Human Action recognition using STIP Evaluation techniques

Latha BM\textsuperscript{1}, Manjula BK\textsuperscript{2}, Venkata Sumana CH\textsuperscript{3} and Hemalatha KL\textsuperscript{4}

\textsuperscript{1,2,3} Department of Electronics and Communication Engineering GM Institute of Technology Davangere, India
\textsuperscript{4} Department of Information Science, Sri Krishna Institute of Technology, Chikkabanavara, Karnataka, India

\begin{itemize}
\item \textsuperscript{1}lathabm@gmit.ac.in, \textsuperscript{2}manjulabk@gmit.ac.in, \textsuperscript{3}sumanavch@gmit.ac.in, \textsuperscript{4}hema.skit@gmail.com
\end{itemize}

\textbf{Abstract.} Human action recognition plays a big role in human-to-human interaction and interpersonal relations. As a result, it provides information regarding the identity of someone, their temperament, and condition, it's tough to extract. Recognition of actions within the video isn't a matter for human sensory system. The identification of the actions of the person by the system wants someone special mechanisms. The identification of the actions by the systems are going to be useful in computer vision method. This method is divided into low level action recognition process and high level recognition process. Recognizing the actions victimization the one feature values extracted goes under low level action recognition method. These method is simple to implement and that they don't seem to be reliable all the time. The high level action recognition method needs some special hardware’s to discover the actions within the video. The experimental results proved that the proposed methodology achieved better performance in terms of accuracy, sensitivity and specificity.

\textbf{Keywords:} Human Action Recognition; Human Sensory System; Interpersonal relations; Low-level and High-level Action Recognition; Video.

\section{1. Introduction}
In recent decades, the human ability to acknowledge another person’s activities is one among the most subjects of study of the scientific areas of computer vision and machine learning [1]. As a results, the human action recognition is applied in many applications; human-computer interaction, video surveillance systems and robotics. Where, these applications require a multiple action recognition system. In image and video analysis, human action recognition is a very important analysis direction [2,3]. Within the past a few years, a large range of papers are printed on human action recognition in video and image sequences. Information could also be lost once either intensity or chromatic representations are thought-about in isolation [4]. The method victimization motion trajectories involve following and dense multi-scale optical flow computation, the associated process quality is often over that of STIP-based approaches. Here, we offer a comprehensive survey of the recent development of the
techniques, as well as ways, systems and quantitative analysis of the performance of human action recognition [5].

The videos containing actions performed by persons were collected. The videos were converted into frames and also the frames were pre-processed. Pre-processing is completed by applying median filter [6]. The median filter finds the noises within the frame and replaces the noise by replace the element victimization the median of the neighbor pixels. Options were extracted from the frames. The extracted options are going to be accustomed acknowledge the action of the person within the video [7]. Harris SPIT, Dennis Gabor SPIT and HOG SPIT were used to extract the feature values from the video frames. The options were then classified victimization SVM classifier supported the kernel operate [8]. The action is recognized supported the label came back by the classifier. The recognition of actions in the video is helpful in computer vision. This may be useful to develop a system which will acknowledge the actions in the surroundings [9]. This paper has scope in varied fields such as a spread of systems that involve interactions between persons and electronic devices like human-computer interfaces, surveillance systems, patient monitoring systems, Military purpose and Sports. The performance of the planned technique is obtained by activity the performance of the classifier [10, 23]. The Accuracy, Sensitivity and Specificity of the classifier is that the most ordinarily lived performance measure for the classifier. The experimental results show that our technique will considerably improve classification, interpretation and retrieval performance for the video pictures.

Section 2 discuss the studies of existing techniques for human action recognition. Section 3 describes the methodology for each module, where Section 4 provides the validation of proposed methodology with existing studies. Finally, the conclusion of the research study is presented in Section 5.

2. Literature Survey

T. Kalsum, et al, [19] used hybrid features for recognizing the human emotions from the facial expression. In this study a combination of features was utilized for feature extraction like bag of features, speeded up robust transform and scale invariant feature transform. These hybrid features improve the ability of facial expression recognition. While utilizing more number of features, the complexity of the system will increase. Further, Z. Gao, et al, [20] used a new natural inspired algorithm; Adaptive Fusion and Category-level Dictionary Learning (AFCDL) for human action recognition. In this paper, the developed AFCDL model performance was assessed on UCLA dataset. The simulation outcome showed that the developed model achieved 87.8% of accuracy. Still, the undertaken AFCDL algorithm need to focus on feature extraction to achieve better result in action recognition.

M. Liu, et al, [21] developed a new algorithm for human action recognition. Here, the motion and visual enhancement approaches were used on color images to improve the local patterns. Then, convolutional neural network was utilized for extracting the discriminative and robust feature vectors from the color images. The extensive experiment showed that the developed model attained 92.61% of recognition accuracy on UCLA dataset. Additionally, D.K. Vishwakarma, [22] extracted the decisive pose using two fold transformations (ridgelet and Gabor wavelet transform) for better human action recognition. However, these existing models showed low performance in the conditions like poor lighting, head pose variations, etc.

3. Methodology

3.1. Flow Chart

Initially, flow diagram of the proposed model is shown in Figure 1. To get rid of the noise from the videos, the input video frames are preprocessed by median filtering approach. The proposed model performance is improved by reducing the noise within the video. Many forms of noise are present in frames/images. The commonly obtained noise is salt and pepper noise, where it occurs in the white and black pixels.

At first, the unwanted pixels are removed from the video frames by applying median filter. In this research work, the noised pixel in the image is detected by median filter. The identified noisy pixel is
replaced by the median value of the neighboring pixel [11]. Figure 2 represents the videos are converted into frames and the Figure 3 states the denoised image.

**Figure 1.** Flow diagram of the proposed model

**Figure 2.** Graphical illustration of converting video into frames.

**Figure 3.** Pre-processed video frame

### 3.2. Feature extraction
From the preprocessed video frames, the feature values are extracted using STIP descriptors of various sorts. The un-normalized descriptors are extracted from the cuboids around STIP detection within the set of undistorted videos. Hence, the features like Harris STIP, Gabor STIP and HOG STIP are extracted from the frames.
3.2.1. Harris and Gabor STIP
The Harris STIP is used for detecting the corners in the frames [12]. It is used for detecting corner in every pixels of the image by considering the differential of the corner with reference to direction. The nearby edges will look similar, if the pixel in a region of uniform intensity.
In addition, the Gabor STIP is used to find the corners from the exact location of the object by Gabor wavelets. The Gabor function delivers a local spectral energy density for a given position and frequency in a certain direction [13]. The convolution of two perpendicular directions is performed with variously dilated wavelets [14]. The Gabor function provides a local spectral energy density, when s=2 and 4.

![Image](image_url)

Figure 4. The original input images (Left) and the subsampled filter result (Right)

3.2.2. HOG STIP
In human action recognition, HOG is an effective feature descriptor that is calculated using orientation and magnitude of the denoised images [15]. In HOG feature descriptor, the vertical $G_v$ and horizontal $G_h$ gradients of the images are determined by the equations (1) and (2).

$$G_v = I_N \times [-1,0,1]^T$$

$$G_h = I_N \times [-1,0,1]$$

The determined vertical $G_v$ and horizontal $G_h$ gradient values are utilized for estimating the gradient magnitude $m(x,y)$ and angular orientation $\theta(x,y)$ of the images by the equations (3) and (4).

$$m(x,y) = \sqrt{G_v^2(x,y) + G_h^2(x,y)}$$

$$\theta(x,y) = \tan^{-1}\left(\frac{G_v(x,y)}{G_h(x,y)}\right)$$

The gradient magnitude $m(x,y)$ and angular orientation $\theta(x,y)$ partition the pre-processed images into several cells. Then, the orientation related to the same cells are quantized and integrated into histogram bins. The obtained histogram bins are merged into final histogram. Total features $T_{hog}$ are determined using equation (5).

$$T_{hog} = B_{img} \times B_s \times N_b$$

Where, $B_s$ is indicated as block size $B_{img}$ is stated as blocks per image, and $N_b$ is denoted as total bins. Figure 5 shows the feature extraction techniques in HAR system.
3.3. Classification using SVM

SVM is a discriminative classifier, which is denoted by a hyper-plane. Recently, SVM classifier is highly utilized in several image processing applications, because it performs effectively in high dimension data. Additionally, the SVM performs well in two class problem, which are related to structure principles and vapnik–Chervonenkis theory [16-17].

The linear discriminant function is calculated using the formula \( w \cdot x + b = 0 \). However, an optimal hyper plane is utilized in SVM for classification that is denoted in equation (6).

\[
pi[w \cdot x + b] - 1 \geq 0, i = 1, 2, \ldots N
\]  

(6)

Then, diminish \( \|w\|^2 \) in equation (6) in order to resolve the optimization issue. Hence the ideal discriminant function is mathematically indicated in equation (7).

\[
f(x) = \text{sign}(\sum_{i=1}^{N} a_i^* \cdot pi(x_i^* - x) + b^*)
\]  

(7)

In order to reduce the complexity in high dimensional data, interchange the interior product \( (x_i^* - x) \) by linear kernel function \( k(x, x') \). So, the ideal discriminant function is re-written as indicated in equation (8). Graphically, the SVM classifier is indicated in Figure 6.

\[
f(x) = \text{sign}(\sum_{i=1}^{N} a_i^* \cdot pi. k(x, x_i) + b^*)
\]  

(8)
4. Results and Discussion

4.1. Performance measure
In this study, MATLAB (2019a) software is applied for experimental simulation with 4 GB RAM, Intel i3 processor and windows 10 operating system. In this research work, the proposed model performance is related with a few previous research models such as Enhanced Skeleton Visualization (ESV) and Adaptive Fusion and Category-level Dictionary Learning (AFCDL). In this scenario, the performance of the proposed model is evaluated in light of specificity, accuracy, and sensitivity. The mathematical expressions of specificity, accuracy, and sensitivity are indicated in the equations (9), (10) and (11).

\[
\text{Specificity} = \frac{TN}{FP + TN} \tag{9}
\]

\[
ACC = \frac{TP + TN}{(FP + TN) + (TP + FN)} \tag{10}
\]

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \tag{11}
\]

Where, TP is indicated as true positive (correctly identified), FP is denoted as false positive (incorrectly identified), TN is denoted as true negative (correctly rejected) and FN is stated as false negative (incorrectly rejected).

4.2. Quantitative analysis
In this study, UCLA dataset is used to validate the proposed and existing models (ESV and AFCDL) performance. Here, 80% of the data are utilized for training and 20% of the data are utilized for testing. In this research, Harris STIP, Gabor STIP and HOG STIP features are used to extract the feature vectors from the pre-processed data. The snapshot shown in Figure 7 is the result of Harris STIP, which identifies the corners by testing each pixel in the image.

![Figure 7. Extracting features using Harris STIP from the video frames](image)

In addition, Figure 8 illustrates Gabor function that gives a local spectral energy density concentrated around a given position and frequency in a certain direction. Additionally, Figure 9 indicates HOG feature extraction that compile a histogram of gradient directions or edge orientations for the pixels within the cell.
Figure 8. Feature extraction by Gabor STIP from video frames

Figure 9. HOG feature points from video frames

Action of the given input video sequence is recognized by the Multi SVM identifier, which is depicted in the Figure 10 that shows surfing is the action identified from the input video processing. Additionally, Figure 11 indicates the GUI performance of the processed video.

Figure 10. “Surfing” is one of the actions identified from the video sequences
4.3. Comparative analysis

Here, the proposed model performance is analyzed in light of specificity, sensitivity and accuracy. The UCLA dataset comprise of twenty actions such as jogging, two hand wave, horizontal arm wave, draw tick, tennis serve, high arm wave, hand clap, forward kick, forward punch, hand catch, pickup, throw, hammer, side-boxing, high throw, draw circle, bend, side kick, draw x, and golf swing. In UCLA database, the proposed model averagely attained 95.3% of accuracy, 100% of specificity and 92.5% of sensitivity that is shown in Table 1. From the experimental outcome, the proposed model showed better performance in human action recognition. The proposed model performance is graphically represented in Figure 12.

Table 1. Performance analysis of the proposed model

| Actions         | Accuracy (%) | Specificity (%) | Sensitivity (%) |
|-----------------|--------------|-----------------|-----------------|
| Jogging         | 96           | 100             | 94              |
| Two hand wave   | 95.4         | 100             | 92              |
| Horizontal arm wave | 95.9     | 100             | 93.5            |
| Draw tick       | 97           | 100             | 93              |
| Tennis serve    | 94           | 100             | 93.45           |
| High arm wave   | 94           | 100             | 92              |
| Hand clap       | 96.80        | 100             | 92              |
| Forward kick    | 92           | 100             | 92.41           |
| Forward punch   | 95.8         | 100             | 90              |
| Hand catch      | 97           | 100             | 95.65           |
| Pickup          | 97.32        | 100             | 93              |
| Throw           | 95.6         | 100             | 92.5            |
| Side-boxing     | 95.6         | 100             | 94              |
| Hammer          | 95           | 100             | 95.31           |
| High throw      | 96           | 100             | 92              |
| Draw circle     | 96.02        | 100             | 92              |
| Bend            | 95           | 100             | 91.63           |
| Side kick       | 95.21        | 100             | 92.5            |
| Draw x          | 96.28        | 100             | 92.25           |
| Golf swing      | 94.17        | 100             | 92.10           |
| Mean            | 95.3         | 100             | 92.5            |
In this scenario, Table 2 shows the comparative analysis of proposed and existing models such as AFCDL and ESV in terms of recognition accuracy. Z. Gao, et al, [20] utilized a new natural inspired algorithm; AFCDL for human action recognition. In this literature paper, the developed AFCDL model performance was investigated on UCLA database. In the experimental phase, the developed model achieved 87.8% of recognition accuracy. Similarly, M. Liu, et al, [21] has utilized motion and visual enhancement approaches on color images for improving the local patterns. Then, convolutional neural network was utilized for extracting the discriminative and robust feature vectors from the color images. Extensive experiment shows that the developed model attained 92.61% of recognition accuracy on UCLA dataset. Compared to the existing research works (AFCDL and ESV), the proposed model achieved 95.3% of accuracy that showed maximum of 7.5% and minimum of 2.69% of improvement in human action recognition. The comparative analysis is graphically stated in Figure 13.

| Methodology | Dataset | Accuracy (%) |
|-------------|---------|--------------|
| AFCDL [20]  | UCLA    | 87.80        |
| ESV [21]    |         | 92.61        |
| Proposed    |         | 95.3         |

**Figure 12.** Graphical evaluation of proposed model

**Figure 13.** Graphical comparison of proposed and existing models
5. Conclusion
In this research, the proposed model acknowledges the action of the persons within the video based on the options extracted using color strips. The extracted features area unit based on the STIP that is combined with several of the opposite ideas, so the feature extraction method is simpler. The recognition of the action is completed using the kernel function of SVM classifier. The proposed model provides better accuracy than the existing models, which showed that the misclassification area unit reduced to large extend. STIP descriptors are reformulated into photometric channels that leads to color strips. Here, the proposed model results in an improved balance between discriminative power and photometric invariance, as chromaticity provides a lot of information based on the higher representations. Color strips are effectively evaluated for recognizing human actions on difficult video benchmarks. The proposed model exactly recognizes the action of the person within the video even in the conditions; illumination variations, contrast variations, Abrupt motions and Scaling of the persons in the video.

References
[1] Atiqur R A, Tan J K, Kim H S and Ishikawa S 2008 Human Activity Recognition: Various Paradigms, International Conference on Control, Automation and Systems
[2] Lin S C F, Wong C Y, Ren T R and Kwok N M 2012 A Comparison Study on Human Action Recognition from Video Streams 5th Int. Congress on Image and Signal Processing 1162-1166
[3] Ong C A, and Theng L B 2014 Human Activity Recognition: A Review IEEE international conference on control system, computing and engineering (ICCSCE 2014) 389-393
[4] Xue L, Sheng Q Z, Pang C, Zhao X, and Wang S 2012 Effective Approaches in Human Action Recognition Int. Conf. on Advanced Computer Science and Information Systems (ICACSI) 1-7
[5] Luo D, Ekene H K, and Jun O 2012 Human Gesture Analysis using1 Multimodal features IEEE Int. Conf. on Multimedia and Expo Workshops.
[6] Chen C, Liu K and Kehatarnavaz N 2013 Real-time human action recognition based on depth1 motion maps Springer
[7] Suraj V and Babu R V 2014 Real-time Human Action Recognition from Motion Capture Data, 4th National Conf. on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCPVRIG) 1-4
[8] Nilam N, Sjarif A and Shamsuddin S M 2015 Human Action Invariancess for Human Action Recognition 7th Int. Conf. on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT) 263-266
[9] Nezhia J, Boujnah N, Htiwich O and Bouhllel MS 2016 Human Action Recognition to Human Behavior Analysis 7th Int. Conf. on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT) 263-266
[10] Chandni J D, Tushar V R 2016 A Survey on Human Action Recognition from1 Videos, Online Int. Conf. on Green Engineering and Technologies (IC-GET) 1-5
[11] Alexandre P, Tabia H, Declercq D and Zanotti A 2016 Feature covariance for human action recognition 6th Int. Conf. on Image Processing Theory, Tools and Applications (IPTA) 1-5
[12] Guillermo C, Albuquerque A, Ara’ujo 2009 Harris-SIFT Descriptor for Video Event Detection based on a Machine Learning Approach, 11th IEEE International Symposium on Multimedia.
[13] Feng Z, Xian-Da Z, Hu Y 2009 Gabor Filter Approach to Joint Feature Extraction and Target Recognition, IEEE Transactions on Aerospace and Electronic Systems 45.
[14] Yuanyuan H, Haomiao Y and Ping H 2012 Action Recognition Using HOG Feature in Different Resolution1 Video Sequences Int. Conf. on Computer Distributed Control and Intelligent Environmental Monitoring 85-88
[15] Mohamed I. K, El-Yacoubi M A and Dorizzi B 2012 Human Action Recognition using Continuous Hmms and HOG/HOF Silhouette Representation ICPRAM 503-508
[16] Huimin Q, Yaobin M, Wenbo X, Zhiquan W 2012 Recognition of human activities using SVM multi-class classifier *Pattern Recognition Letters* **31** 100-11

[17] Adithyan P, Bhargavi R, and Vaidehi V 2012 Abnormal Human Activity Recognition Using SVM Based Approach *Int. Conf. on Recent Trends in Information Technology* 97-102

[18] Fam B L and Jaward M H 2015 Spatio-Temporal Descriptor for Abnormal Human Activity Detection *14th IAPR Int. Conf. on Machine Vision Applications (MVA)* 471-474

[19] Tehmina K, Anwar S M, Majid M, Khan B, Ali S M 2018 Emotion recognition from facial expressions using hybrid feature descriptors *IET Image Process* **12** 1004-1012

[20] Gao Z, Xuan H Z, Zhang H, Wan S and Choo K K R 2019 Adaptive fusion and category-level dictionary learning model for multiview human action recognition *IEEE Internet of Things Journal* **6** 9280-9293

[21] Liu M, Li, H and Chen C 2017 Enhanced skeleton visualization for view invariant human action recognition *Pattern Recognition* **68** 346-362

[22] Vishwakarma D K 2020 A two-fold transformation model for human action recognition using decisive pose *Cognitive Systems Research* **61** 1-13

[23] Parameshachari BD, Kiran Rashmi P, Supriya MC, Rajashekarappa and Panduranga HT 2019 Controlled partial image encryption based on LSIC and chaotic map. *In ICCSP* 60-63