Received: January 1, 2019


International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

# View-Invariant Feature Representation for Action Recognition under Multiple Views

**Kumbala Pradeep Reddy[1]*** **Gullipalli Apparao Naidu[2]** **Bulusu Vishnu Vardhan[3]**

[1]*Department of Computer Science Engineering, Tirumala Engineering College, Telangana, India*
[2]*Department of Computer Science Engineering, JB Institute of Engineering and Technology, Telangana, India*
[3]*Department of Computer Science Engineering,*
*Jawaharlal Nehru Technological University College of Engineering, Manthani, Telangana, India*
* Corresponding author's Email: kumbalapradeepreddy@gmail.com

**Abstract:** Robust Action Recognition under multiple views has gained a significant research interest recently. To enhance the performance of Multi-view Action Recognition, we propose a novel Feature extraction and Feature Selection mechanism that allows building a mutual relationship between the actions sequences of multiple views. The feature extraction considered multiple features which are invariant to scale and orientation. Three different features such as Intensity Features, Orientational Features and Contour Features are used to represent every action. Further, the feature selection is accomplished through self-similarity matrix and is very much helpful in the provision of a perfect discrimination between actions sequences of different views. The proposed method is validated over the standard multi-view IXMAS dataset and experimental results confirm that the proposed method outperforms the conventional approaches with respect to Recognition Accuracy.

**Keywords:** Action recognition, Multiple views, Scale invariant features, Orientation invariant features, Self-similarity matrix, Accuracy.

## 1. Introduction

In recent years, Human Action Recognition (HAR) based understanding from image and video has gained a great research interest in computer vision due to its widespread applicability in various applications including Robotics, Human-computer Interactions, Behavior analysis [1], Content based Retrieval, Video Indexing, Gesture Recognition and Visual Surveillance [2]. The main objective of a HAR system is to identify actions in a video sequence under different situations like occlusion, cluttering and different lighting conditions. The main center of this system is the computational algorithms which understand the human actions. Similar to the human vision system, these computational algorithms ought to produce a label after the analysis of partial or entire action in the video sequence [3, 4]. Developing such algorithms is typically addressed in the computer vision research, which studies how to make the computers to gain high level understanding regarding human actions from digital images and videos? Various solutions are developed in earlier for action recognition over years. In any HAR system, first the action needs to be represented in a machine understandable format. Space-time shapes [5], Covariance Features [6], Time Evolution based Human Silhouettes [7], and Local 3D Patch Descriptors [8] are some of the most popular techniques used for action representation. The further representation used feature descriptors such as Space-time Interest Points (STIP) [9], and Self-Similarity Matrices (SSM) [10] based approaches.

Recently, Multi-View Human Action Recognition (MVHAR) have gained a significant interest due to its effective tackling capability by the accomplishment of multiple cameras and observing an action in multiple views. MVHAR system is more robust than single view HAR system for view

changes; MVHAR considers the view point changes which has a significant impact on the action understanding. Hence the extraction of view invariant features is important. Based on this aspect, a novel MVHAR system is developed in this paper which is more effective in the extraction of features which are view invariant and also scaling and translational invariant. Towards the feature extraction, a hybrid technique is proposed based on Intensity, Orientation and Contour features. Further to reduce the computational burden, this paper accomplished SSM based key frames selection. Extensive simulations conducted over the developed system shows the outstanding performance with respect to accurate action recognition for multiple views.

Rest of the paper is organized as follows; Section 2 illustrates the details of literature survey. Section 3 illustrates the details of proposed mechanism. Section 4 illustrates the details of simulation experiments and finally the conclusions are provided in section 5.

## 2. Related work

Since the main stream of this paper is Multi-view Action recognition, the related work carried out here is related to MVHAR only. Both 2D and 3D based approaches have been addressed for multi-view action recognition and the details are discussed as follows;

J. Liu et al. [12] developed a cross view human action recognition system based on view knowledge transfer. In cross views there exist some high-level features which can share information regarding cross views and they help to build a connection between cross views. To extract such features and to formulate two view-dependent vocabularies, a bipartite graph model is proposed. The bipartite graph partitions the two vocabularies into visual-word clusters called as bilingual words. These words can bridge a semantic gap across view-dependent vocabularies. However this codebook-to-codebook correspondence at video level won't have any guarantee that a pair of videos of two different views has similar feature representation.

J. Zheng and Z. Zhiang [13] presented a joint learning mechanism to learn a common dictionary and a set of view-specific dictionaries for cross view human action recognition. For every view both (Cross-view and view-specific) dictionary features are learned. Here the view-specific features are only specifies about a single and the common dictionary features are for cross-views. This approach mainly focused on the information transfer across views but not jointly modeled the relations among multiple views.

A. Farhadi and M. K. Tabrizi [14] considered the correlations between actions acquired at different views. They used a cluster of code words based split mechanism for each view. These splits are transferred between different views and they are learned to recognize the action. However, the correlation signifies a linear relation but not non-linear relationship between actions at different views.

Gao et al. [15] proposed a new 'multi-view discriminative and structured dictionary learning with group sparsity and graph model (GM-GS-DSDL)' for MVHAR based on the fusion of features obtained in multiple views. For each, view, STIPs are extracted as a feature set and then formulated into Multi-view Bag of Words (MVBoW). Though the STIPs are more prominent, they can't alleviate temporal correlation between the frames of an action under single view.

Zho et al. [16] developed a MVHAR algorithm based on local similarity random forest and sensor fusion. Multi-Sensor fusion is applied to remove the disparities under multiple views and random forest algorithm is applied at segment level to attain a less complexity. However, a simple sensor fusion can't provide a sufficient discrimination between actions under different views.

Z. Gao et al. [17] proposed Multi-Dimensional HAR system based on the image set and group sparsity. Initially, for every view, this approach extracts a dense trajectory feature and then constructs a shared codebook by *k*-means for all views. Further, employs a weighted BoW to code the dense trajectory feature by the codebook for every view. Though the group sparsity provides an effective sparse features for every view, the redundant information is not focused much which creates an additional computational burden over the classifier.

Among the action representation features, Histogram based features are already proved their efficiency in the action recognition and most of the single view HAR techniques have been used the Histogram Oriented gradients (HOG-3D) [18, 19] as feature descriptor. Chun, S. and Lee, C. S [20] proposed a novel descriptor called as Histogram of Motion Intensity and Direction (HMID), to capture the local motion characteristics of Human Action in Multiple Views. Support vector machine is accomplished for classification. Another method based on 2D motion templates, Motion History Images, and HOG was proposed in [21]. HOGs are used as an efficient descriptor of the MHIs and the classification is done through the nearest neighbor

(NN) Classifier. However, the local motion features constitute an additional complexity at the classifier.

Combining the advantages of HoG-3D with STIPs [23] and MoSIFT [24], a novel cross-view recognition framework is developed by Zan Gao et al. [22]. Though this approach achieved a greater recognition results, the feature extraction phase constitutes an abnormal computational time. Furthermore, the redundant information like background under multiple views is needed to remove, which is not focused in this method.

Human silhouette is an effective action representation attribute through which the single view actions are recognized more perfectly. Some authors used silhouette in MVHAR also. For instance, Chaaraoui A A et al. [25, 26] proposed to represent the key poses of human action with the help of contour points of Human silhouette. Though this approach achieved an effective recognition results with less complexity, the silhouette is not robust to scaling and rotation variations. To overcome this problem, Kushwaha A et al. [27]; proposed to extract rotation-invariant local binary patterns (LBPs) and contour points from the silhouette. The classification is done with the help of Multi-class Support Vector Machine (MC-SVM). Further in [28], the coarse silhouette features are combined with motion features and radial-grid based features for multi-view action recognition. However, the Human silhouettes have too much difference for multiple views which results in higher false positives.

Some more approaches are there in which a new matrix called, Self-Similarity Matrix (SSM) is accomplished for Multi-view Action Recognition. For instance, Imran N Junejo et al. [29] proposed a view-independent action recognition framework based on new action descriptor that captures the structure of temporal similarities and dissimilarities within action sequence. However, the alone temporal self-similarities are not effective for view-independent action recognition. Next, J wang and H Zheng [30] found that the SSMs capture global time information which is useful for action recognition in multiple views. Further, Dynamic Time Warping (DTW) is applied for the complete utilization of time information in SSMs. K-nearest neighbor classifier is accomplished for action classification. Though the temporal similarities are extracted effectively with DTW with SSM, the redundant information is not much reduced due to the non-consideration of spatial similarities.

Considering the advantages of wavelet transform, A. A et al. [31] proposed a novel method for MVHAR by integrating the wavelet transform with silhouette. Initially, the contour of human silhouette is extracted and a distance signal is measured. In the next step, the wavelet transform is applied to extract the features of a single view and they are combined with features of multiple views. Finally a hierarchical classifier using SVM and Naïve Bayes classifiers are accomplished for classification of actions. However, the wavelet transform is non-invariant to scaling due to the presence of down sampler. In MVHAR, the feature set must be in such as way that it has to cover all scaling and rotation variations. Next, Kuan Pen Chou et al. [32] proposed to extract the scale-invariant features and used to model the global spatial-temporal distribution. However this method is not robust for inter and intra class variations. Next, A B Sargano et al. [33] presented a novel feature descriptor for MVHAR based on region based geometrical and Hu-moments extracted from the Human silhouette. MC-SVM is accomplished for classification. However, this approach is not focused on the redundant information by which the false positives can be high in number. Furthermore, this approach is also not extracted the temporal or spatial self-similarities by which inter and intra class relations between actions can be obtained.

## 3. Proposed approach

This section describes the details of proposed method for multi-view human action recognition (MVHAR). The block diagram of proposed Human Action Recognition system is depicted in Fig. 1, below.

As shown in the Fig. 1, the proposed HAR system consists of three principal stages, 1) Key Frame Selection, 2) Feature extraction, and 3) Action Recognition. In the key frame selection, only informative and discriminative frames are extracted from every video sequence acquired. Next, in the feature extraction stage, a set of features are extracted from every frame and finally in the third sage, the extracted features are processed for action recognition through machine learning algorithm.

### 3.1 Key frames selection

In this section, the process of key frames selection is described. In MVHAR, for every action, there exists multiple views and they are acquired through multiple cameras. However, in every view, only few frames are informative and remaining are redundant, i.e., consists of same type of information. This paper focused to select only those informative frames and tries to remove the redundant frames

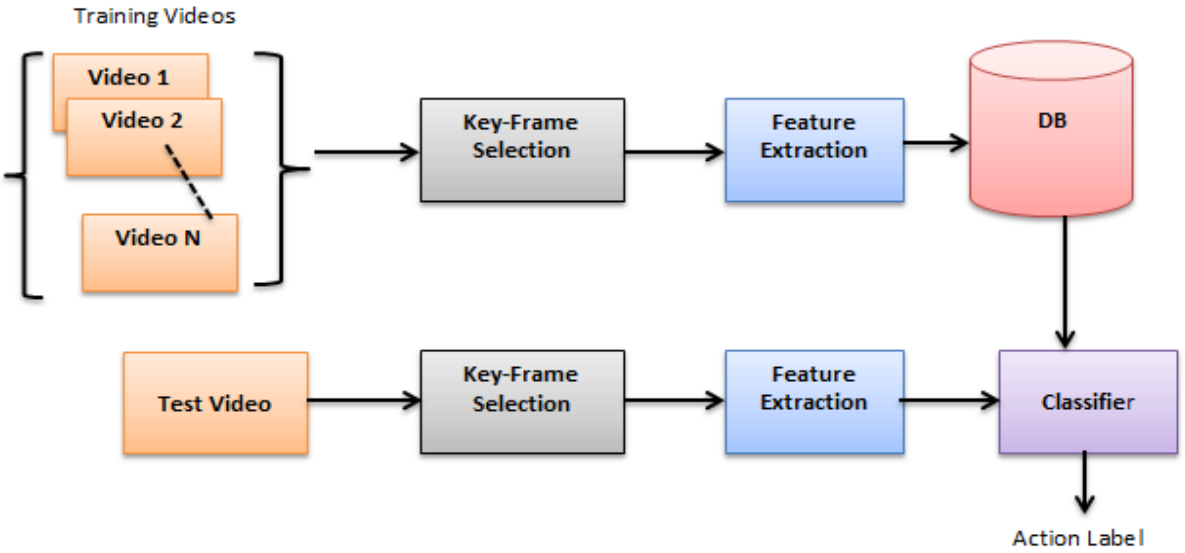


Figure.1 Block diagram of Proposed HAR system

through the key frame selection process. Towards such selection, his paper adopts Self-Similarity Matrix (SSM) Assessment.

### 3.1.1. SSM assessment

Here the main intention of SSM Assessment is to find the similarities in an action video sequence acquired in multiple views with multiple cameras. SSM depicts the similarity observations frame-by-frame for an action video sequence. SSM was initially introduced by Junejo et al. [29] as descriptor for feature extraction which is robust to view changes. Given an Action video sequence $I$ with $T$ frames, $I = [I_1, I_2, \dots\dots\dots\dots\dots, I_T]$, an SSM is obtained as,

$$SSM(I) = \begin{bmatrix} 0 & d_{12} & d_{13} & d_{14} & \dots & d_{1T} \\ d_{21} & 0 & d_{23} & d_{24} & \dots & d_{2T} \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ d_{T1} & d_{T2} & d_{T3} & d_{T4} & \dots & 0 \end{bmatrix} \quad (1)$$

Where $d_{ij} = \|p_i - p_j\|^2$ is the Euclidean distance between pixel intensities of frames $I_i$ and $I_j$. Obviously the diagonal elements in the above matrix are zero which denotes a self-similarity between the same frames. Here the term $d_{ij}$ denotes the similarity between the frames $I_i$ and $I_j$ through their pixel intensities.

The main advantage of SSM assessment is to find the frames which have almost similar information. For example, in an action video sequence, initial frames won't carry any significant information. The frames acquired after the starting

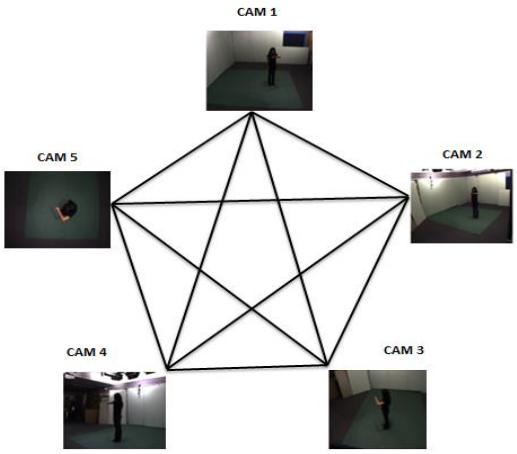


Figure.2 Graphical representation of Multi views

of an action are more informative because, they have variations due to the movements. But in the initial frames, we can observe an almost zero variations which denote that they are redundant. Further there exist some frames which have almost same motion movements. Such types of frames are also considered as redundant. In this paper, the SSM is accomplished to find out the redundant frames. For an every action acquired under multiple views, there exists some common frames and they need to be removed such that the computational time will also reduce. Hence to reduce the number of frames followed by computational time, the SSM is accomplished. Based on this SSM metric, the frames which have almost same information are removed and only few frames are extracted which have much significant information regarding the Human action.

According to Fig. 2, here the SSM is measured between all views. For this purpose, initially the action videos are processed for frame extraction for

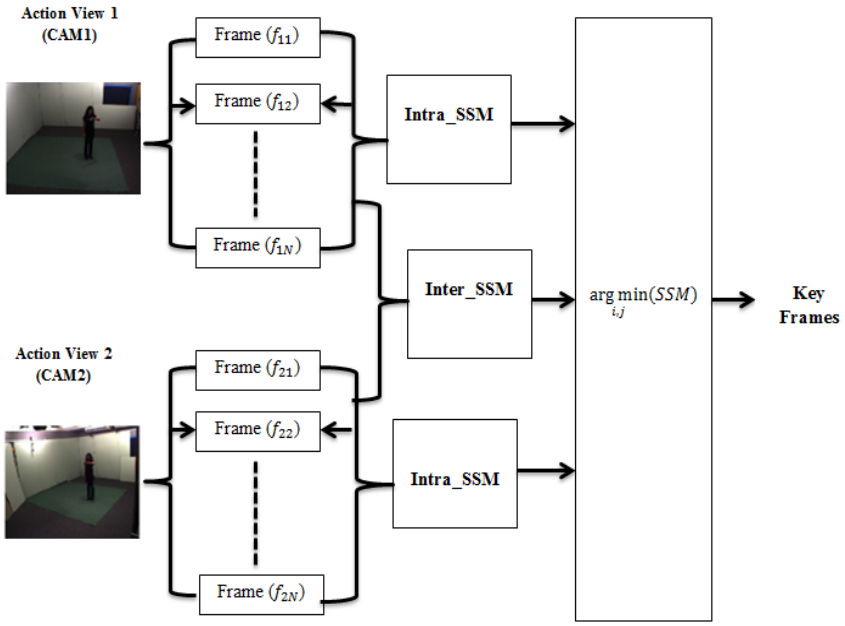


Figure.3 Key frames selection

$N$ number of frames extraction. After frames extraction, the SSM is measured between the same frames of different views. For a given action with multiple views, initially the Intra-SSM is measured between the frames of a single view. Further the Inter-SSM is measured between the same frames of different views. Fig. 3 illustrates the process of key frames selection more clearly.

Over the obtained SSM Matrix, key frames are selected by finding the minimum differences (i.e., minimum $d_{ij}$). For example the first frame of first View is processed for subtraction from the first frame of remaining Views and among the obtained values, a minimum value is evaluated. Based on that minimum values, last N-1 frames are only selected which has lower minimum value, i.e., maximum difference. Simply, we select the frames which have maximum difference with first frame in the case of first frame as reference. Only one frame is excluded which have minimum difference with reference frame. This process is accomplished for second frame and also for further frames. Mathematically, it is performed as

$$[P, V] = \min(SSM(d_{i=1,\dots N, j=1,\dots N})) \quad (2)$$

Where $P$ represents the position of frame which has minimum difference and $V$ represents its minimum value. In this manner, only few key frames are extracted from every View and they are only processed for feature extraction.

### 3.2 Feature extraction

Once the key frames are extracted from every View, they are processed for feature extraction. After extracting feature set from every key frame, they are concatenated and are formulated into a 1-D feature vector. Further to obtain the only informative features from all the features, Principal Component Analysis (PCA) is accomplished to reduce the high dimensional feature vector into a low dimensional feature vector by keeping 99% principal components.

In this phase, totally three types of features are extracted for every frame. Given a frame, we need to represent it with and effective feature set or we need to describe it with effective features. Here the proposed feature descriptor is intended to capture intensity, orientation and contour information. These three features are more important and with these features, a moving object or a human action can be discriminated effectively. To extract these three features for every frame, three different techniques such as Gaussian Pyramid, Gabor Filter and Wavelet Pyramid are applied [36].

#### 3.2.1. Intensity features

The main objective of the intensity feature is to provide sufficient discriminative information for action recognition system through the intensities of different human actions of same or different views. In this phase, Gaussian filter is applied to for intensity features extraction. To extract the intensity features from every key frame, initially we construct a seven level Gaussian pyramid and at each level the

frame is convolved with Gaussian filter with variance $\sigma = 2$. At first level of Gaussian pyramid, the frame is convolved with Gaussian filter as it is and at the second level, the same Gaussian filter is convolved with a down sampled frame [34]. Similarly, in the same manner, for an increasing level of Gaussian Pyramid, the frame is down sampled and convolved with Gaussian filter. The mathematical representation of Gaussian Pyramid is shown as

$$W(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{x^2+y^2}{2\sigma^2}\right)} \quad (3)$$

$$G_l(i,j) = \sum_{x=1}^{X}\sum_{y=1}^{Y} W(x,y) G_{l-1}(2i+x, 2j+y) \quad (4)$$

Where $l$ denotes the level of Gaussian Pyramid, and $(i,j)$ represents the co-ordinates of the pixel in the frame.

After this process, the Gaussian based intensity features are extracted by the accomplishment of Feature-by-feature subtraction of initial and final levels of Gaussian Pyramid. The initial and final levels are linked by $Final\ Level = \max level - initial\ level$. For example, if we consider the initial level as *2*, the final level will be *7-2=5*, so level *5* is final level. In this case, the Gaussian Feature map is obtained by the subtraction of Level *2* and Level *5*.

**Note:** for a subtraction process, the sizes of two matrices must be same but here the sizes of initial and final frames won't be same because, as the Gaussian pyramid increases, the size of frame decreases gradually due to down sampling. Hence to obtain the same size of final level, it is interpolated into the size of the frame at initial level and then subtracted [36]. A simple representation about the Gaussian Pyramid feature map is shown in Fig. 4, below.

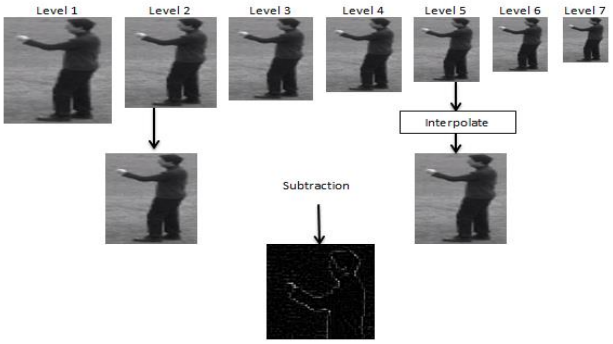


Figure.4 Gaussian pyramid feature map extraction

In this work, we considered totally three maps obtained by the subtraction of final levels such as level 6, level 5 and level 4 from initial levels such as level 1, level, 2 and level 3 respectively.

### 3.2.2. Orientational features

Orientational features play an important role in the scene classification. Based on these features, the HAR system will become robust to scale and translation invariance. These features are effective in the provision of a proper discrimination between the actions captured under multiple views. Particularly these features are very much helpful in the MVHAR system.

Towards Orientational features extraction, this work accomplished Gabor Filter due to its effectiveness in the feature extraction in different orientations. Since the Gabor filter extracts the features which are scale- and orientation-invariant, this paper considered it for Orientational features extraction. Here the Gabor filter is accomplished in different scales such as $5 \times 5$, $7 \times 7$, $9 \times 9$, and $11 \times 11$, and eight orientations such as $0^0$, $45^0$, $90^0$, $135^0$, $180^0$, $225^0$, $270^0$, and $315^0$. So totally for each frame, we will get $4 \times 8 = 32$ feature maps. The mathematical formula for Gabor filter is shown as

$$G(x,y) = exp\left(\frac{X^2+\gamma Y^2}{2\sigma^2}\right) \cos\left(\frac{2\pi}{\lambda}x\right) \quad (5)$$

Where

$$X = xcos\theta - ysin\theta,\ \ Y = xsin\theta + ycos\theta \quad (6)$$

Where $(x,y)$ is position relative to the center of filter. A simple representation of Gabor Filter accomplishment over a frame is depicted in the following Fig. 5, below.

However the obtained 32 feature maps are high in number and create much computational burden.

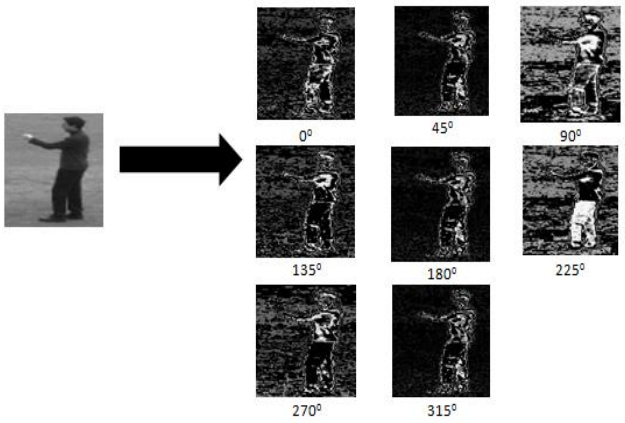


Figure.5 Gabor filter accomplishment

Hence, to reduce this computational burden, here a max pooling mechanism [37] is applied over the 32 feature maps of every frame. In other words, we will pick a maximum value from all feature maps with filter scale in each orientation. The max pooling in different scales is performed as

$$F_{max} = \max_{(x,y,\theta_s)} \begin{bmatrix} F_{5\times5}(x,y,\theta_s), F_{7\times7}(x,y,\theta_s), \\ \dots\ F_{11\times11}(x,y,\theta_s) \end{bmatrix} \quad (7)$$

Where $F_{max}$ is the maximum feature map obtained through the max-pooling, $F_{k\times k}(x,y,\theta_s)$ is the feature map at $k \times k$ scale and at $\theta_s$ orientation. In this manner, we will get totally eight feature maps, one from each orientation.

#### 3.2.3. Contour features

The contour features are more important in determining the pose of Human action. A complete and effective contour itself provides efficient information about the Human action. Here the Wavelet Transform is accomplished to obtain the contour feature from every frame. Generally, for a frame/image, a 2-D wavelet transform decomposed it into four sub bands such as Approximations *(CA)*, Horizontals *(CH)*, Verticals *(CV)* and Diagonal Details *(CD)* [35]. Here the wavelet Transform itself down samples the input image to its half the size and hence the obtained Sub bands are half of the size of an input from which they derived. Based on this advantage, this work accomplished a five level decomposition. Hence totally we will get 20 sub bands, four bands from each level of decomposition. Here only the approximations are processed for further decomposition.

Next, similar to the Gaussian Pyramid feature map extraction process, here also the subtraction is accomplished at adjacent levels but only between approximation sub bands. To further compute the feature map, we evaluate the difference between two each wavelet approximation sub bands. Here also the adjacent levels have no same size to perform subtraction process. Hence the approximation band at high level of wavelet pyramid is interpolated to the size of an approximation band at low level. Simply $f_i = S_i - S_{i+1}$ , where $S_i$ is the approximation band at $i_{th}$ level and $S_{i+1}$ is the approximation band at $i + 1_{th}$ the level.

Consider an example frame of size $2048 \times 2048$ . If this frame was subjected to wavelet decomposition, the obtained sub bands are having the size of $1024 \times 1024$. Here in the second level of wavelet transform, an approximation band (CA) of size $1024 \times 1024$ is processed as input and the obtained output sub bands are of size $512 \times 512$. This continues up to five levels and then pyramid feature map is constructed by point-to-point subtraction of adjacent levels. A simple representation of wavelet pyramid feature map construction is shown in Fig. 6, below.

Finally the obtained feature maps through Gaussian, Gabor Filter and Wavelet pyramid are formulated into a 1-D feature vector. This 1-D feature vectors are extracted for every frame and these features are very much helpful in the provision of much discrimination between different human actions. Furthermore, the Gaussian Pyramid and Wavelet Pyramid explore the multi-resolution property of image, and then the proposed approach achieves invariance to scaling.

### 3.3 Action recognition

In the third phase, i.e., action recognition, the obtained feature set is processed for action recognition. Before this, the system is trained through the Support Vector Machine algorithm. Initially, the HAR system is trained with database videos having different actions acquired at multiple views. The same feature extraction process is applied at the training phase also to overcome the memory constraint of database. For every training action, initially a set of key frames are extracted based on the Key frames selection (Section 3.1) and further for the obtained key frames, the three set of features (Section 3.2) are extracted and then processed for PCA for dimensionality reduction. Further obtained principal components of every training action are trained through SVM algorithm. Next, in the testing phase, the videos of same action but captured under multiple views are tested one by one and here the SVM classifier is accomplished for classification. The SVM classifier classifies the input test action into one of the human action types and produces the label to which it belongs to. For


Figure.6 Wavelet Laplacian pyramid feature map

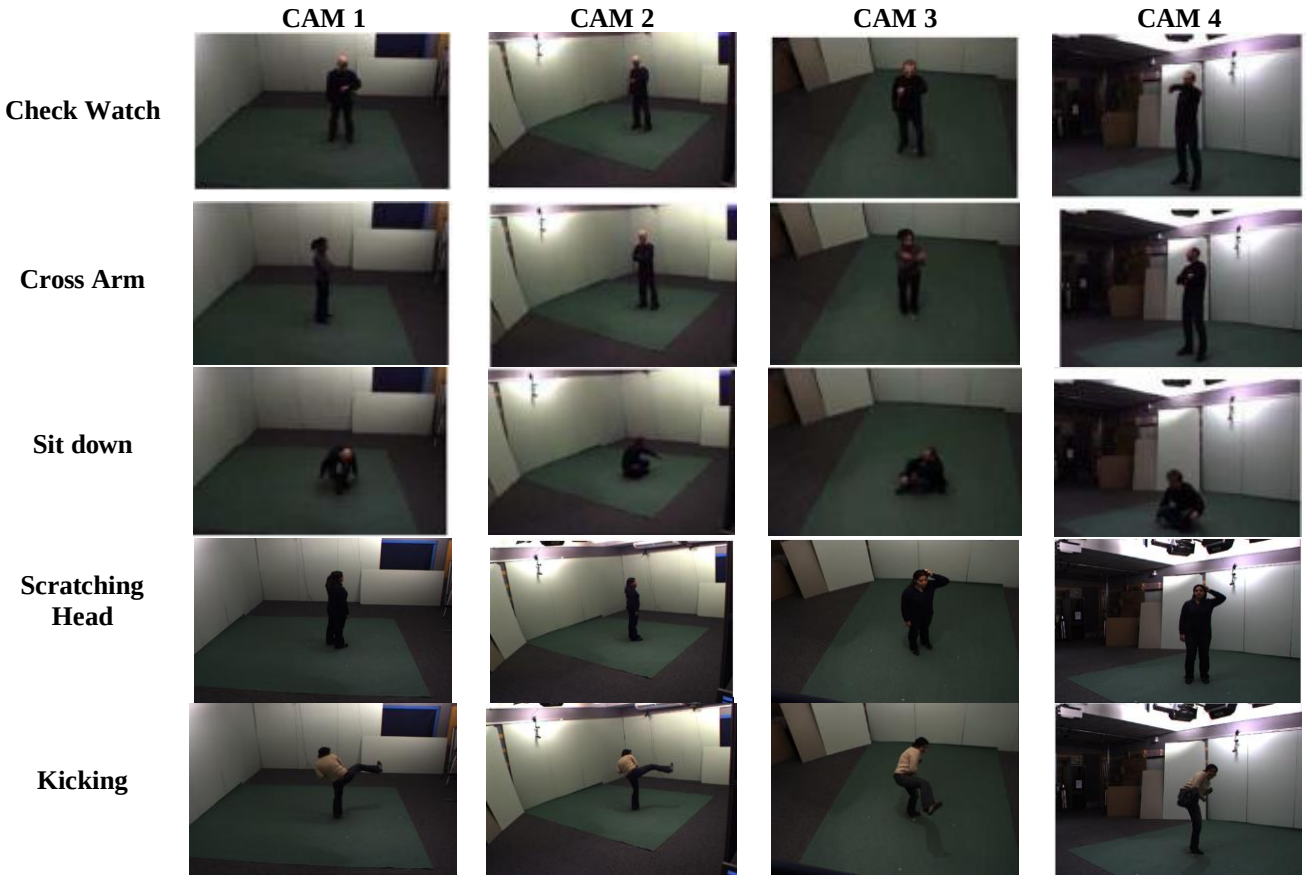


Figure.7 Few samples of IXMAS dataset

SVM classification, we train non-linear SVMs using $\chi^2$ kernel and adopt one-against all approach for multiclass classification.

## 4. Simulation results

This section describes the details of simulation experiments conducted over the proposed recognition model. Under this section, a standard database is considered for simulation. After simulation, the performance is measured through the performance metrics and then a comparative analysis is conducted to alleviate the performance effectiveness of proposed approach in the recognition of different actions. To simulate the developed model, MATLAB2014a software is used. Initially the training process is performed through different videos having different action sequences and also in different views. After the completion of training, testing is performed through different actions sequences and with different views.

### 4.1 Dataset details

For the evaluation of proposed system, comprehensive experiments are conducted over the well-known multi-view INRIA Xmas Motion Acquisition Sequences (IXMAS) dataset. IXMAS is a challenging dataset, acquired with multiple actors under multiple camera views. This dataset is more popular among the HAR methods for testing view independent action recognition algorithms, including both cross-view and multi-view action recognition. This dataset consists of 12 action classes such as *check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point and pick up*. Each action is performed three times and 12 different subjects are recorded with five cameras, four are fixed at four sides and one is fixe on the top. These five cameras capture five views such as left, right front back and top. The frame rate is 23 frames per second and the size of frame is $390 \times 291$ pixels. Fig. 7 shows some samples of different actions under multiple views. Each row represents different action and each column represents different views.

### 4.2 Results

In this paper, the performance metrics namely Accuracy, Precision, Detection Rate or Recall, False Negative Rate (FNR) and F-Score are considered to evaluate the performance of proposed approach. After testing different actions with different view,

Table 1. Performance metrics for different actions

| Action/Metric | Recall (%) | Precision (%) | F-Score (%) | FNR (%) |
|---|---|---|---|---|
| Check Watch | 93.4574 | 93.5532 | 93.5053 | 6.5423 |
| Cross arms | 93.6145 | 93.8454 | 93.6510 | 6.3855 |
| Scratch head | 92.8741 | 92.9699 | 92.9220 | 7.1259 |
| Sit down | 90.4785 | 90.5743 | 90.5264 | 9.5215 |
| Get up | 91.0025 | 91.0983 | 91.0504 | 8.9975 |
| Turn around | 89.3658 | 89.4616 | 89.4137 | 10.6342 |
| Walk | 92.4314 | 92.5272 | 92.4793 | 7.5686 |
| Wave | 93.4647 | 93.5605 | 93.5126 | 6.5353 |
| Punch | 87.4571 | 87.7785 | 87.6175 | 12.5429 |
| Kick | 88.7496 | 88.8954 | 88.8224 | 11.2504 |
| Point | 89.4963 | 90.1124 | 89.8033 | 10.5037 |
| Pick up | 90.1247 | 90.2247 | 90.1747 | 9.8753 |

Table 2. Performance metrics under different views

| CAM/Metric | Recall (%) | Precision (%) | F-Score (%) | FNR (%) |
|---|---|---|---|---|
| CAM 1 | 90.8606 | 91.2354 | 90.5481 | 9.1394 |
| CAM 2 | 89.2239 | 90.2147 | 88.9114 | 10.7761 |
| CAM 3 | 90.2895 | 89.2421 | 89.9770 | 9.7105 |
| CAM 4 | 89.3228 | 90.4538 | 89.0103 | 10.6772 |

the obtained results are formulated into a confusion matrix. Depends on the obtained numerical results, True Positives (TPs), True Negatives (TNs), False Positives (FPs), and False Negatives (FNs) are measured. Based on the obtained TP, TN, FP and FN values from the confusion matrix, performance metrics are evaluated and the respective mathematical representation is given as;

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

$$F-Score = \frac{2\times Recall\times Precision}{Recall+Precision} \tag{10}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

$$False\ Negative\ Rate = \frac{FN}{TP+FN} \tag{12}$$

Here totally 12 actions are considered for every actor under different views. After the simulation of different actions sequences through the proposed approach, the obtained results are represented as above.

Table 1 depicts the details of performance metrics evaluation under different actions. Here, all types of actions of the IXMAS dataset are processed for simulation. For every action, the developed system displays a label to which it belongs. Based on the label, the correctly classified results are measured and they are called True Positives and the incorrectly classified results are called True Negatives. For example, if the action sequence of ***check watch*** is processed for testing and the system had displayed a label of Scratch Head, then it is counted under True negative. In this manner, for every action, the total number of positively and negatively classified results are measured and upon the substitution of these value in to the Eq. (8), we obtained the Detection Rate (Recall). Similarly, the further metrics are also measured for every action.

Next, the performance evaluation is done under different views, i.e., all actions captured under different views such as CAM 1, CAM 2, CAM 3 and CAM 4 are processed for testing and the obtained Recall, Precision, F-Score and FNR are shown in Table 2. In this testing process, the system displays the output label as CAM 1, CAM 2, CAM 3 and CAM 4. Under this simulation, for a given test action sequence belongs to CAM 1, if the system shows the label as CAM 1, then it is considered as True Positive otherwise True Negative.

Further the proposed approach is compared with the conventional approaches such as Chun et al. [20] and Sargano et al. [33] through the obtained performance metrics such as Recall, Precision and Accuracy. Initially the comparison is carried out between proposed and conventional approaches through different views. Further the comparison is accomplished through different actions. Figs.8, 9, and 10 describe the comparison details between proposed and conventional approaches through recall, precision and accuracy respectively.

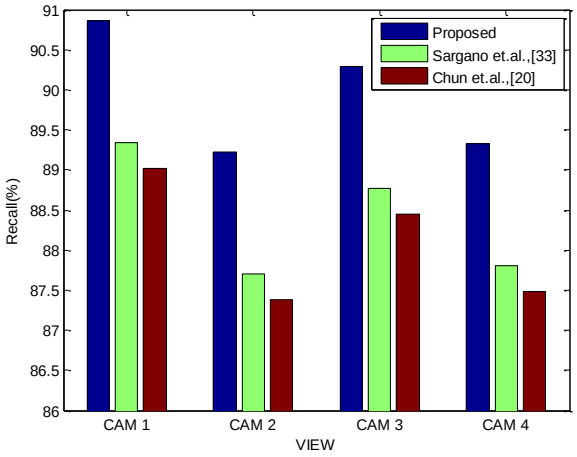


Figure.8 Recall comparison under different views

As it can be seen from the above Fig. 8, the Recall of proposed approach is high compared to the conventional approaches in all views. Since the proposed approach extracted and trained a view-invariant features which have robustness in all views, the developed system classified most of the action sequences correctly. A higher value of TP results in higher recall rate and it can be observed from the above figure. On an average, the recall of proposed approach is obtained as 89.9242% and for the conventional approaches it is 88.2037% and 88.0767% for Sargano et al. [33] and Chun et al. [20] respectively. Since the conventional approaches not focused much on the Orientational features and scale-invariant features which has a significant effect in the MVHAR, they are not able to detect the all the actions perfectly. And also the conventional approaches focused only one type of features like Sargano et al. considered Human silhouettes and Chun et al. considered Histogram of Motion Intensity, which are not able to provide a perfect discrimination between multiple views of a single action. Further the conventional approaches also not concentrated on the key feature selection which is most impact in the proposed approach.

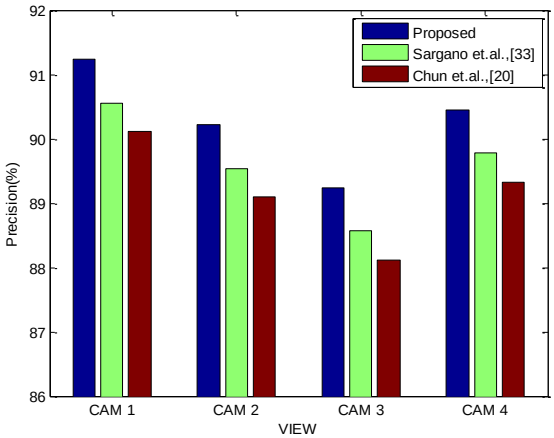


Figure.9 Precision comparison under different views

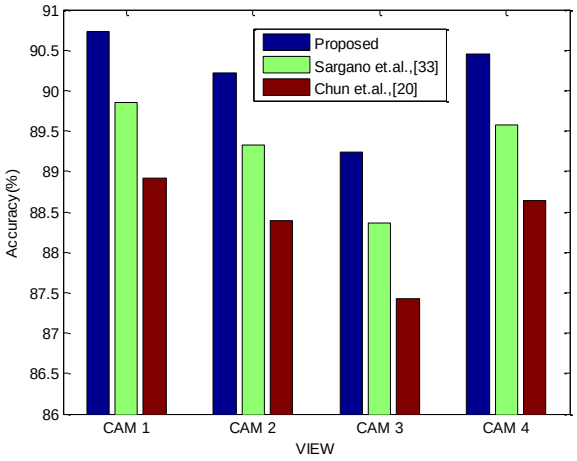


Figure.10 Accuracy comparison under different views

As it can be seen from the above Fig. 9, the Precision of proposed approach is high compared to the conventional approaches in all views. A higher precision value indicates more precise classification. It also denotes less false positives, means for a given non-required input action sequence, the output will be non-required only but not required. A higher value of false positives will result a lower precision. In the above figure, the precision is measured with respect to views, i.e., the actions under different views are tested and here the output is view only. For a given action sequence belongs to CAM 1, the output may or may not be CAM 1. A perfect classification will increase the TP otherwise it increases FP. On an average, the Precision of proposed approach is obtained as 90.2865% and for the conventional approaches it is 88.7408% and 88.6432% for Sargano et al. [33] and Chun et al. [20] respectively. Due to the extraction of key frames followed by view-invariant features from every key frame, the proposed system can classify the all views more precisely. A more precise classification requires more knowledge about the movement features of an action. In the proposed approach, this provision is done through the extraction of multiple features such as Intensity, Orientational and Contour, whereas the conventional Sargano et al. focused only on the contours and Chun et al. focused only on the histogram features.

Fig. 10 reveals the Accuracy details of proposed and conventional approaches under different views. As it can be observed from above figure, the accuracy of proposed approach is high compared to the conventional approaches. Since the proposed approach is extracted scale-invariant features through Gaussian Pyramid and Orientational-invariant features through Gabor filter, for any action sequence with different orientation and

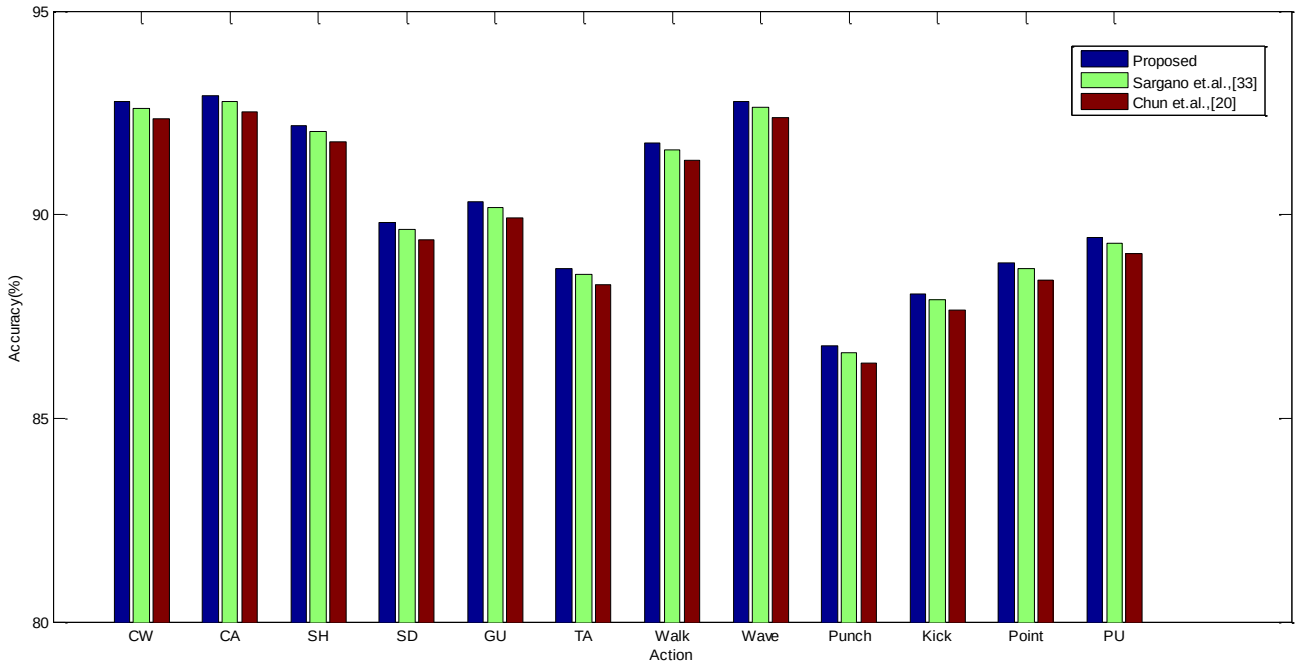


Figure.11 Accuracy comparison under different Actions

scaling, the proposed system can effectively match with the key features trained to it and produces a proper output, i.e., label to which view the action sequence belongs. Due to this property the proposed approach can classify any action sequence acquired in view with more accuracy, while the conventional approaches have more false positives which leads to lower accuracy. On an average, the Accuracy of proposed approach is obtained as 90.1615% and for the conventional approaches it is 89.2767% and 88.3427% for Sargano et al. [33] and Chun et al. [20] respectively.

The further comparison is done with respect to action sequences. In the above Fig. 11, the accuracy of proposed and conventional approaches is compared under the classification of different action sequences. A short form representation is presented under X-Label of above figure for every action like *CW – Check Watch, CA-Cross Arms, SH-Scratch Head, SD-Sit Down, GU-Get Up, TA-Turn Around, Walk, Wave, Punch, Kick, Point and PU-Pick Up*. This figure deals with an in details performance analysis through accuracy at action level. A sit can be seen from the above figure, the highest accuracy is observed at Waving Action and lowest is observed at Punch action. For Check watch and Cross arms actions, the accuracy is almost equal because they have almost similar contour pattern. For further action classes also, the proposed approach obtained a significant accuracy and it is high compared to the conventional approaches. Due to the provision a perfect discrimination between the movements of an action, the proposed approach has obtained better results even in the actions with similar characteristics. For instance, when two actions namely CW and CA are considered, they have similar hand movements and the sensitive difference between these two actions need to be captured and this is achieved through the proposed approach due to the multi-feature set analysis. In the case of Sargano et al., the human silhouettes cannot provide a proper difference between these two actions by which the accuracy is less. On an average, the Accuracy of proposed approach is obtained as 91.0431% and for the conventional approaches it is 89.7500% and 83.0300% for Sargano et al. [33] and Chun et al. [20] respectively.

## 5. Conclusion

A novel Multi-View Human Action Recognition system is developed in this paper which is robust different multiple views. With an aim of optimal feature set extraction from every action sequence, this paper proposed new feature extraction approach combining with Gaussian Features, Gabor Features and Wavelet Features. Furthermore, this paper also proposed to select only key frames which have more significant information about the actions through self-similarity matrix accomplishment. Extensive simulations carried out over the proposed approach through different actions acquired under different views had shown an outstanding performance. The performance analysis accomplished through the performance metrics such as Accuracy, Recall, and Precision reveals the effectiveness in the precise

classification of any type of action under any view. Compared to the conventional approaches, the proposed model has better performance in the classification of multi-view human actions. On an average the proposed approach gained an improvement in the accuracy is of 1.2931% and 8.0131% from conventional approaches Sargano et al. [33] and Chun et al. [20] respectively.

The further scope of this paper is towards the development of a new Feature Descriptor Method through which major obstacles like Occlusions and Background Movements can be neutralized. Furthermore, a new frame selection technique which can able to derive both linear and non-linear relationships between action sequences is intended to develop.

## References


[1] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse Spatio-temporal features", In: *Proc. of IEEE Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65–72, 2005.

[2] T. Ko, "A survey on behavior analysis in video surveillance for homeland security applications", In: *Proc. of the 37th IEEE Applied Imagery Pattern Recognition Workshop*, pp. 1–8, 2008.

[3] A. Bobick and J. 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.

[4] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos", In: *Proc. of International Conf. on Computer Vision*, pp.1-5, 2011.

[5] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation", In *Proc. of the IEEE Conf. on Computer Vision and. Pattern Recognition*, pp. 984–989, 2005.

[6] K. Guo, P. Ishwar, and J. Konrad, "Action recognition from video using feature covariance matrices", *IEEE Transactions on Image Processing*, Vol. 22, No. 6, pp. 2479–2494, 2013.

[7] D. Weinland, M. Özuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes", In: *Proc. of the 11th European Conf. on Computer Vision*, pp. 635–648, 2010.

[8] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies", In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[9] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as Space-Time Shapes", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 12, pp. 2247-2253, 2007.

[10] Z. Lin, Z. Jiang, and L. S. Davis, "Recognizing actions by shape motion prototype trees", In: *Proc. of the 12th IEEE International Conf. on Computer Vision*, pp. 444–451, 2009.

[11] I. N. Junejo, E. Dexter, I. Laptev, and P. Perez, "View-independent action recognition from temporal self-similarities", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 33, No. 1, pp. 172–185, 2011.

[12] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer", In: *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3209–3216, 2011.

[13] J. Zheng and Z. Jiang, "Learning view-invariant sparse representations for cross-view action recognition", In: *Proc. of IEEE International Conf. on Computer Vision*, pp. 3176–3183, 2013.

[14] A. Farhadi and M. K. Tabrizi, "Learning to recognize activities from the wrong view point", In: *Proc. of European Conf. on Computer Vision*, pp. 154–166, 2008.

[15] Z. Gao, H. Zhang, G. P. Xu, Y. B. Xue, and A. G. Hauptmann, "Multi-view discriminative and structured dictionary learning with group sparsity for human action recognition", *Signal Processing*, Vol. 112, No.1, pp. 83-97, 2015.

[16] F. Zho, L. Shao, and M. Lin, "Multi-view action recognition using local similarity random forests and sensor fusion", *Pattern Recognition Letters*, Vol. 34, No. 1, pp. 20-24, 2013.

[17] Z. Gao, Y. Zhang, H. Zhang, Y. B. Xue, and G. P. Xu, "Multi-dimensional human action recognition model based on image set and group sparisty", *Neurocompting*, Vol. 215, No. 26, pp. 138-149, 2016.

[18] A. Klaser and M. Marszalek, "A Spatio-temporal descriptor based on 3D gradients", In: *Proc. of International Conf. on British Machine Vision*, pp. 995–1004, 2005.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", In: *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 886–893, 2005.

[20] S. Chun and C. S. Lee, "Human action recognition using histogram of motion intensity

and direction from multiple views", *IET Computer Vision*, Vol. 10, No.4, pp. 250–257, 2016.

[21] F. Murtaza, M. H. Yousaf, and S. Velastin, "Multi-view Human Action Recognition using 2D Motion Templates based on MHIs and their HOG Description", *IET Computer Vision*, Vol. 10, No.7, pp. 758-767, 2016.

[22] Z. Gao, W. Nie, A. Liu, and H. Zhang, "Evaluation of local spatial–temporal features for cross-view action recognition", *Neurocomputing*, Vol. 173, No. 15, pp.110-117, 2016.

[23] I. Laptev, "On space-time interest points", *International Journal of Computer Vision*, Vol. 64, No.2, pp. 107–123, 2005.

[24] M. Chen and A. Hauptmann, "MoSIFT: Recognizing human actions in surveillance videos", Technical report, *Carnegie Mellon University*, Pittsburgh, USA, pp.1-16, 2009.

[25] A. A. Chaaraoui, P. C. Pérez, and F. F. Revuelta, "Silhouette-based human action recognition using sequences of key poses", *Pattern Recognition Letters*, Vol. 34, No.1, pp. 1799–1807, 2013.

[26] A. A Chaaraoui and F. F. Revuelta, "A Low-Dimensional Radial Silhouette Based Feature for Fast Human Action Recognition Fusing Multiple Views", *International Scholarly Research Notices*, Vol. 2014, pp.1-11, 2014.

[27] A. K. S Kushwaha, S. Srivastava, and R. Srivastava, "Multi-view human activity recognition based on silhouette and uniform rotation invariant local binary patterns", *Multimedia Systems*, Vol.23, No.4, pp.451-467, 2016.

[28] S. Pehlivan and D. A. Forsyth, "Recognizing activities in multiple views with fusion of frame judgments", *Image and Vision Computing*, Vol. 32, No.4, pp. 237–249, 2014.

[29] N. Imran Junejo, E. Dexter, I. Laptev, and P. Perez, "View-Independent Action Recognition from Temporal Self-Similarities", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 33, No. 1, pp.172-185, 2011.

[30] J. Wang and H. Zheng, "View-robust action recognition based on temporal self-similarities and dynamic time warping", In: *Proc. of IEEE International Conf. on Computer Science and Automation Engineering (CSAE)*, pp.1-5, 2012.

[31] A. Aryanfar, R. Yakob, and A. A. Halin, "Multi-View Human Action recognition Using Wavelet data Reduction and Multi-class Classification", In: *Prof. of International Conf. on Soft Computing and Software Engineering*, pp.585-592, 2015.

[32] K. P. Chou, M. Prasad, D. Wu, N. Sharma, D. L. Li, Y. F. Lin, M. Blumenstein, W. C. Lin, and C. T Lin, "Robust Feature-Based Automated Multi-View Human Action Recognition System", *IEEE Access*, Vol. 6, pp.15283-15296, 2018.

[33] A. B. Sargano, P. Angelov and Z. Habib, "Human Action Recognition from Multiple Views Based on View-Invariant Feature Descriptor Using Support Vector Machines", *Applied Sciences*, Vol.6, No.309, 2016.

[34] C. Yang, M. Schmalz, W. Hu, and G. Ritter, "Center-surround filters for the detection of small targets in cluttered multispectral imagery: Background and filter design", *Detection Technologies for Mines and Minelike Targets*, Vol. 2496, pp. 637–648.

[35] A. Cohen, I. Daubechies, and J. Feauveau, "Bi-orthogonal bases of compactly supported wavelets", *Communications and Pure Applied Mathematics.*, Vol. 45, No. 5, pp. 485–560, 1992.

[36] L. Liu, L. Shao, X. Zhen, and X. Li, "Learning Discriminative Key Poses for Action recognition", *IEEE Transactions on Cybernetics*, Vol. 43, No. 6, pp.1860-1870, 2013

[37] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex", *Natural Neuroscience*, Vol.2, pp. 1019–1025, 1999.