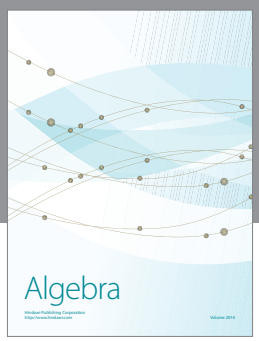
[36] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

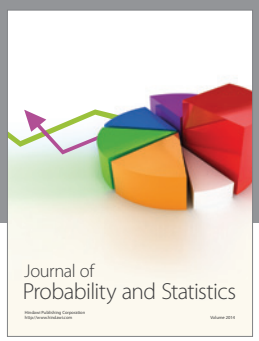
[37] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, vol. 2, pp. 1395–1402, Beijing, China, October 2005.

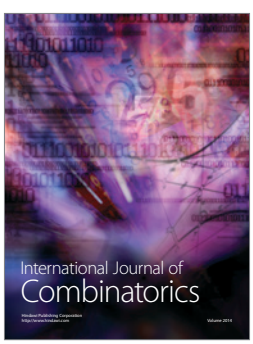
[38] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR' 08)*, pp. 1–8, Anchorage, Alaska, USA, June 2008.

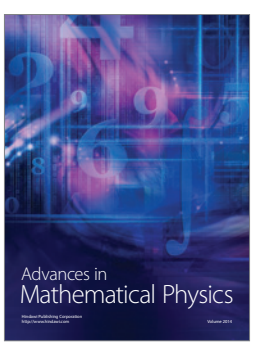


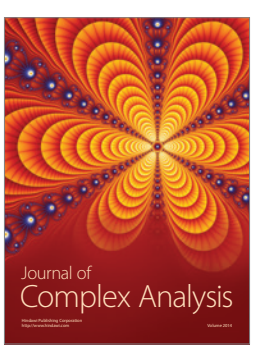


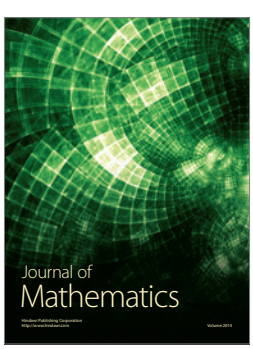


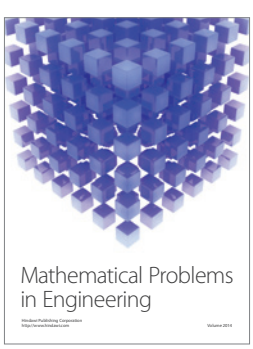


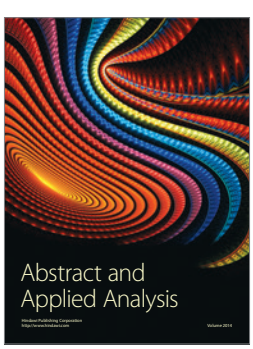


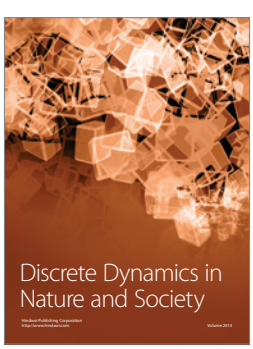


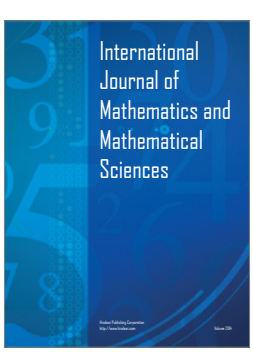












Hindawi

Submit your manuscripts at
http://www.hindawi.com