International Journal of Machine Learning and Cybernetics (2019) 10:2329–2340
https://doi.org/10.1007/s13042-018-0870-1

ORIGINAL ARTICLE

Human activity recognition using mixture of heterogeneous features and sequential minimal optimization

Humza Naveed[1] · Gulraiz Khan[1] · Asad Ullah Khan[1] · Aiman Siddiqi[1] · Muhammad Usman Ghani Khan[1]

Received: 23 August 2017 / Accepted: 27 August 2018 / Published online: 24 September 2018

Abstract
Automated detection and tracking of a person's actions plays a vital role in surveillance systems. Human activity detection has been carried out by using a variety of features; including flow-based, spatio-temporal and interest points based. We have created a fusion of features by incorporating those which give better results. LBP, HOG, Haar wavelets, SIFT, velocity and displacement being the major ones. By employing the time efficiency and optimality of SMO to train SVM, we have trained our system for both single person and multi-human action classification with improved accuracy. A generalized hierarchy of actions has been presented in this paper to demonstrate the extension of our methodology. We have achieved an accuracy of 91.99% on combination of KTH and Weizmann dataset and 86.48% on multi-human dataset. We have introduced our self-generated multi-human activity dataset in the following paper.

Keywords Human activity detection · HOG · SIFT · Velocity and displacement

1 Introduction

Human activity detection has numerous applications in the fields of security, surveillance, robotics and interactive systems. By keeping a track of suspicious activities performed by individuals and predicting acrimonious actions beforehand can dilute the effects of unpleasant events. Security personnel have been pretty good with this task of tracing suspected persons for decades. However, humans are prone to error and may result in false accusations. It is for this reason that automated surveillance systems are in the spotlight for aiding the surveillance process. Increased intensity of criminal activities all around the globe entails establishment of better automated systems to speed up the surveillance and output more accurate results. A well-developed system which detects, recognizes and follows the actions of an individual is needed. The system is expected not only to maintain record of past events but also to predict the upcoming events based on the irregularity in the normal trail of a person.

This paper focuses on presenting a comprehensive system to detect 13 single human actions and their combinations to depict multi-human scenarios, considering two people to cover multi-human case. Our technique identifies human activities irrespective of which primary body parts are involved in performing the action. The proposed methodology is extendable to work over a range of actions for any number of people present in the view.

Spatio-temporal, flow-based and keypoints dependent methodologies have been known to work well for human actions recognition, either as individual attributes or as combinations [1–5]. Most researchers have preferred using ready-made silhouettes for feature extraction to model shape-based changes in the human poses over a series of frames and also minimize the computational complexity [6, 7]. Among the flow and trajectory based features, optical flow has been found useful to maintain the track of interest points over a sequence of frames [8–10]. For classification, wide usage of support vector machines (SVM) [11, 12], hidden markov models (HMM) [13–16] and artificial neural networks (ANNs) [17] have been observed.

By making use of the positive aspects of features, being used commonly for human activity detection, and adding a few more, we have created an amalgam of features which

✉ Humza Naveed
humza.naveed@kics.edu.pk

✉ Gulraiz Khan
gulraiz.khan@kics.edu.pk

[1] Computer Vision and Machine Learning Lab, Al-Khawarizmi Institute of Computer Science, University of Engineering and Technology Lahore, Lahore, Pakistan

results in a comprehensive set of features to detect a wide range of human activities. Velocity and displacement, scale-invariant feature transform (SIFT) [1], histogram of oriented gradients (HOG) [18], Haar wavelets and radial histogram have been observed to provide a significant boost in the accuracy of detection. We have dealt with thin, high-dimensional data by training SVM via sequential minimal optimization (SMO) [19]. To account for realistic scenarios, we have not restricted our system to mere single human actions detection, but have also incorporated the recognition of multiple people performing activities independently. We have presented a generalization of actions, Fig. 1, to account for those which have not been elaborated in this paper but can be classified using our proposed methodology. The training and evaluation of single human actions has been performed on KTH and Weizmann datasets, however, multi human activity has been evaluated over self-generated dataset. The dataset covers 13 human actions performed by two persons independently in each video sequence. We discussed the dataset in Sect. 4 (see Fig. 8).

The paper gives a detailed survey of previous work in Sect. 2. Section 3 contains a discussion on methodology including preprocessing performed on the data, a discussion on extracted features, and classification techniques employed in our system. Sections 4 and 5 consist of experimentation and results respectively.

2 Literature survey

Human actions recognition techniques can be broadly classified into four categories. Flow-based methods, spatio-temporal templates, tracking and centered around interest point [1, 20]. Spatio-temporal templates make use of unique poses of the body to encode the action information. Optical flow and SIFT or speeded up robust features (SURF) features are the examples of flow-based method and interest point based techniques respectively. Moussa et al. [1] in his work, makes use of fine-tuned SIFT interest points, K-means clustering and later builds a normalized codebook using cluster indices to be passed onto the next stages of classification with SVM. He used min–max normalization technique to normalize the codebook of visual words. M. Moussa has made a successful effort in classifying six actions of the KTH dataset and ten belonging to Weizmann. Human pose presents a lot of information about the action being performed when it is used in addition with other poses tracked before or after the observation. Tharau and Hlavac [21] have used such approach in their attempt for activity recognition. Their methodology works well for both static images and a bunch of video frames since it does not involve dynamic feature computations. Poses are assigned weights according to the credibility of their occurrences. The testing phase is based on these assigned weights, classifying the action category using the information of training data weights.

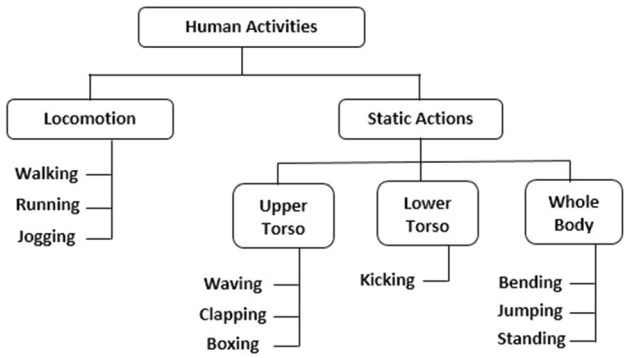

Fig. 1 Broad classification of Action Classes

Ming-Yu and Alexander [22] used spatio-temporal features along with the interest points to model local motion. They applied optical flow on SIFT features and retained only those features that showed significant optical flow. Using Bi-Gram bag of words approach for clustering, they achieved an accuracy of 95.83% on KTH dataset. Manosha et al. [23] combined optical flow and silhouette based shape features to train their system on Weizmann and UIUC datasets. They used one-vs-one multiclass SVM to achieve the accuracy of 97.45% on UIUC dataset and perfect 100% on 10 actions of Weizmann [24]. Umakanthan et al. [18] found the problems in standard multi-class SVM based approaches for complex datasets and proposed binary tree SVM based activity recognition using Gaussian Mixture Models. They used HOG and motion boundary histogram at the feature extraction stage to obtain 58.2% accuracy on challenging Hollywood dataset. Recently, Liu [25] et al. used a Bayesian network based generative framework to model the complex activities upon primitive actions recorded from sensors. Their probabilistic network modeled the uncertainties arising from missing or incorrect data from the sensors in temporal domain well. They have 78% accuracy on complex OSUPEL [26] dataset. Abdul Azim et al. [27] proposed optical flow variant trajectory features. They achieved 94.90% accuracy for KTH online database and 95.36% for weizmann dataset.

Multi-person actions recognition is more challenging than single person. One of the primary reasons for this difficulty is the non-availability of simple datasets for multi-person action recognition [28]. They are generally more complex than that used for single human actions.

Gilbert [29] used two-dimensional Harris corner interest points for multi-class action recognition on multi-KTH dataset [30]. Multi-KTH dataset has same six actions as of KTH [31]. Gilbert used similar approach to that used for pose estimations but their templates are mixture of spatio-temporal features. They classify actions using the maximum-likelihood approach based on the count of

matching action templates. In [32], they extended their methodology to include information from neighborhood to low-level features to make hierarchical classifiers and they achieved more speed and accuracy than provided in [29]. This approach relies too much on the interest points and does suffer with inaccuracies. Our methodology uses interest points in combination with other low level features encoding the spatio-temporal information of given frames to classify actions. The details about our features are given in Sect. 3.

The methods have been performed over either one of the KTH [31] or Weizmann [33] datasets for evaluation, consisting of unifomity across all the data instances. Our approach performs better in a way that it is independent of this uniformity. It includes train data from more than one datasets and accordingly accounts for non-uniform specifications of the data. Moreover, we have devised a system to classify multi human actions which has been evaluated on our in-house, self-generated dataset. For this dataset too, the accuracies are in agreement with our results for public datasets. Our contribution includes the generalization of basic action categories to extend the action recognition to other actions not covered in this paper and a faster approach for action recognition in general making use of SMO [19] to train polykernel. Our dataset includes more actions than covered in multi-KTH and is more challenging than multi-KTH [30].

3 Methodology

Our proposed system comprises of three major stages: preprocessing, feature extraction and classification. We employed different features extraction approaches to describe activities in series of frames. Considering significance of each feature, heterogeneous features are used to cover all poses. An overview of proposed framework is shown in Fig. 2.

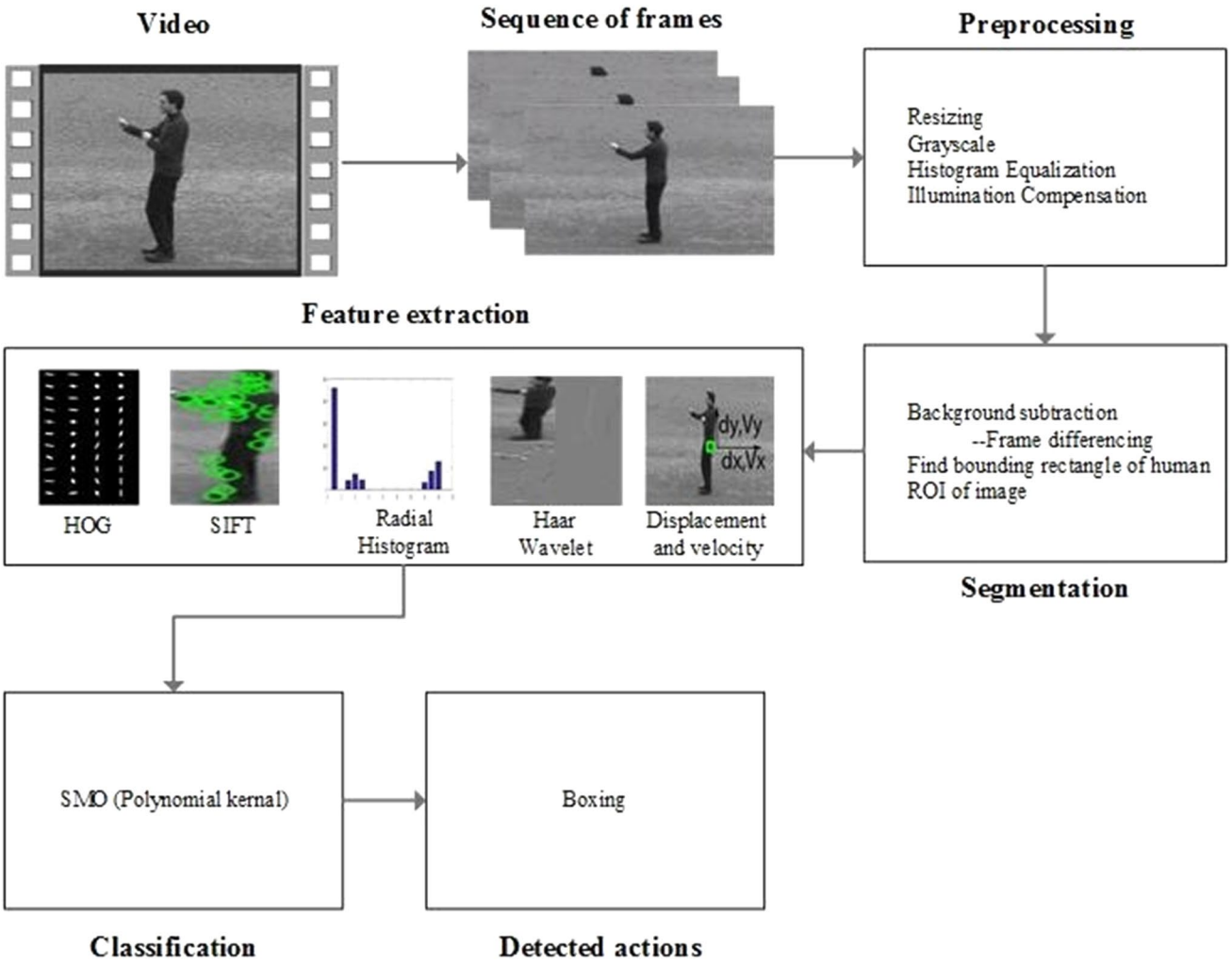

Fig. 2 Framework for proposed activity detection system

3.1 Preprocessing

Extracted frames from video are resized to 128×64 for calculating fixed number of features. We exploited stable background to segment out human body using background subtraction. We extracted silhouettes by performing frame differencing. Bigger connected components are used to filter out human silhouette from other connected components (Fig. 3). In case of multi-human activity, multiple silhouettes are extracted from the image depending upon the number of humans in the frame. Extracted silhouettes are forwarded to the system for centroid detection. Region of interest from image is selected on the basis of silhouette boundary and cropped out for feature calculation as shown in Fig. 4.

3.2 Features

This approach towards human activity detection incorporates fusion of features.

Details of each feature are specified in this sub-section along with their importance in classification.

Velocity and displacement are calculated using silhouette center point marked as P1 and P2 for two frames respectively. Movement of these two points is employed to calculate velocity and displacement. In our method, these two features are averaged out for series of frames.

(a) Single human silhouette **(b)** Region of interest

Fig. 3 Single human frame preprocessing

Displacement is simply the difference between points P1 and P2, while velocity is calculated by dividing displacement with time between two frames. These two features are helpful in distinguishing between activities, for example boxing and walking where former involves negligible body movement in comparison to later.

Histogram of oriented gradients (HOG) gives us the count of localized gradient directions in a specified image block. The bins of histogram represent the distribution of edge directions. HoG works locally by dividing image into cells. These cells represent portions over which the histogram of gradients is created. These histograms are included into a bigger block for normalization to avoid effect of lighting variations. Normalization is carried by using maximum value of a block. Histograms of each separate block are concatenated to form one HoG representation of the image. Two basic steps of HoG are:

1. Computation of gradient values
2. Bins formation of these computed orientations

HOG features have shown their significance in pedestrian detection. Based on this fact, we used importance of these features in combination with other well-known features. We built poselets of human actions using HOG descriptor. It gives us a global edge descriptor for activity detection. In proposed method, we used sequence of frames from a video. For each frame HOG determines the frequency of gradients. An average of all the extracted features is taken. We have used 24×128 window size and [–1, 0, +1] mask for filtering. Histogram is calculated for individual cells of 8×8 grid size. Histograms of these cells are normalized by incorporating them into a larger block of 16×16 pixels using Eq. 1.

$$f = \frac{m}{\sqrt{\|m\|_2^2 + c}} \tag{1}$$

'm' represents a non-normalized vector containing histograms of a block, whereas 'c' denotes a constant.

Fig. 4 Multi human frame preprocessing

(a) Multi human frame preprocessing **(b)** First person ROI **(c)** Second person ROI

Local binary pattern (LBP) is a texture analysis tool used to recognize human activities in [34–36]. It is robust to illumination changes. It extracts local features of whole image considering a 3×3 neighborhood at a time. A constant binary number is generated for each pixel in a window, carried out by comparing the center pixel $I(x_c,y_c)$ with pixels $I(x_{c0}, y_{c0}), I(x_{c1}, y_{c1}), \ldots I(x_{c7}, y_{c7})$ surrounding the center. Let us assume pixel value is represented by v and function that transform intensity into binary value is $s(x)$ given by Eq. 2:

$$s(x) = \begin{cases} 1 \text{ if } x > 0 \\ 0 \text{ if } x < 0 \end{cases} \tag{2}$$

where $x = v_i - v_c$ for $0 \le i \le 7$.

The LBP pattern of $I(x_c, y_c)$ is represented by the decimal number equivalent to binary pattern obtained by Eq. 3:

$$LBP(x_c, y_c) = \sum_{i=0}^{7} s(v_i - v_c)2^i \tag{3}$$

Total 8192 LBP features are extracted from a single image. These features in combination with others are used for human activity classification.

Haar wavelet is known for data analysis and image compression. Information in input image is divided into approximation and detail sub-signals. Approximation sub-signal is calculated by applying low pass filter in horizontal and vertical direction (LL) to produce top left segment in Fig. 5. Detail sub-signals involve three types of high frequency components. First sub-signal is obtained by using horizontal high pass and vertical low pass filter (HL) for top right block. Low pass for horizontal and high pass filter (LH) for vertical direction is applied to get second bottom left block. Finally, high pass filter (HH) is applied in both directions to get third sub-signal i.e. lower right segment of output image in Fig. 5.

It can be seen in Fig. 5, top left segment contains maximum energy and least in lower right. Edges information are preserved only in LH and HL, therefore we used LL, LH, and HL as features for classification. Forward wavelet transform is defined using Eq. 4.

$$\begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} = F \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \tag{4}$$

where x_{ij} is input matrix, y_{ij} is output matrix and F is a translation filter and can be defined as in Eq. 5

$$F = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \tag{5}$$

Scale invariant feature transform (SIFT) molds image data into a constant scale of coordinates with respect to regional features. Computation of SIFT points includes following basic steps:

1. *Extrema detection using Gaussian* function makes use of the scale-space to detect same object from more than one view angles, irrespective of the change in scale. The extrema are evaluated by employing difference of Gaussian function, after which each pixel is compared to its 8 neighbors: this value being either greater or lesser than all the neighboring values indicates a local extrema.
2. *Keypoint localization* involves filtering points of interest out of all the detected keypoints by simple thresholding.
3. *Orientation assignment* of detected keypoints comprises up of magnitude and orientation of gradient calculation to bring consistency in local orientation of keypoints.

The SIFT features resulted in a dynamic number of keypoints for every frame independent of the previous frames. We manipulated SIFT features such that it provides us constant number of keypoints for every frame. To achieve the goal, we made use of Algorithm 1 where upper bound (u) is 20 and lower bound (l) is 13. Detected keypoints are used as features in association with other features to enhance classification performance.

Angular histogram is helpful for various human actions involving different poses, each of them possessing a

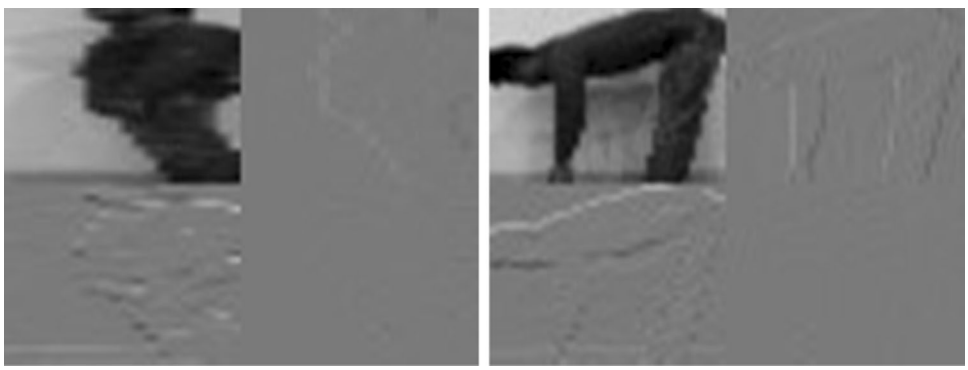

Fig. 5 Wavelet features of activities standing and bending (left to right)

unique spread in space. To incorporate this information, we employed radial/angular histogram which is an efficient shape descriptor and describes varying body postures effectively.

Algorithm 1 SIFT Key-points manipulation

```
procedure KEY-POINTS MANIPULATION Input: Key points(KPs), Upper bound (u), Lower bound (l)
    Output: Specific length features
      if length(KPs) > u then
        KPs = Sort(KPs)(0 : u)
    else if length(KPs) > l and length(KPs) < u then
        i ← 0
        for j == length(KPs) :: u do
            KPs(j) = Avg(KPs(i),KPs(i + 1))
            i = i + 2
        end for
        else if length(KPs) < l then
            KPs(length(KPs) : u) = 0
        end if
    end procedure
```

The rectangular ROI consisting of the detected human body silhouette has been divided into a grid of four blocks to create radial histogram. Angles of each white pixel are calculated by taking the center of each window as the origin. Four histograms of 18 bins are created, one for each of the four blocks. Every bin covers 20°. All four of them are then averaged to obtain a single histogram for one silhouette, reducing its dimension from 72 bins to only 18 (Fig. 6).

3.3 Fusion of features

Recognizing complex human activities require multiple local and global features detection and their interaction for defining motion accurately. It has been accepted by research community that merging different features give good classification results [37]. Activity of a human can be characterized by locomotion of a human body. There can be two possibilities for motion. First is the movement of body parts without relocating complete body i.e. clapping, waving, boxing and bending. Second is the movement of complete body from one point to other i.e. walk, run, jump and side walk. For incorporating all aspect of activity detection different features need to be combined.

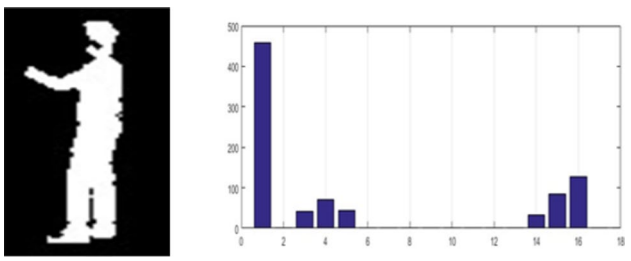

Fig. 6 Human body silhouette on left side and its angular histogram on right side

Algorithm 2 Complete fusion of features algorithm

```
procedure FUSION OF FEATURES                Input: Video sequence, Background frame
    Output: Predicted action for each group of 25 frames
    FramesList ← videoToFrame(V ideo)
    backgroundSubtraction(FramesList,backgroundFrame)
    for i == 0 : length(FramesList) do
        group ← FramesList(i : i + 25)
        groups ← group
    end for
    for i == 0 : length(groups) do

        for j == 0 :: length(groups(i)) do
            frame ← group[j]
            HOGList = avg(HOG(frame),HOGList)
            SIFTList = avg(SIFT(frame),SIFTList)
            RList = avg(R(frame),RList)
            WList = avg(W(frame),WList)
            if j == 0 then
                P1 = Centroid(Silhouette)
            else if j == length(group) − 1 then
                P2 = Centroid(Silhouette)
            end if
        end for
        d_x = |P1.X − P2.X|
        d_y = |P1.Y − P2.Y|
        v_x = d_x / Timefor25frames
        v_y = d_y / Timefor25frames
        F = Append(d_x, d_y, v_x, v_y, HOGList, SIFTList, RList, WList)
        Result = SMO_Classifier(F)
        Action_List.push(Result)
    end for
end procedure
```

Combination of different features is carried out by initially dividing complete video into multiple groups of 25 frames each, followed by feature calculations for every individual frame in a group. Velocity, displacement, HoG, SIFT, radial histogram and wavelet features are calculated for each individual frame and averaged out for single group. Finding all feature sets for complete group is followed by combining them into a single list of features, used in classification. Complete steps are shown in algorithm 2.

3.4 Classification

We have a large number of sparse features with dimensions exceeding up to 5000. Training a huge data using classical quadratic optimization problem is costly and inefficient. Therefore, we selected SMO [19] which works best on sparse data with high dimensions [38]. Classification results were improved by using it. SMO not only speeded up the training process, it also reduced classification error. The accuracy we obtained using SVM radial basis function (RBF) was 80.486% and 91.80% for SVM polynomial on single human activity dataset. It improved up to 91.99% when SVM trained using SMO. Smaller discrepancies are shown, when SVM polynomial is trained using SMO. This is occurring because SMO optimizes lagrange multipliers analytically, in opposite to traditional SVM training that considers optimizing all multipliers at a time. A global maximum (or minimum) is achieved by

SMO, while SVM with polynomial kernel has value closer to maxima (or minima) [19].

SMO uses the same quadratic optimization problem (QP) that SVM uses and does little manipulation to achieve reduced training time and optimized lagrange multipliers. It breaks down the data into best possible smaller QP problem chunks and tries to optimize two lagrangians at a time analytically. The SVM dual quadratic maximization problem is represented as:

$$\begin{aligned} &\max_{\lambda} \sum_{j=1}^{m} \lambda_j - \frac{1}{2} \sum_{j=1}^{m} \sum_{k=1}^{n} \lambda_j \lambda_k y_j y_k x_j x_k \\ &0 \leq \lambda_j \leq C, \forall_j \\ &\sum_{j=1}^{m} y_j \lambda_j = 0 \end{aligned} \tag{6}$$

where, λ is represented by lagrange multiplier, x is input data, whereas y represents class label.

The working of SMO is such that it optimizes two lagrangians at a time, keeping all other lagrangians constant, while satisfying equality and box constraints. Considering two lagrange multipliers λ_1 and λ_2 to be optimized and keeping all other multipliers $l_3, l_4, \ldots, l_m$ to be constants, the equation becomes:

$$\lambda_1 y_1 + \lambda_2 y_2 = -\sum_{j=3}^{m} \lambda_j y_j \tag{7}$$

We can re-write the right hand side of equation by replacing it with some constant c:

$$\lambda_1 y_1 + \lambda_2 y_2 = c \tag{8}$$

The linear Eq. (8) will be used to optimize over λ_1 and λ_2. The optimal value of λ_1 is achieved by finding $\lambda_{1new,unclipped}$, restricting it with in upper bound U and lower bound L limits, mentioned in [19].

$$\lambda_1 = \begin{cases} U, & \text{if } \lambda_1^{new,unclipped} > U \\ \lambda_1^{new,unclipped} & \text{if L} \leq \lambda_1^{new,unclipped} \leq H \\ L & \text{if } \lambda_1^{new,unclipped1} \leq L \end{cases}$$

The similar procedure is employed to find other optimal lagrange multipliers $l_1, l_2, \ldots, l_m$. Classifier decision boundaries are determined by these optimized multipliers.

4 Experimentation

We have used three datasets for training and testing our algorithm:- KTH human action dataset [31] and Weizmann human action dataset [39] for single human activity recognition, multi-human dataset for multiple human's activity recognition. The accuracy comparative analysis is presented in this section including parameters setting for classifier and dataset description.

4.1 Parameters setting

The classifier accuracy tends to vary with different parameters setting. The parameters that we used are mentioned here. Polynomial kernel function is used for classification of our large non-linear data with cross-validation of 10 folds. The normalization step was carried out before training. Round off error was set to 1.0E−12.

4.2 Datasets

KTH dataset comprises of 600 videos of 6 human actions (boxing, handwaving, hand-clapping, jogging, running and walking). These action were performed by 25 different subjects in 4 scenarios: outdoor, indoor, outdoor with scale variation, and different clothes. The frame rate of videos is 25 with 160×120 resolution, examples are shown in Fig. 7. Each video has variable duration ranging from 10 s to a minute. In each video, there are 300–1500 frames and actions on average are repeated after 30 frames. The maximum accuracy achieved on this dataset is around 92–94% [40, 41, 27].

Weizmann dataset has total 93 videos of 10 human actions performed by 9 different persons, look at Fig. 7. There are 40–120 frames per video with resolution 180×140. The dataset videos length ranges between 1 and 5 s. The highest accuracy on this dataset reported in literature is 100% [42, 24].

Multi-human dataset We created our own dataset of 13 human activities (walking, running, jogging, jump, pjump, side, boxing, hand clapping, bend, jack, skip, wave1 and wave2) of two persons performing activities independently. The dataset has 130 videos of 10 actors; each subject performed all 13 activities where each video has similar activity, for example, in boxing video both persons conducted boxing. The dataset was captured with static camera under varying lighting conditions. Sample images are shown in Fig. 8.

5 Results and discussion

5.1 Single human activity recognition

Both KTH and Weizmann human activity recognition datasets were used for training and testing. Selected actions from KTH are boxing, walking, hand-clapping, running and jogging, while from Weizmann bend, jack, jump, pjump, side, skip, wave1 and wave2 are selected. We randomly picked 4 videos of these actions from Weizmann dataset and 9 videos

Fig. 7 Sample images from KTH and Weizmann dataset

Fig. 8 Sample images from our multi-human dataset

from KTH dataset for training. We extracted features of 25 frames representing one human action and trained SMO on these features. The confusion matrix is shown in Table 4.

We have tested activity recognition by adding and removing different features. The accuracy results vary with different combinations of velocity (V), displacement (D), SIFT, HOG, LBP, radial histogram (R) and Haar wavelet (W). We selected a best possible feature set that showed the highest accuracy. A comparison of features combinations and respective accuracies is shown in Table 1. The complete feature set achieves 91.038% accuracy. The classification accuracy reduced to 88.878% by removing HOG features and increased to 91.99% by taking out only LBP features. The best results are obtained without LBP features. Feature dimensions are also reduced to 5910 from 14,102 dimensions with exclusion of LBP. Table 1 shows detailed accuracy comparison with different features.

Comparison among other well-known classification algorithms is also provided in Table 1. We used same feature set to see the performance variation on other algorithms. It can be seen that SMO is out-performing other classifiers. SVM with polynomial kernel has performance closer to SMO. All

Table 1 Accuracy comparison of different features on single human dataset

	No. of feature	SMO (%)	SVM (RBF) (%)	SVM (Poly) (%)	Decision Tree (%)	Random Forest (%)	Naive Bayes (%)
V + D + SIFT + R + HoG + W (ours)	5910	91.99	80.486	91.80	67.279	84.835	77.895
V + D + SIFT + R + LBP + W	10,322	88.878	89.60	89.121	64.705	75.965	75.0
V + D + SIFT + R + HoG + LBP	12,055	91.085	89.597	90.923	67.141	83.18	74.126
V + D + SIFT + R + HoG	3863	91.682	67.601	91.524	68.336	85.248	74.586
V + D + SIFT + R + LBP	8275	88.741	86.451	87.879	64.154	75.459	69.76
V + D + R + HoG + W	5851	91.912	87.546	90.298	67.417	84.826	78.079
V + D + SIFT + R + HoG + W + LBP	14,102	91.038	89.30	91.0	76.378	82.169	76.378

other algorithms: decision tree, random forest, naive bayes failed to produce good results on our high dimensional sparse feature vector.

An accuracy of 100% has been achieved on classes boxing, jack, pjump, and wave2, while six classes bend, hand clapping, jump, side, skip, and wave1 showed 97% accuracy. Locomotive actions (walking, running, and jogging) have the highest amount of confusion among each other, hence shown the lowest accuracy, Table 4.

In Table 1, V is velocity, D is displacement, R represents radial histogram, and W corresponds to haar wavelet.

Proposed approach performs better than most of the recent publications. We have very less accuracy variation on KTH and Weizmann dataset that shows the robustness of our method. In comparison to our methodology, Klaser et al. [43] have accuracy decreased to 84.3% for Weizmann dataset while it is 91.4% on KTH dataset. Laptev et al. [44] has equivalent accuracy to our approach on Weizmann dataset. The comparative results are shown in Table 2.

Table 2 Accuracy comparison on KTH and Weizmann with other publications

Method	KTH (%)	Weizmann (%)
Proposed method	92.286	91.695
Fathi et al. [42]	90.5	100
Niebles et al. [4]	83.3	90.0
Abdul-Azim et al. [45]	94.90	95.36
Ji et al. [46]	90.2	–
Raja et al. [47]	86.6	–
Laptev et al. [44]	–	91.8
Klaser et al. [43]	91.4	84.3

5.2 Multi-human activity recognition

For multi-human activity training and testing was carried out on our own dataset in a similar way that was performed on single human activity recognition with similar features. The recognition for each person works separately. Two distinct feature sets are calculated for each person, and classified independently. The confusion matrix for multi-human activity is shown in Table 5. Similar to single human activity walking, running and jogging have lowest accuracies because of maximum confusion. The classes hand clapping and pjump had the maximum accuracy of 100%. Bend, boxing, jack, and wave1 are more than 94% accurate, see Table 5 for detailed accuracy comparison of each class. Accuracy varies for different number of features similar to single human activity. Feature set without LBP has attained 86.48% classification accuracy. In opposite to this, 0.214% and 0.429% improvements in accuracy have been shown without LBP + radial histogram and SIFT + LBP + radial histogram respectively. Table 3 shows accuracy comparison with different number of features. Akin to single human activity, all other classifiers have degraded performance in contrast to SMO, although SVM with polynomial kernel is approximately equal to SMO in performance.

Tables 3 and 1 depict the accuracy for different sets of combinations of features. Rationale behind using this diverse sets of features is diverse dynamic nature of activity. Some of the activities include only performing some specific action on same place while others include locomotion. Activities with some locomotion are best depicted using velocity and displacement measures. SIFT provide interesting local features of an image irrespective of the spatial variation of the action. Haar wavelet and local binary pattern are used for appearance attributes of activity.

From Tables 3 and 1, we can conclude the effect of different features in activity detection. As shown by the accuracies of Tables 3 and 1; velocity, displacement, SIFT, angular histogram, HOG and wavelets are the best descriptor for activity detection as they combine the local and global properties of a human body. LBP is discarded as it considerably slows down the system without giving a significant boost in terms of accuracy. LBP is useful for applications where detailed texture description is required, as in face recognition. So we preferred a combination of velocity, displacement, SIFT, angular histogram and HoG for activity detection.

The confusion matrices in Tables 4 and 5 show the test set accuracies obtained for the best performing system proposed on 13 action classes. The actions with similar body and limb

Table 3 Accuracy comparison of different features on multi-human dataset

	No. of features	SMO (%)	SVM (RBF) (%)	SVM (Poly) (%)	Decision tree (%)	Random forest (%)	Naive Bayes (%)
V + D + SIFT + R + HoG + W (ours)	5910	86.48	62.875	86.30	49.570	77.038	71.888
V + D + SIFT + R + LBP + W	10,322	81.331	80.042	81.0	38.412	70.171	63.948
V + D + SIFT + R + HoG + LBP	12,055	85.407	80.043	85.339	48.068	75.536	66.094
V + D + SIFT + R + HoG	3863	86.051	42.704	85.351	55.15	77.897	67.54
V + D + SIFT + R + LBP	8275	80.687	80.043	80.121	38.412	70.171	60.944
V + D + R + HoG + W	5851	86.481	75.965	86.512	50.0	74.892	72.317
V + D + SIFT + R + HoG + W + LBP	14,102	84.335	80.043	83.989	47.210	72.345	66.952

Table 4 Confusion matrix for KTH and Weizmann human activity dataset

	Bend	Boxing	Hand clapping	Jack	Jogging	Jump	Pjump	Running	Side	Skip	Walking	Wave1	Wave2
Bend	1	0	0	0	0	0	0	0	0	0	0	0	0
Boxing	0.0023	0.997	0	0	0	0	0	0	0	0	0	0	0
Hand clapping	0	0.008	0.77	0.15	0.067	0	0	0	0	0	0	0	0
Jack	0	0	0.15	0.76	0.076	0	0	0	0	0	0	0	0
Jogging	0	0	0.049	0.067	0.88	0	0	0	0	0	0	0	0
Jump	0	0	0	0.02	0.03	0.95	0	0	0	0	0	0	0
Pjump	0	0	0	0	0	0.015	0.985	0	0	0	0	0	0
Running	0	0	0	0	0	0	0	1	0	0	0	0	0
Side	0	0	0	0	0	0	0	0.02	0.98	0	0	0	0
Skip	0	0	0	0	0	0	0	0	0.02	0.98	0	0	0
Walking	0	0	0	0	0	0	0	0	0	0	0.02	0.98	0
Wave1	0	0	0	0	0	0	1	0	0	0	0.03	0.97	0
Wave2	0	0	0	0	0	0	0	0	0	0	0	0.02	0.98

Table 5 Confusion matrix for Multi-human activity dataset

	Bend	Boxing	Hand clapping	Jack	Jogging	Jump	Pjump	Running	Side	Skip	Walking	Wave1	Wave2
Bend	0.985	0	0	0.015	0	0	0	0	0	0	0	0	0
Boxing	0.04	0.94	0	0.02	0	0	0	0	0	0	0	0	0
Hand clapping	0	0	1	0	0	0	0	0	0	0	0	0	0
Jack	0.02	0	0	0.96	0	0.02	0	0	0	0	0	0	0
Jogging	0	0	0	0	0.7	0.1	0	0.05	0	0.1	0.05	0	0
Jump	0	0	0	0	0.083	0.75	0	0	0	0.083	0.083	0	0
Pjump	0	0	0	0	0	0	1	0	0	0	0	0	0
Running	0	0	0	0	0.5	0	0	0.375	0	0	0.125	0	0
Side	0	0	0	0.02	0.02	0.02	0	0	0.84	0.06	0.04	0	0
Skip	0	0	0	0	0.06	0.1	0	0.03	0.16	0.6	0.03	0	0
Walking	0	0	0	0	0.087	0.0217	0	0	0	0.065	0.82	0	0
Wave1	0	0	0.05	0	0	0	0	0	0	0	0	0.95	0
Wave2	0	0.025	0.025	0	0	0.025	0	0	0	0	0	0.025	0.9

movements become a cause of slight overlap among few of these actions. Boxing and hand clapping show relatively similar movements of arms. This leads to misclassification in some instances. On the similar pattern, jogging and walking both include significant lower body movements. In case of multi-human actions: jumping, jogging, skipping and walking show minor intermixing due to similarity in the body movements of these actions. Waving by one hand

and waving using two hands show intermixing but since the second action involves use of another arm, which acts as a classifying trait, the overlap between the two actions is insignificant. The system presents a good overall accuracy and only a minute number of false classifications on the test data with 91.99% and 86.48% accuracies in case of single and multi-human action datasets respectively.

6 Conclusion

The methodology makes use of an amalgam of features i.e., velocity, displacement, HOG, Radial histogram, LBP, haar wavelets and SIFT feature points. Multiple combinations of these features have been tested over the classifier to select the one with best results. LBP and haar wavelet have been observed to lower the accuracy for which they have been excluded from the final combination.

Support vector machine and sequential minimal optimization have been employed for training purposes. The method applies to both single human and multi human action scenarios. The system has been trained and evaluated over 13 human actions in both cases. The evaluation and experimentation has been carried out on standard KTH and Weizmann action datasets for single human actions and on a self-generated in-house dataset for multi human dataset. An accuracy of 91.99% has been achieved in case of single human actions and 86.48% upon testing for multi human actions. This system can be extended for human–human interactions and human-object interactions.

References

1. Moussa MM et al (2015) An enhanced method for human action recognition. J Adv Res 6(2):163–169
2. Kolekar MH, Dash DP (2016) Hidden markov model based human activity recognition using shape and optical flow based features. In: Region 10 conference (TENCON) (2016 IEEE) IEEE
3. Yang J, Cheng J, Lu H, Human activity recognition based on the blob features. In: IEEE international conference on multimedia and expo, pp 358361, 2009
4. Niebles JC, Fei-Fei L (2007) A hierarchical model of shape and appearance for human action classification. In: IEEE conference on computer vision and pattern recognition, 2007. CVPR'07. IEEE
5. Ke S-R et al (2013) A review on video-based human activity recognition. Computers 2(2):88–131
6. Li W, Zhang Z, Liu Z (2010) Action recognition based on a bag of 3d points. In: 2010 IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW). IEEE
7. Li W, Zhang Z, Liu Z (2008) Expandable data-driven graphical modeling of human actions based on salient postures. IEEE Trans Circ Syst Video Technol 18(11):1499–1510
8. Wang H et al (2013) Dense trajectories and motion boundary descriptors for action recognition. Int J Comput Vis 103(1):60–79
9. Kellokumpu V, Pietik ainen M, Heikkila J (2005) Human activity recognition using sequences of postures. MVA, pp 570–573
10. Wang H, Schmid C (2013) Action recognition with improved trajectories. In: Proceedings of the IEEE international conference on computer vision
11. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. In Proceedings of the 17th IEEE international conference on pattern recognition (ICPR), Cambridge, UK, 2326 August 2004; vol 3, pp 32–36
12. Vapnik V (1999) The nature of statistical learning theory. Springer, New York
13. Hoang LUT, Ke S, Hwang J, Tuan PV, Chau TN (2012) Quasi-periodic action recognition from monocular videos via 3D human models and cyclic HMMs. In: Proceedings of IEEE international conference on advanced technologies for communications (ATC), Hanoi, Vietnam, 1012 October 2012; pp 110–113
14. Yamato J, Ohya J, Ishii K (1992) Recognizing human action intime- sequential images using hidden markov model. In: IEEE computer society conference on computer vision and pattern recognition, pp 379–385
15. Brand M, Oliver N, Pentland A (1997) Coupled hidden markov models for complex action recognition. In: Proceedings of IEEE computer society conference on computer vision and pattern recognition (CVPR), San Juan, PR, USA, 1719 June 1997; pp 994–999
16. Duong TV, Bui HH, Phung DQ, Venkatesh S (2005) 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), San Diego, CA, USA (2005) June 2005; vol 1, pp 838–845
17. Fiaz MK, Ijaz B (2010) Vision based human activity tracking using artificial neural networks. In: Proceedings of IEEE international conference on intelligent and advanced systems (ICIAS), Kuala Lumpur, Malaysia, 1517 June 2010; pp 15
18. Umakanthan S, Denman S, Fookes C, Sridharan S (2014) Activity recognition using binary tree SVM. In: IEEE workshop on statistical signal processing, pp 248–251
19. Platt J (1998) Sequential minimal optimization: a fast algorithm for training support vector machines. Technical Report MSR-TR-98-14, Microsoft Research
20. Kolekar MH, Sengupta S (2004) Hidden markov model based structuring of cricket video sequences using motion and color features. In: Indian conference on computer vision graphics and image processing, pp 632–637
21. Thurau C, Hlavc V (2008) Pose primitive based human action recognition in videos or still images. In: IEEE conference on computer vision and pattern recognition, 2008. CVPR 2008. IEEE
22. Chen M, Hauptmann A (2009) MoSIFT: recognizing human actions in surveillance videos. In: CMU-CS-09-161, Carnegie Mellon University
23. Chathuramali KG, Manosha, Rodrigo R (2012) Faster human activity recognition with SVM. In: 2012 international conference on advances in ICT for emerging regions (ICTer). IEEE
24. Chathuramali KM, Rodrigo R (2012) Faster human activity recognition with SVM. In: International conference on advances in ICT for emerging regions, pp 197–203
25. Liu L, Wang S, Su G, Huang ZG, Liu M (2017) Towards complex activity recognition using a Bayesian network-based probabilistic generative framework. Pattern Recogn 68:295–309
26. Brendel W, Fern A, Todorovic S. Probabilistic event logic for interval-based event recognition. In: 2011 IEEE conference on computer vision and pattern recognition (CVPR) 2011 Jun 20, pp 3329–3336. IEEE
27. Abdul-Azim HA, Hemayed EE (2015) Human action recognition using trajectory-based representation. Egypt Inform J 16:187198

28. Chaquet JM, Enrique J, Carmona, Fernndez-Caballero A (2013) A survey of video datasets for human action and activity recognition. Comput Vis Image Underst 117(6):633–659
29. Gilbert A, Illingworth J, Bowden R (2008) Scale invariant action recognition using compound features mined from dense spatio-temporal corners. In: European conference on computer vision. Springer, Berlin
30. Uemura H, Ishikawa S, Mikolajczyk K. Feature tracking and motion compensation for action recognition. In: Proc. of BMVA british machine vision conference
31. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. In: proceedings of the 17th international conference on pattern recognition, 2004. ICPR 2004, vol 3. IEEE
32. Gilbert A, Illingworth J, Bowden R (2009) Fast realistic multi-action recognition using mined dense spatio-temporal features. In: 2009 IEEE 12th international conference on computer vision. IEEE
33. Gorelick L et al (2007) Actions as space-time shapes. IEEE Trans Pattern Anal Mach Intell 29(12):2247–2253
34. Uddin Md, Zia et al (2013) An indoor human activity recognition system for smart home using local binary pattern features with hidden markov models. Indoor Built Environ 22(1):289–298
35. Mattivi R, Shao L (2009) Human action recognition using LBP-TOP as sparse spatio-temporal feature descriptor. In: International conference on computer analysis of images and patterns. Springer, Berlin
36. Kellokumpu V, Zhao G, Pietikinen M (2011) Recognition of human actions using texture descriptors. Mach Vis Appl 22(5):767–780
37. Bruzzone L, Persello C (2009) A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability. IEEE Trans Geosci Remote Sens 47(9):3180–3191
38. Joachims T. Training linear SVMs in linear time. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2006
39. Blank M et al (2005) Actions as space-time shapes. In: Tenth IEEE international conference on computer vision, ICCV 2005, vol 2. IEEE, 2005
40. Schindler K, Van Gool L (2008) Action snippets: how many frames does human action recognition require?. In: IEEE conference on computer vision and pattern recognition, 2008. CVPR 2008. IEEE
41. Jhuang H et al (2007) A biologically inspired system for action recognition. In: IEEE 11th international conference on computer vision, ICCV 2007. IEEE, 2007
42. Fathi A, Mori G (2008) Action recognition by learning mid-level motion features. In: IEEE conference on computer vision and pattern recognition, 2008. CVPR 2008. IEEE
43. Klaser A, Marszaek M, Schmid C (2008) A spatio-temporal descriptor based on 3D-gradients. In: BMVC
44. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: IEEE CVPR
45. Abdul-Azim HA, Hemayed EE (2015) Human action recognitionusing trajectory-based representation. Egypt Inform J 16:187198
46. Ji S, Xu W, Yang M, Yu K (2013) 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
47. Raja K, Laptev I, Perez P, Oisel L (2011) Joint pose estimation and action recognition in image graphs. In: International conference on image processing, Brussels, Belgium, Sept. 2011, pp 2528