JOURNAL OF MECHANICS OF CONTINUA AND
MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-15, No.-1, January (2020) pp 128-146 ISSN (Print) 0973-8975

VIEW-ROBUST HUMAN ACTION RECOGNITION BASED ON SPATIO-TEMPORAL SELF SIMILARITIES

K. Pradeep Reddy[1], G. Apparao Naidu[2], B Vishnu Vardhan[3]

[1]Research Scholar, Department of Computer Science Engineering, Tirumala Engineering College, Telangana, India.

[2]Professor, Department of Computer Science Engineering, JB Institute of Engineering and Technology, Telangana, India.

[3]Professor, Department of Computer Science Engineering, Jawaharlal Nehru Technological University College of Engineering, Manthani, Telangana, India.

Email: [1]kumbalapredeepreddy@gmail.com , [2]apparaonaidug@gmail.com , [3]mailvishnu@yahoo.com

https://doi.org/10.26782/jmcms.2020.01.00010

Abstract

Multi-View Human Action Recognition, as a hot research area in computer vision, has many more applications in various fields. Despite its popularity, more precise recognition still remains a major challenge due to various constraints. Extracting the robust and discriminative feature from video sequence is a crucial step in the Human Action Recognition system. In this paper, a new feature extraction technique is proposed based on the integration of three different features such as intensity, Orientation and Contour features. Unlike the earlier approaches which applied feature extraction directly over actions videos, this approach applies the feature extraction only over key frames which are extracted from a large set of frames. The key frames selection is accomplished based on a new mechanism, called Gradient Self-Similarity Matrix (GSSM). GSSM is proposed as an extension to the most popular Self-Similarity Matrix (SSM) by evaluating the gradients of actions frames before SSM accomplishment. Once the key frames are extracted, the hybrid feature extraction mechanism is applied and the obtained features are processed for classification through Support Vector Machine Classifier. The proposed framework is systematically evaluated on IXMAS dataset and NIXMAS dataset. Experimental results enumerate that our method outperforms the conventional techniques in terms of recognition accuracy.

Keywords: Computer Vision, Human Action Recognition, Multiple Views, Self-Similarity Matrix, Gaussian, Gabor, Wavelet, Accuracy.

I. Introduction

Automatic recognition of human actions has gained significant research interest in computer vision during past few years. The ever growing interest in the human action characterization is in part due to the increasing number of real world applications such as Human computer interaction (HCI), video annotation, smart home system, activity monitoring in surveillance environments and intelligent video surveillance [XXIX] etc. For example, detection of falling actions of older aged people is very important and it is carried out through smart home system. Further, the detection of abnormal behavior of humans in time is a most important in intelligent video surveillance systems.In general, major of the earlier developed Human Action Recognition (HAR) techniques considered the single view of human action for recognition [XXXIV, XIX, III, XXVI]. In these methods, the test video is of only a single sided view. However, the real world videos have so many challenges towards single view HAR, because, the visual appearance of actions are greatly affected by self-occlusion and view point changes.Hence there is a need to consider multiple views for an HAR, which is called as Multi-View HAR (MVHAR) [XXXVII].

In MVHAR, the recognition of human action is accomplished based on the multiple views. Initially, the system is trained after the extraction of a set of features from every action. These set of feature are view invariant and helps in the recognition of action under any view. The main advantage of MVHAR is view invariant recognition under multiple view points and also under self-occluded scenarios. Since self-occlusion problems can be handed by deploying the multiple cameras, MVHAR methods are more robust than the Single View HAR methods. But, the action recognition is generally accomplished through the motion trajectories with respect to the camera view point, the changes in view point has a significant impact on the action recognition [VII, IV]. Hence, the extraction of view invariant features is an important task in the MVHAR system.

To achieve an efficient action recognition performance under multiple view-points, this paper proposed a novel HAR framework. In this framework, to extract key framesof action under multiple views, a new variant of SSM, called GSSM is accomplished. Further to extract view invariant features of an action, this paper extracted three different features such as intensity features, orientational features, and contour features. Finally Principal Component analysis is accomplished to reduce the dimensionality of obtained feature vector. Simulation is conducted over two standard benchmark datasets, namely IXMAS and NIXMAS to test the performance of proposed framework.

Rest of the paper is organized as follows; Section II reveals the details of literature survey. Section III reveals the complete details of prosed framework. Section IV discusses the details of simulation results and finally the concluding remarks are discussed in section V.

II. Literature Survey

Several approaches are proposed in earlier for HAR. In the HAR system, action representation and feature extraction are most important tasks and most of the earlier approaches focused in that direction only. Based on the methodology followed for action representation, the earlier approaches are broadly classified as Trajectory based approaches [XII, XV, XX], Space-Time Shapes [XXII, XXVIII], 3D Patch Analysis [V] and Silhouettes [XXIII]. Further, the feature extraction approaches are classified as local features or Space-Time Interest Points (STIPs) [XVI, XX] based, appearance based [XI, XXXVIII], and motion based [XXVII].Among these, STIPs based methods are more effective which ensures robust action recognition even under camera movements, low resolution inputs and also under noisy inputs. However, these methods assume that the individual spat-time descriptors provide sufficient discrimination capability to classify different actions; however, they ignored the information related to the global Spatio-Temporal distribution. Thus, using the STIPs,smooth motions are not captured due to the lack of temporal information.Next, in the appearance based action recognition, the appearance of a test action is not match with the appearance of training actions. Generally, the motion based approaches considered the optical flow vectors as features for HAR.However, these methods are more prone to noise variations which consequences to false recognition.

Recently, the SSM basedaction recognition has gained a significant recognition results. Imran Junejo et al. [XIV] introduced the concept of SSM. In [XIV], every action is represented with a set of SSMs. In this approach, initially, the action video is represented with low level features. Further the SSM is constructed based on the evaluation of Euclidean distance between the extracted features of all frames in a pairwise fashion. E. Shechtman and M. Irani [XI] introduced a new image/video matching technique based on the internal self-similarities. This approach assumes that the images / videos are similar in their internal layouts even though the patterns generating those similarities are different. These internal self-similarities are effectively captured by Local Self-Similarity (LSS) descriptor. The LSS is applied densely throughout the video at multiple scales and extracted the matching entities even under various geometric distortions.Further, one more method is proposed by Stark et al. [XXV] in which the shape is considered as a local object and the similarity is accomplished between the shapes. Some more authors focused to combine the Bag of Visual Words (BoVW) with LSS and introduced a new similarity metric called as, bag-of-local-self-similarities (BOLSS) [XVIII, VI].However, the LSS is never capture the global similarities in the entire image by which the image matching will be less effective.Inspired with the LSS, I. Junejo et al. [XIII] applied the local self-similarity surfaces for action recognition. These surfaces are constructed by performing the matching between patches, centered at a pixel. Once the surfaces are computed, [XIII] it was proposed to transform these surfaces into Histogram of Gradients (HoG) and then processed for training through Conditional Random Fields (CRFS).

To overcome the LSS problem, T. Deselaers and V. Ferrari [XXXVI] proposed the Global Self-Similarity (GSS)and explored its advantages over LSS.This captures the spatial arrangements of self-similarities within the entire image. This approach also had shown the effective utilization of GSS based descriptors to detect the objects Branch and Bound Framework and also in a sliding window framework. Furthermore, [XXXVI] also introduced two different global descriptors, namely, self-similarity hyper cubes (SSH) and bag-of-correlation surfaces (BOCS). Experimental validation is carried out on the Pascal VOC 2007 dataset [XXIV] and shown a better classification performance than the LSS based descriptors. However, the major drawback of GSS is that it is very expensive to compute if done directly.Jing Wang et al. [XXVI] proposed a new HAR method based on SSM and Dynamic Time Warping (DTW). Here the SSM captured the Global Time information which was useful in the action recognition under viewpoints. DTW is applied further for the full pledged utilization of SSM information. K-Nearest Neighbor Classifier (KNNC) is accomplished for classifications.

SamySadek et al. [XXX] proposed a novel HAR framework based on temporal self-similarities and fuzzy log-polar histograms. Initially, in this approach, the reliable key points, i.e., action snippet is extracted. Next, the local temporal self-similarities are accomplished to extract the descriptors based on the fuzzy log-polar histograms and finally the SVM classifier is accomplished for action recognition. Considering the RGBD video for action recognition, Y P Hsu et al. [XXXIV] proposed to construct a Spatio-temporal matrix (STM) based on the Euclidean distance between Spatio temporal feature vectors. Further, to recognize the action, this approach described the local tendency of the STM using pyramid-structural BoW (BoW-pyramid) and SVM classifier is trained for classification.

K P Chou et al. [XXI] proposed a novel MVHAR framework based on the Gaussian Mixture Modelling (GMM) and Gabor filter. Here the GMM is accomplished for background subtraction and Gabor filter is applied for interest point extraction. Three different classifiers namely, Nearest Mean Classifier (NMC), GMM classifier (GMMC) and Nearest Neighbor Classifier (NNC) and three different datasets namely, Weizmann [XXII], KTH [VIII] and MUHAVI [XXII] are considered for classification and simulation respectively. S. Karungaru et al. [XXXIII] proposed a new method to recognize multi-view actions captured in an indoor work environment. Histogram of Gradients (HOG) features are extracted from every action view ad they are learned through AdaBoost classifier. As a preprocessing stage, this method accomplished to measure the distance between detected area in successive frames to recognize a mobile or a stationary person. Several fuzzy rules are applied to detect the human action based on the height of person and the direction he/she is facing.

Next, considering the advantages of wavelet transform, A. A et al. [I] 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.

III. Proposed HAR System

Overview of Framework

The overview of the proposed framework is shown in Figure.1. The overall framework is accomplished in two phases they are training and testing. In the training phase, a larger number of actions are trained to the HAR system and in the testing phase, the HAR system is tested by giving different actions as inputs.

For a given inputaction sequence the proposed system finds key frames based on SSM. Next, from the obtained frames, three different set of features are extracted. The novelty of the proposed HAR system is key frames extraction and it is done through a new SSM, called gradient SSM.Similar to our first contribution [XXI], this paper also focused on three set of features such as intensity features, Orientational features and contour features. The main objective behind the consideration of three features is to achieve a better recognition accuracy under all possible (both ideal and real time) constraints. The intensity feature set provides more information about the basic variations in actions through the pixel intensities. Next, the orientational feature set gives more information about the scale and rotation variances. Since the proposed system is focused on multiple views, the rotation invariant feature is more important and the Gabor filter is an efficient filter in such provision. Next, the contour features are also important and they provide information about the trajectory variations at different scales. After extracting individual features, a composite feature vector is formulated by concatenating all these features. Once the features are extracted from test action video, they are subjected to classification through Classifier.

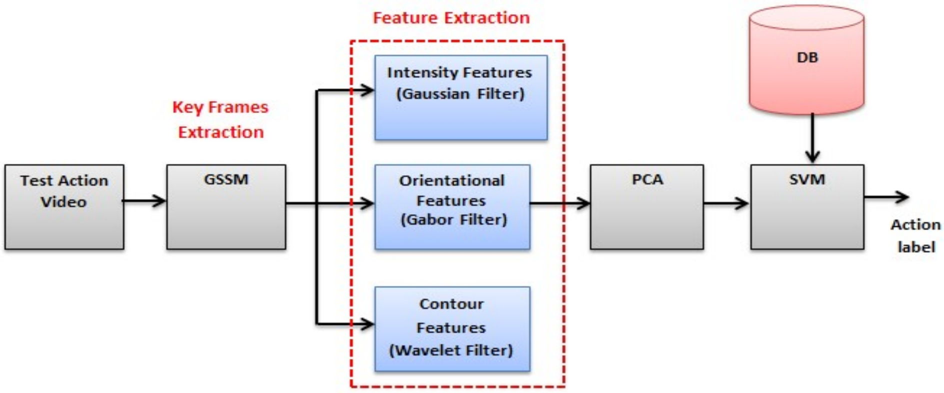

Feature Extraction
DB
Key Frames Extraction
Intensity Features (Gaussian Filter)
Test Action Video
GSSM
Orientational Features (Gabor Filter)
PCA
SVM
Action label
Contour Features (Wavelet Filter)

Figure.1 Block diagram of proposed HAR system

Key frames selection

Key frames selection is most important in the MVHAR system. The action videos captured under multiple views consist of almost similar data, and when they are considered as it is for recognition, it results in an unnecessary computational complexity followed by an increased computational time. Hence, this approach focused to extract only key frames for every action at both training and testing phases. To extract the key frames, this paper considered the SSM as a base reference and proposed a new version of SSM, called as Gradient SSM (GSSM).

Overview of SSM

SSM is matrix which explores the temporal similarities between frames of a same action. The main advantage with SSM is the study of temporal dynamics of an action. For a given action having N number of frames, the SSM is constructed as a

$N \times N$ matrix and the elements of SSM matrix are obtained based on Euclidean distance evaluation. For a sequence of frames, $F = \{F_1, F_2, F_3, \ldots., F_N\}$ is discrete (x, y, t) −space, a SSM of F is obtained as

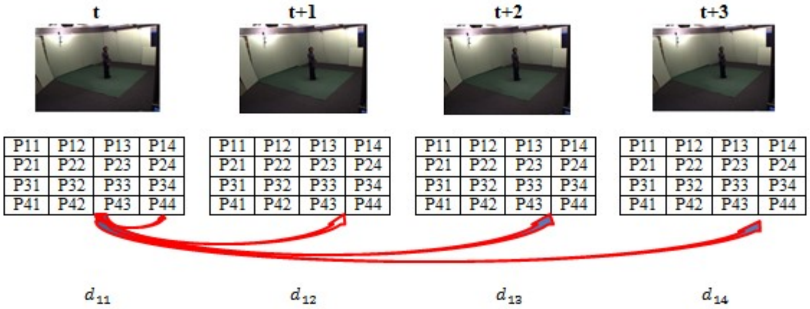
$$SSM(F) = \begin{bmatrix} 0 & d_{12} & d_{13} & d_{14} & \ldots & d_{1T} \\ d_{21} & d_{22} & d_{23} & d_{24} & \ldots & d_{2T} \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots \\ d_{T1} & d_{T2} & d_{T3} & d_{T4} & \ldots & d_{TT} \end{bmatrix} \qquad (1)$$

Where d_{ij} is the Euclidean distance between two frames, $F_i$and F_j. As shown in the above matrix, the diagonal elements are zero and these zeroes denote the similarity between same frames. For instance, the zero at the first row and the first Colum denotes the similarity between the first frame and itself. Next the term d_{12} is the Euclidean distance between first frame and second frame. This process continues for all frames and based on these observations, the most important two properties of SSM are defined as;

1). $d_{ij} = d_{ji}$

2). $d_{ij} \leq 0$ (Non-Negative)

SSM is much effective in the selection of frames based on observation of temporal dynamics, i.e., they key frames can be selected on the basis of constraint, $\max(SSM)$. It means the frames with minimum similarity can be considered key frames. A sample representation of temporal dynamics analysis for first four frames through SSM is shown in Figure.2. In figure.2, totally four successive frames with time instants *t, t+1, t+2*, and *t+3* are considered and first frame is processed to discover the temporal dynamics with respect to the frames at further time instants. Here, the term d_{12} is the Euclidean distance between two frames such as frame at time *t* and frame at time *t+1*. Next, the term d_{13} is the Euclidean distance between two frames such as frame at time *t* and frame at time *t+2* and the term d_{14} is the Euclidean distance between two frames such as frame at time *t* and frame at time *t+3*. These distance metrics alleviates the similarity between two frames at different times and thus the SSM can be used to analyze the temporal dynamics.



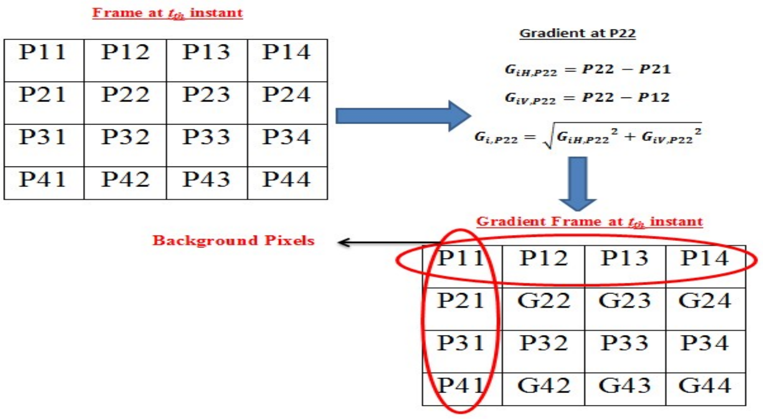
P11	P12	P13	P14	P11	P12	P13	P14	P11	P12	P13	P14	P11	P12	P13	P14
P21	P22	P23	P24	P21	P22	P23	P24	P21	P22	P23	P24	P21	P22	P23	P24
P31	P32	P33	P34	P31	P32	P33	P34	P31	P32	P33	P34	P31	P32	P33	P34
P41	P42	P43	P44	P41	P42	P43	P44	P41	P42	P43	P44	P41	P42	P43	P44

Figure.2 Temporal dynamics analysis through SSM

K. Pradeep Reddy et al [XXII] proposed a new key frame selection mechanism based on the observation of both inter-class and intra-class variations between frames of a single action captured under multiple views. However, this approach evaluated the SSM based on the pixel intensities and it is not effective in the case of a video with more occlusions and moving backgrounds. Hence this paper proposed an adaptive SSM, called as Gradient SSM in which the SSM is applied over the trajectories of action frames.

Gradient SSM

Here, the main intention of GSSM is to extract the key frames with respect to the trajectories of an action sequence. For a given action, a continuous trajectory is extracted through it's gradients. Unlike the conventional SSM which is focused only on the exploration of temporal dynamics, the proposed GSSM focuses on the spatial dynamics also. The GSSM effectively discovers the similarity between two frames in the occlusion and moving background. The gradients evaluation for a frame att_{th} instant is shown in Figure.3.


Frame at t_{th} instant
Gradient at P22
$G_{iH,P22} = P22 - P21$
$G_{iV,P22} = P22 - P12$
$G_{i,P22} = \sqrt{G_{iH,P22}^{2} + G_{iV,P22}^{2}}$
Gradient Frame at t_{th} instant
Background Pixels

P11	P12	P13	P14
P21	P22	P23	P24
P31	P32	P33	P34
P41	P42	P43	P44

P11	P12	P13	P14
P21	G22	G23	G24
P31	P32	P33	P34
P41	G42	G43	G44

Figure.3 Gradients Evaluation of a sample frame

Here, in the proposed GSSM evaluation, Laplacian operator is used to find the gradients. Since the Laplacian kernel is achieved a greater performance in edge enhancement in digital image processing, we adopt Laplacian operator to capture the spatial similarities. Mainly there are two reasons behind the consideration of Laplacian operator. (1) In the action image, the Laplacian operator can enhance the features with sharp discontinuities, highlights edges, and also can find the fine details. (2) Since the Laplacian operator is the second order derivate, it is more effective than the first order derivative in the analysis of finer details of image. Due to these two reasons, the Laplacian operator is considered here to extract the edge details and to highlight the finder details.

Consider a frame F_i from the available N frames $F = \{F_1, F_2, F_3, \dots, F_N\}$ of an action sequence, it is represent with a set of feature as $F_i = \{f_1, f_2, f_3, \dots, f_M\}$, where $f_i \in R^d$ is the ith feature. First apply gradient operator ∇ on the frame F, resulting in a first order gradient sequence as

$$G = \nabla F = \{G_1, G_2, G_3, \ldots., G_M\} \tag{2}$$

Where $G_i = dF_i/dt = f_i - f_{i-1}$. Since the input considered here is a 2-D image, the gradients are applied in the two directions, i.e., horizontal and vertical directions. Let $G_{iH}$and G_{iV} be the horizontal and vertical gradients respectively of an feature f_i, the final gradient can be obtained as

$$G_i = \sqrt{{G_{iH}}^2 + {G_{iV}}^2} \tag{3}$$

For every pixel/feature of a frame, this operation is performed and the entire frame is represented with first order derivatives. Next, apply the gradient operator $\nabla$to G, resulting in an another sequence as

$$L = \nabla \mathrm{G} = \{L_1, L_2, L_3, \ldots., L_M\} \tag{4}$$

Where $L_i = dG_i/dt = G_i - G_{i-1}$. Since the input considered here is a 2-D image, the gradients are applied in the two directions, i.e., horizontal and vertical directions. Let $L_{iH}$and L_{iV} be the horizontal and vertical gradients respectively of feature G_i, the final second order gradient can be obtained as

$$L_i = \sqrt{{L_{iH}}^2 + {L_{iV}}^2} \tag{5}$$

The resultant sequence L is the second order difference of a frame F. Simply L can be represented as $L = \nabla^2 F$. Initially, every frame is processed for gradients evaluation and then the resultant gradient frames are processed for SSM evaluation. Similar to the case when the Laplacian operator applied over an image/frame, the steep changes, edges and finer details are enhanced which are more helpful in the detection of key frames from a larger set of frames. For a key frame, to discriminate between two actions, the system should have sufficient knowledge and it is provided by the enhanced edges, sharp points and steep changes.

Once the key frames are extracted from every view, they are processed for key frame extraction according to the method described in [XXI].

Feature Extraction

Similar to the feature extraction method described in [XXI], this paper also considers three different features such as Intensity Features, Orientational Features and Contour Features.

To extract the intensity features, this work applied Gaussian pyramid filtering. Here the Gaussian filter is applied in a pyramidal fashion up to seven levels and the intensity features are extracted by subtracting the high level Gaussian convolved features from low level Gaussian convolved features. For an every level, the input frame is down sampled and convolved with Gaussian filter and then subtracted from its previous level frame.

To extract the Orientational features, this paper accomplished Gabor filter in various orientations and scales. For a given action frame/image, the Gabor filter is applied at totally eight orientations such as 0^0, 45^0, 90^0, 135^0, 180^0, 225^0, 270^0, and 315^0 and at different scales such as 5×5, 7×7, 9×9, and 11×11. Hence totally we will get 32 feature maps. Only eight important feature maps are extracted from these 32 by applying max pooling over them. For every orientation, we will have four feature maps at different scale and among those four only one feature map is extracted through max pooling, hence totally we will get eight Orientational feature maps at this phase.

Further, in the contour features extraction phase, Discrete Wavelet Transform (DWT) is applied up-to five levels. Initially, the input action key frame is decomposed into four sub bands such as Approximations (A), Horizontals (H), Verticals (V) and Details (D). In the further decompositions, the approximation band is considered as input and further decomposed into four sub bands. In this manner, the DWT is accomplished up-to five levels. Similar to the intensity feature extraction mechanism, the contour features are extracted by subtracting the high level sub-band features from low level sub-band features. Here, to ensure the dimensionality equality, interpolation is applied on the high level sub bands. The subtraction is applied only between approximation bands.

Finally, a 1-D feature vector is constructed by combining all the three features and applied Principal Component Analysis (PCA) to reduce the dimensionality. Over the obtained principal components 90% of components are considered as final set of features.

IV. Simulation Results

In this section, we assess the proposed HAR framework on two publicly available datasets namely, IXMAS [IX] and NIXMAS [X]. The quantitative evaluation is done through various performance metrics under varying environments. To simulate the developed HAR 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 action sequences and with different views.

Dataset details

We consider totally two different datasets. The first dataset is INRIA Xmas Motion Acquisition Sequences (IXMAS) dataset. This dataset consists of 12 action classes such as *point out (PO), pick up (PU), wave (WV), punch (P), turn around (TA), walk (WA), sit down (SD), get up (GU), Cross arms (CA), scratch head (SH), Kick (K)* and *check watch (CW)*. Each action is performed three times and 12 different subjects are recorded with five cameras, four are fixed at four sides and one is fixed 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 390×291 pixels. Some action samples of this dataset are shown in Figure.4.

The next dataset is NIXMAS having new videos with same action as of IXMAS dataset. The overall sequences present in this dataset are 1148. The actions recorded under this dataset are with different camera, actors and viewpoints. Moreover, the two to third ratio of videos are having the objects that occlude the actors. The major difference between IXMAS and NIXMAS is background only. In the IXMAS action videos, for all views, the background is constant and non-varying in nature, but in the NIXMAS dataset, the background is varying and consists of various objects. Some action samples of this dataset are shown in figure.5.

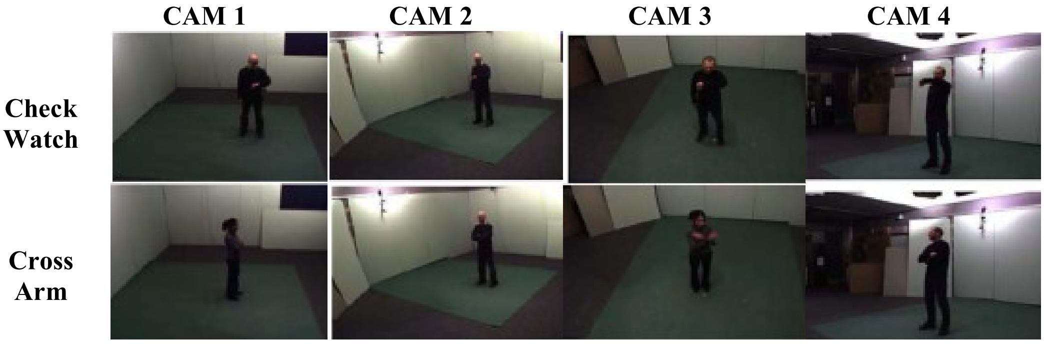


Figure.4 IXMAS dataset Action sample under multiple views

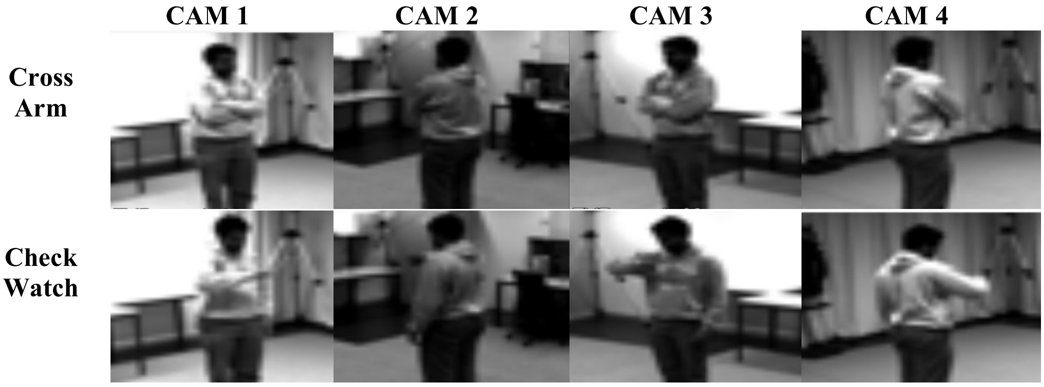


Figure.5 NIXMAS dataset Action samples under multiple views

Quantitative Evaluation

Under this evaluation, the performance is measured through various performance Metris such as Recall or Detection Rate or True Positive Rate (TPR), True Negative Rate (TNR), precision or positive predictive value (PPV), False Positive Rate (FPR), False Negative Rate (FNR), and Accuracy. These performance metrics are obtained based the following mathematical formulations as;

$$True\ Positive\ Rate\ (TPR) = \frac{TP}{TP+FN} \tag{6}$$

$$True\ Negative\ Rate\ (TNR) = \frac{TN}{TN+F} \tag{7}$$

$$Precision\ (PPV) = \frac{TP}{TP+FP} \tag{8}$$

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{FP+TN} \tag{9}$$

$$False\ Negative\ Rate\ (FNR) = \frac{FN}{FN+TP} \tag{10}$$

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

Here in the simulation, totally 12 actions are tested and the observed performance metrics are shown below. Initially, the simulation is conducted over the IXMAS dataset and then over the NIXMAS dataset. After testing all the action sequences, a confusion matrix prepared, according to the Table.1. The model of sample confusion matrix is shown in table.1, for a test case of CW only. In the case of signals testing are related to CW class, the actions labeled as CW are counted as TP and remaining are counted as FN. Similarly this is applied for remaining classes also. Based on this confusion matrix, the above specified performance metrics are measured and outlined in Table.2.

Next, the simulation is conducted through different view-points, i.e., the actions under different views are processed and the obtained class labels are observed. Under this case, along with the class label, the system also produces an action sequence attached with that label. Based on the obtained action, view, the confusion matrix is prepared and the performance is measured through the performance metrics. The obtained metrics under this simulation study is represented in Table.3.

Table.1 Model of Confusion Matrix

		Predicted											
		CW	**CA**	**SH**	**SD**	**GU**	**TA**	**WA**	**WV**	**P**	**K**	**PO**	**PU**
Actual	**Check Watch**	**TP**	FN	FN	FN	FN	FN	FN	FN	FN	FN	FN	FN
	Cross Arms	FP	**TN**	FN	FN	FN	FN	FN	FN	FN	FN	FN	FN
	Scratch Head	FP	FN	**TN**	FN	FN	FN	FN	FN	FN	FN	FN	FN
	Sit Down	FP	FN	FN	**TN**	FN	FN	FN	FN	FN	FN	FN	FN
	Get Up	FP	FN	FN	FN	**TN**	FN	FN	FN	FN	FN	FN	FN
	Turn Around	FP	FN	FN	FN	FN	**TN**	FN	FN	FN	FN	FN	FN
	Walk	FP	FN	FN	FN	FN	FN	**TN**	FN	FN	FN	FN	FN
	Wave	FP	FN	FN	FN	FN	FN	FN	**TN**	FN	FN	FN	FN
	Punch	FP	FN	FN	FN	FN	FN	FN	FN	**TN**	FN	FN	FN
	Kick	FP	FN	FN	FN	FN	FN	FN	FN	FN	**TN**	FN	FN
	Point Out	FP	FN	FN	FN	FN	FN	FN	FN	FN	FN	**TN**	FN
	Pick Up	FP	FN	FN	FN	FN	FN	FN	FN	FN	FN	FN	**TN**

Table.2 Performance Metrics for different actions under IXMAS dataset

Action/Metric	TPR (%)	TNR (%)	PPV (%)	FPR (%)	FNR (%)	F-Score (%)
Check Watch	94.2721	95.2095	94.3679	4.7909	5.7278	94.3200
Cross Arms	94.4292	95.3665	94.6601	4.6338	5.5707	94.5445
Scratch Head	93.6888	94.6257	93.7846	5.3742	6.3111	93.7367
Sit Down	91.2932	92.2301	91.3890	7.7698	8.7067	91.3411
Get Up	91.8172	92.7541	91.9130	7.2458	8.1827	91.8651
Turn Around	90.1805	91.1174	90.2763	8.8825	9.8194	90.2284
Walk	93.2461	94.1830	93.3419	5.8169	6.7538	93.2940
Wave	94.2794	95.2163	94.3752	4.7836	5.7205	94.3273
Punch	88.2718	89.2087	88.5932	10.7911	11.7281	88.4322
Kick	89.5643	90.5012	89.7101	9.4987	10.4356	89.6371
Point Out	90.3110	91.2479	90.9271	8.7520	9.6889	90.6180
Pick Up	90.9394	91.8763	91.0394	8.1236	9.0605	90.9894

Table.3 Performance Metrics for different actions under Different Views

Camera/Metric	TPR (%)	TNR (%)	PPV (%)	FPR (%)	FNR (%)	F-Score (%)
CAM 1	91.5161	92.3247	91.8900	7.6753	8.4839	91.7027
CAM 2	89.8794	90.6880	90.8702	9.3120	10.1206	90.3721
CAM 3	90.9450	91.7536	89.8976	8.2464	9.0550	90.4183
CAM 4	89.9783	90.7869	91.1093	9.2131	10.0217	90.5403

Table.2 gives the details of performance metrics evaluated after the simulation of proposed approach over IXMAS dataset. These metrics are obtained after testing the entire 12 actions one by one. In table.2, the TPR is measured as the ratio of total number of action sequences recognized correctly to the total number of action sequences given as input. For example, for a given Check Watch Action, the TPR is measured as total Check watch actions recognized correctly for a total check watch action sequences processed for testing. Next, the TNR is measured as the ratio of total number of action sequences recognized as true negatives for a sum of total number of false positive actions and true negative actions. In the case of Check watch action, any other actions given as input and recognized as correctly is considered as true negative and the total of such actions is measured a TNR. Further, the PPV or Precision measures the precise recognition for all actions. It is measured as the total number of actions recognized correctly to the total number of action processed for testing in accrual. The further two metrics such as FNR and FPR are opposite two TPR and TNR respectively. Finally, the F-Score is measured as the harmonic mean of recall and precision. In the table.2, except for punch and kick actions, the TPR for remaining actions is observed t be high and it is above 90% and for cross arms action, the TPR is observed to be maximum and for punch action, it is observed as minimum. Similarly the PPV also has obtained the same characteristics and it is maximum for cross arms action and minimum for punch action.

The further simulation study is done through multiple view-points. Under this simulation, the actions of multiple views are processed for testing and the obtained results are verified with input views. For a given action under left side view, the obtained same action and same view is only considered as true positive and the total such count is considered as TPR. Further, the TNR is measured as total number of negative view recognized correctly. The obtained results are shown in table.3, and the best performance is obtained for actions given under CAM 2. The Maximum TPR is obtained for CAM 1 and minimum is obtained for CAM 2. Further the maximum precision is occurred for action tested under CAM 1 and minimum is for CAM 3. Simultaneously, the highest FPR and FNR is observed for actions tested under CAM 2 and CAM 2 respectively and lowest is observed at CAM 1 and CAM 1 respectively. Next, the simulation is conducted for actions of NIXMAS dataset and the results are enumerated in table.4 and table.5. Similar to the above simulation strategy, same actions are processed but with varying actors, and captured environments.

Table.4 Performance Metrics for different actions under NIXMAS dataset

Action/Metric	TPR (%)	TNR (%)	PPV (%)	FPR (%)	FNR (%)	F-Score (%)
Check Watch	91.5660	92.5034	91.6618	7.4965	8.4339	91.6139
Cross Arms	91.7231	92.6604	91.9540	7.3395	8.2768	91.8384
Scratch Head	90.9827	91.9196	91.0785	8.0803	9.0172	91.0306
Sit Down	88.5871	89.5240	88.6829	10.475	11.4128	88.6350
Get Up	89.1111	90.0480	89.2069	9.9519	10.8888	89.1590
Turn Around	87.4744	88.4113	87.5702	11.5885	12.5255	87.5223
Walk	90.5400	91.4769	90.6358	8.5230	9.45994	90.5879
Wave	91.5733	92.5102	91.6691	7.4897	8.42664	91.6212
Punch	85.5657	86.5026	85.8871	13.4973	14.4342	85.7261
Kick	86.8582	87.7951	87.0040	12.2048	13.1417	86.9310
Point Out	87.6049	88.5418	88.2214	11.4581	12.3950	87.9121
Pick Up	88.2333	89.1702	88.3333	10.8297	11.7666	88.2833

Table.5 Performance Metrics for different actions of NIXMAS dataset under Different Views

Camera/Metric	TPR (%)	TNR (%)	PPV (%)	FPR (%)	FNR (%)	F-Score (%)
CAM 1	88.4552	88.9988	89.4512	11.0012	11.5448	88.9504
CAM 2	85.4885	87.9333	85.7489	12.0667	14.5115	85.6185
CAM 3	89.0263	86.2141	86.7314	13.7859	10.9737	87.8639
CAM 4	86.2353	87.8796	87.3939	12.1204	13.7467	86.8107

Table.4 depicts the details of performance evaluation of proposed approach over the NIXMAS dataset. Under this case, the actions are occluded and also the background carries varying objects like tables, and chairs etc. All the actions are tested one by one, and then based on the obtained class labels; the performance is measured through the performance metrics. Compared to the performance analysis shown in table.2, the performance analysis is observed to be poor. For example, let's consider an action Turn Around. The obtained TPR under normal case is 90.1805, whereas the

TPR under Occluded case is 87.4744. Further, the precision is observed as 90.2763 and 87.5702 under the normal and Occluded simulation respectively. This is mainly due to the presence of occlusion in the actions by which the system may get confused and results in a wrong classification. These occlusions results in lower TPR, TNR and PPV and higher FPR and FNR.

The next simulation is accomplished through multiple view-points and the results are shown in table.5. Compared to the results obtained in table.3, the results of table.5 are poor. In this case, for a given action, under multiple views, the background objects such as tables, and chairs also changes which results more confusion for recognition system. Particularly, it can be observed that the view which has less variations has obtained maximum TPR or detection rate. From the table.5, the highest TPR is observed for CAM 3 and the lowest is for CAM 2. Since the actions performed under CAM 2 are backside views, the main hands movement is not much disclosed by any feature extraction techniques and hence resulted in a less TPR. Furthermore, the precision is also less for CAM 2 only and it is high for CAM 1 due to the clear visualization of hand movements. Since CAM 1 is a front view, the TNR and PPV are observed as more for the all actions captured through CAM 1.

Comparative Analysis

Under this subsection, the performance enhancement of proposed approach is exposed by comparing it with some conventional HAR methods such as A. Aryanfar et al. [I] and K. P. Reddy et al. [XXI]. The main performance metric considered for this comparison purpose is Recognition Accuracy. The obtained accuracy results are shown in the figure.6.

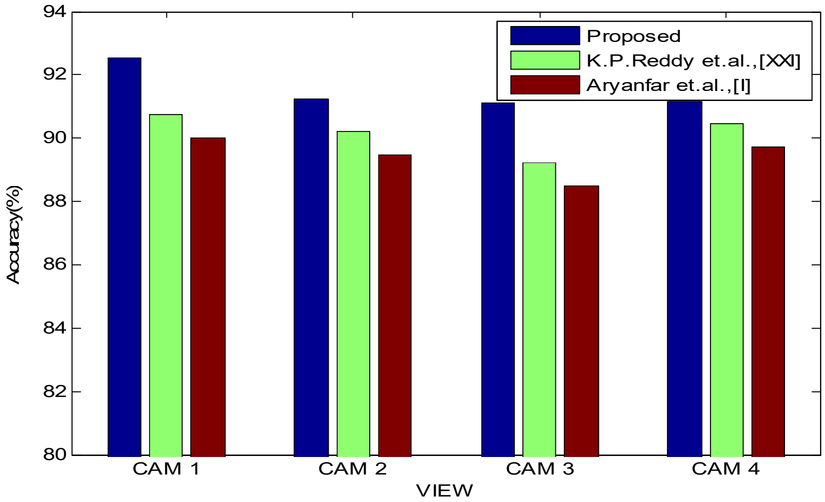


Figure.6 Accuracy comparison under multiple views

The main theme of HAR method proposed in [I] is the Multi View action recognition with wavelet based feature extraction. Next, in the HAR method proposed by K.P Reddy et al. [XXI], the main objective is to obtain maximum recognition accuracy under multiple views for all actions. For this purpose, initially, the SSM is accomplished for key frame selection and then applied a three stage feature extraction technique for extracting all possible discriminating features from every action. In an average, the accuracy of proposed approach is observed as 91.5118% and for conventional approaches; it is 90.1615% and 89.5768% for K.P Reddy et al. and Aryanfar et al.,respectively. Due to the consideration of multiple features followed by a novel key frame selection mechanism, the accuracy of proposed approach is high. Whereas, in the Method proposed by Aryanfar et al., Only Wavelet features are considered for feature extraction and no preprocessing mechanism for the selection of key frames. Moreover, there is no preprocessing at which the noise or redundant data is focused to reduce. Next, in the method proposed by K.P Reddy et al., though the key frames are extracted from a set of large number of frame, the SSM considered for key frames selection is directly applied over frames. Due to this type of frame selection, the frames with varying background objects are also extracted as key frames which results in less accuracy, especially for action of NIXMAS dataset.

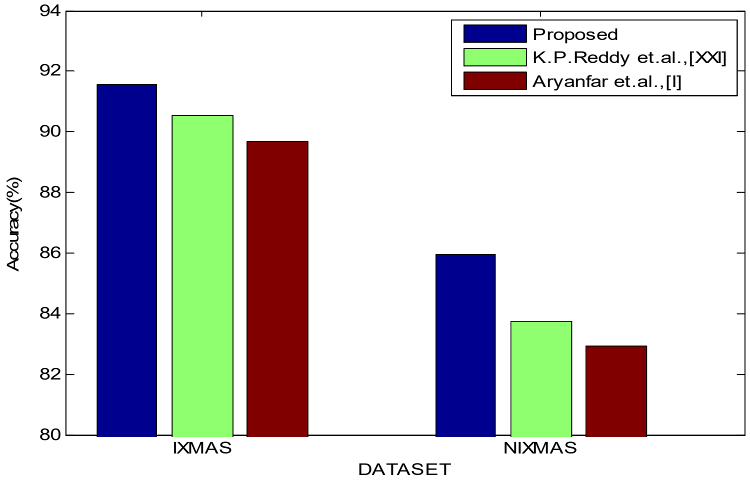


Figure.7 Accuracy comparison under Various Datasets

A further comparison shown in figure.7 is based on the datasets, i.e., the accuracy evaluated after the simulation of two different datasets such as IXMAS and NIXMAS. As it can be seen from figure.7, the accuracy of all method under the simulation of actions of IXMAS dataset is more compared to the accuracy of actions of NIXMAS dataset. The main reason behind this enhancement is the quality of video. In the actions of IXMAS dataset, the video is more qualitative and also no much background variations, whereas in the actions of NIXMAS, the videos are Occluded and also having varying background objects. The accuracy for IXMAS is observed as 91.5541%, 90.5247% and 88.8115% for proposed, K.P. Reddy et al.,and Aryanfar et al., respectively. Similarly, the accuracy for NIXMAS is observed as 85.9685%, 83.7415% and 82.0283% for proposed, K.P. Reddy et al., and Aryanfar et al., respectively. Overall, the proposed approach is obtained a high accuracy for both datasets and this is due to the accomplishment of advanced key frames selection and hybrid feature extraction technique.

V. Conclusion

In this paper, we proposed a new Multi-View Human Action Recognition Framework based on the Self-Similarity Matrix and a hybrid feature extraction technique. This framework selects the key frames based on a Gradient Self-similarity matrix, which is an extended version of Self-similarity matrix. Without any deviation to the basic methodology of SSM, this approach developed a new variant of SSM, called GSSM, based on the gradients in the action frames. This GSSM successfully extracts the key frames in which the variations are more with respect to the moving human body. Due to the accomplishment of the second order derivatives, the variations in the movements of human body are more accurately and helped much in the selection of only important frames for a given action video. Further, the proposed hybrid feature extraction techniques helps in the provision of sufficient discrimination between different actions with respect to intensity, orientation and contours. Extensive simulations conducted over two different datasets such as IXMAS and NIXMAS had shown the effectiveness of proposed approach. Furthermore, the comparative analysis conducted between proposed and conventional approaches had proven the efficacy in the recognition of all possible human actions under varying environments.

References

I. 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*, Berkeley, Suta, pp.585-592, 2015.

II. 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

III. A. Eweiwi, M.S. Cheema, C. Bauckhage, J. Gall, Efficient pose-based action recognition, in: Asian Conference on Computer Vision, Springer, 2014, pp. 428–443.

IV. A. Farhadi and M. K. Tabrizi, “Learning to recognize activities from the A. wrong view point,” in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 154–166.

V. A. Gilbert, J. Illingworth, and R. Bowden, “Scale invariant action recognition using compound features mined from dense Spatio-temporal corners,” in *Proc. ECCV*, 2008, pp. I: 222–233.

VI. C. H. Lampert, H. Nickisch, S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In CVPR, 2009.

VII. C. Rao, A. Yilmaz, and M. Shah, View-invariant representation and recognition of actions, *Int. J. Comput. Vis.*, vol. 50, no. 2, pp. 203–226, 2002.

VIII. C. Schuldt, I. Laptev, and B. Caputo, ‘‘Recognizing human actions: A local SVM approach,’’ *in Proc. 17th Int. Conf. Pattern Recognit.*, vol. 3. Aug. 2004, pp. 32–36.

IX. D. Weinland, E. Boyer, and R. Ronfard, “Action recognition from arbitrary views using 3D exemplars,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–7.

X. D. Weinland, M. Özuysal, and P. Fua, “Making action recognition robust to occlusions and viewpoint changes,” in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 635–648.

XI. E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In CVPR, 2007.

XII. H. Wang, A. Klaser, C. Schmid, and C. Liu, “Action recognition by dense trajectories,” in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3169–3176.

XIII. I. Junejo, “Self-similarity based action recognition using conditional random fields,” in *Information Retrieval Knowledge Management (CAMP), 2012 International Conference on*, march 2012, pp. 254 –259.

XIV. 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. 99, no. PrePrints, 2010.

XV. Ivo Everts, Jan C. van Gemert and Theo Gevers, Evaluation of Color STIPs for Human Action Recognition, *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, 2013.

XVI. Jing Wang, and Huicheng Zheng, “View-robust action recognition based on temporal self-similarities and dynamic time warping”, *IEEE International Conference on Computer Science and Automation Engineering (CSAE*), Zhangjiajie, China, 2012.

XVII. J. Wang, Z. Chen, and Y. Wu, “Action recognition with multiscale Spatio-temporal contexts,” in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3185–3192.

XVIII. K. Chatfield, J. Philbin, and A. Zisserman. Efficient retrieval of deformable shape classes using local self-similarities. In NORDIA Workshop at ICCV 2009, 2009.

XIX. K. G. C. Manosha, Ranga Rodrigo, Faster Human Activity Recognition with SVM, The International Conference on Advances in ICT for Emerging Regions – ICTER 2012 : 197-203.

XX. K. Huang, Y. Zhang, and T. Tan, “A discriminative model of motion and cross ratio for view-invariant action recognition,” *Image Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 2187–2197, 2012.

XXI. K. Pradeep Reddy, G. Apparao Naidu, B.VishnuVardhan, “View-Invariant Feature Representation for Action Recognition under Multiple Views”, *International Journal of Intelligent Engineering and Systems,* Vol.12, No.6, 2019.

XXII. L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *PAMI*, vol. 29, no. 12, pp. 2247–2253, December 2007.

XXIII. M. Abd-el-Kader, W. Abd-Almageed, A. Srivastava, and R. Chellappa, "Silhouette-based gesture and action recognition via modeling trajectories on Riemannian shape manifolds," *CVIU*, vol. 115, no. 3, pp. 439–455, 2011.

XXIV. M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007.

XXV. M. Stark, M. Goesele, and B. Schiele. A shape-based object class model for knowledge transfer. In ICCV, 2009.

XXVI. Paul, S.N.; Singh, Y.J. Survey on Video Analysis of Human Walking Motion. Int. J. Signal Process. Image Process. Pattern Recognit. 2014, 7, 99–122.

XXVII. P. Matikainen, R. Sukthankar, and M. Hebert, "Feature seeding for action recognition," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1716–1723.

XXVIII. Q. Le, W. Zou, S. Yeung, and A. Ng, "Learning hierarchical invariant Spatio-temporal features for action recognition with independent subspace analysis," in *CVPR*. IEEE, 2011, pp. 3361–3368.

XXIX. R. Poppe, A survey on vision-based human action recognition, Image Vision Comput. 28 (6) (2010) 976–990.

XXX. SamySadek, Ayoub Al-Hamadi, Bernd Michaelis, and UsamaSayed, "An Action Recognition Scheme Using Fuzzy Log-Polar Histogram and Temporal Self-Similarity", *Hindawi Publishing Corporation EURASIP Journal on Advances in Signal Processing,* Volume 2011, Article ID 540375, 9 pages.

XXXI. S. Maji, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp.3177–3184.

XXXII. S. Singh, S. A. Velastin, and H. Ragheb, ‘‘MuHAVi: A multi-camera human action video dataset for the evaluation of action recognition methods,’’ *in Proc. 17th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS),* Sep. 2010, pp. 48–55.

XXXIII. Stephen Karungaru, Masayuki Daikoku, Kenji Terada, "Multi Cameras Based Indoors Human Action Recognition Using Fuzzy Rules", JPRR Vol 10, No 1 (2015).

XXXIV. S. Vishwakarma, A. Agrawal, A survey on activity recognition and behavior understanding in video surveillance, Visual Computer 29 (10) (2013) 983–1009

XXXV. S. Wu, O. Oreifej, and M. Shah, "Action recognition in videos acquired by a moving camera using motion decomposition of Lagrangian particle trajectories," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1419–1426.

XXXVI. T. Deselaers and V. Ferrari, "Global and efficient self-similarity for object classification and detection," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1633–1640.

XXXVII. T. SyedaMahmood, M. Vasilescu, and S. Sethi, Recognizing action events from multiple viewpoints, in Proc. EventVideo, 2001, pp. 64–72.

XXXVIII. X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *CVPR 2011*. IEEE, 2011, pp. 489–496.

XXXIX. Yen-Pin HsuChengyin Liu Tzu-Yang Chen Li-Chen Fu, "Online view-invariant human action recognition using RGB-D Spatio-temporal matrix", *Pattern recognition*, Volume 60, December 2016, Pages 215-226.