Human Interactions Recognition using Bag of Words

R.Newlinshebiah, S.P.Sivasubbu, V.Sivasankar

Abstract— **A video surveillance system can be defined as a technological tool that assists humans by providing an extended perception and capability of capturing interesting activities in the monitored scene. This paper describes a methodology for automated recognition of one to one human interactions such as handshake, kicking and hugging. The frame work consists of background subtraction followed by feature extraction (Speed Up Robust Features) and action classification using SVM classifier. It is computationally efficient and invariant to occlusion, lightning. The method produces good categorization accuracy and precision. Human behaviour recognition has various applications such as human-computer interfaces; content based video retrieval, Visual monitoring & surveillance.**

Index Terms— **Human Action Recognition, Bag of Words, SURF Features.**

I. INTRODUCTION

Recognizing the activities of human from video sequence is gaining more attention in applications of computer vision such as surveillance and security agencies. This paper presents bag of words approaches and support vector machine classifier for the representation and recognition of one to one human activities in video sequence. Here the term 'Interactions' refers to the complex sequence of actions performed by two humans who could be interacting with each other in a constrained manner. These interactions are typically characterized by analysing the video frames to know about the interactions and use the training set to identify similar kind of actions. In recognizing human activities, the shape of the human silhouette plays a key role and it can extract from background subtraction blobs.

Background subtraction is a used to detect moving objects in videos from static cameras. Technique used for this process is mixture of Gaussian. In short, the Gaussian mixture model can be effectively modified to describe more complex background. The methods are proposed to build more complicated background models other than GMM. One of the typical methods is that the background model is build as a statistical model. In [1] each pixel is modelled by four parameters, brightness distortion, chromaticity distortion, the variation of the brightness distortion, and the variation of the chromaticity distortion. Brendel *et al.*[2]propose a segmentation method by tracking regions across frames with a new circular dynamic-time warping (CDTW) algorithm, which generalizes the conventional DTW algorithm to match closed boundaries of two regions, for region matching to identify the longest, best matching boundary portion of two regions. The second stage is feature extraction and representation, and the important characteristics of image frames are extracted and represented in a systematical way as features. The algorithm used in this to extract the features of object of interest is Speeded-Up Robust Features (SURF).At first, bag of words model used SIFT to extract the feature from the object which can be referred in [3].The major difference between SURF and SIFT is that SURF descriptor has 128 dimensions while SIFT descriptor only has 64 dimensions and its comparisons can be referred in [4].

In [11] have stated that SURF outperforms SIFT in terms of result and computational time, thus we chose SURF instead of SIFT as our feature extractor. There are also other methods like HOG descriptors, Dalal and Triggs [5] were the first to propose histogram of oriented gradient (HOG) descriptors for human detection. The HOG is derived based on evaluating normalized local histograms of image gradient orientations by considering fine-scale gradient in a dense grid, relatively coarse spatial binning, and fine orientation binning and high-quality local contrast normalization in overlapping descriptor blocks. This is termed as the process of retrieving images on the basis of low-level image features, provided a query image or manually constructed description of these low-level features.

These descriptions have a little relation to the semantic content of the image. Taking the distance information and the width feature of a silhouette, Lin et al. [6] propose a new feature, called nonparametric weighted feature extraction (NWFE), which uses the nearest neighbour classifiers to build histogram vectors for human activity recognition. Lucas-Kanade [7] and Tomasi [8] propose a point tracking method based on the sum of squared intensity differences, named LKT (Lucas-Kanade-Tomasi) feature tracker.
The paper is structured as follows. The next section discusses the Proposed Method. The Section 3, gives the Recognition Results and Discussion. Finally, Section 4 gives the Concluding remarks of the proposed method.

II. PROPOSED METHODOLOGY

The methodology consists of background subtraction of the input video followed by feature extraction and finally the SVM classifier classifies them to various classes such as handshake, kicking and hugging.

R. Newlin Shebiah, Department of ECE, Mepco Schlenk Engineering College, Sivakasi (E-mail: newlinshebiah@yahoo.co.in).
S.P.Sivasubbu, Department of ECE, Mepco Schlenk Engineering College, Sivakasi.
V.Sivasankar, Department of ECE, Mepco Schlenk Engineering College, Sivakasi.

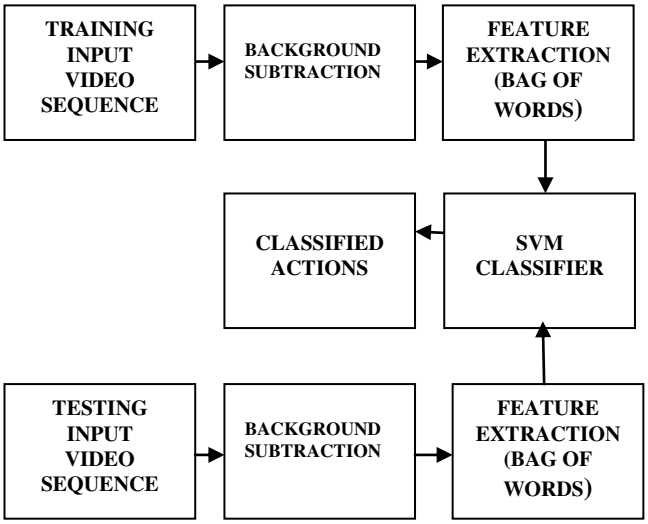


Fig. 1 : Block diagram of Proposed Method

A. *Background subtraction:*

In recognizing human activities,the shape of the human silhouette plays a key role and it can extract from background subtraction blobs.Background subtraction is a widely used technique for detecting moving objects in videos from static cameras.In background subtraction method,we isolate a moving object as foreground from background. For example,the human's interaction activities like handshake ,hugging , kicking can be easily perceived from the background subtracted images.Since background subtraction is a low-level task, two aspectsshould be considered: accuracyand computational resources (both time and memory) the output of the background subtraction is used for other high-level tasks, such as tracking and recognition. For such recognition tasks, the performances affect the erroneous output. Second, computational resourcesused for background subtraction are very crucialbecause the resourcesremaining after this low-level task should be used for highleveltasks, and is preferable as a means of implementing thistask in real-time such as staticip cameras .Therefore, it is important for the background subtractionmethod to obtain high accuracy and low resource requirements.So for this reason we have used open cv platform to perform background subtraction using mixture of gaussian.This method is applicable mostly to RGB color images and grayscale images.To perform background subtraction process,pixelvaue for each current frame and static background frame needed to be calculated. The probability of observing a pixel value, X_t, at time t as follows:

$$P(X_t) = \sum_{i=1}^{k} \omega_{i,t},\eta\,(X_t, \mu_{i,t}, \Sigma_{i,t}) \quad \text{-----} \quad (1)$$

where$K$ is the number of Gaussians, which is default value is between 3 and 5. $\omega_{I,t}$, $\mu_{i,t}$, and $\Sigma_{i,t}$are weight, mean, and the covariance matrix of the i^{th} Gaussian in the mixtureat time t, respectively.K-means approximation is preferred to update this model.Every new pixel value is checked against the K Gaussian distributions to determinewhether this value is within 2.5 standard deviation. If none of the distributions includes this pixel value, theleast probable distribution is replaced with a distribution whose mean, variance, and weight are set to the current pixel value, predetermined high variance, and low weight, respectively. To update the k distributions weight the following function is used:

$$\omega_{k,t} = (1 - \alpha)\,\omega_{k,t-1} + \alpha M_{k,t} \text{--------} \quad (2)$$

whereα is a learning rate, and $M_{k,t}$is 1 for the distributionwhich includes the current pixel value within its 2.5 standard deviation and 0 for the other distributions.The summation turns 1 to renormalize ,after updating the weights. The parameters of the distribution which includes the current pixel value within its 2.5 standard deviation are updated as follows:

$$\mu_{k,t} = (1 - \rho)\mu_{k,t-1} + \rho X_t \text{-------} \quad (3)$$

$$\sigma^2_{k,t} = (1 - \rho)\,\sigma^2_{k,t-1} + \rho(X_t - \mu_{,t})^T(X_t - \mu_{k,t}) \text{ -------} \quad (4)$$

whereρ is a learning factor for adapting distributions. Theparameters of the other distributions remain the same. Todecide whether $X_t$is included in the background distributions,
the distributions are ordered by the value of $\omega_{k,t}/\sigma_{k,t}$andthe first B distributions which satisfy (5) are chosen as thebackground distributions as follows:

$$B = \arg_b \min\,(\sum_{i=1}^{k} \omega_{k,t} > T) \quad \text{-------} \quad (5)$$

where$T$ is a measure of the minimum portion of the data that should be accounted for by the background. If $X_t$is within 2.5 standard deviation of one of these B distributions, it is decided as a background pixel.This methodology is discussed detailed in [9].

B. *Feature extraction :*

The algorithm we used to extract the features of object of interest is Speeded-Up Robust Features (SURF)..Generally, the BoW consists of3 main steps :(for Reference see [10]:

- SURF automatically detects the region of interest points from the object.
- Visual dictionary is created by quantizing the keypoints and descriptors.
- Finally, bag of words histogram is constructed by finding the occurances of each visual word.

Functions of SURF are explained as follows:
Speed-up computations by fast approximation of
(i) Hessian matrix and
(ii) Integral images
SURF algorithm is based on the Hessian Matrix to detect interest points. A Hessian matrix in 2-dimensions consists of a 2 x 2 matrix containing the second-order partial derivatives as follows:

$$H(f(x,y)) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} \quad \text{----------} \quad (6)$$

For the Hessian matrix, the eigenvectors form an orthogonal basis showing the direction of curve(gradient) of the image.
The descriptor uses a distribution of Haar-wavelet responses around the interest point's neighborhood.The next step is to

find major interest points in scale space using non-maximal suppression which can be referred in [6].

Because of the coarse scale of the scale space, we need to interpolate the interest point to arrive at the correct scale (σ)usingtaylorexpansion.Next step is to find the direction of rotationally invariant features.We use Haar Transforms to assess the primary direction of the feature. The intuition is that they give you a sense of the direction of the change in intensity. They are resistant to overall luminance changes .For each feature it Looks at pixels in a circle of 6*σ radius and Compute the x and y Haar transform for each point Use the resulting values as x and y coordinates in a Cartesian map. Weight each point with a Gaussian of 2*σ based on the distance from the interest point. Rotate a wedge of π/3 radians around the circle and Choose direction of maximum total weight .A square descriptor window is constructed with a size of 20*σ centered on each interest point and orientation based on the derived rotation .Divide the descriptor window into 4 x 4 sub-regions .Each sub-region is 5*σ square and Haar wavelets of size 2*σ are computed for 25 regularly spaced points in each sub-region dx and dy are computed at each point in the rotated direction .feature vectors can be computed for each of the 16 sub-regions we compute 4 values Sum of dx ,Sum of dy,Sum of abs(dx) and Sum of abs(dy) Feature vector is a 64 dimensional vector consisting of the above 4values for each of the 16 sub-regions.The final step for the BoW model is to convert vector represented patches to "codewords" which also produces a dictionary. A codeword can be considered as a representative of several similar patches. One simple method is performing k-means clustering over all the vectors. Codewords are then defined as the centers of the learned clusters. The number of the clusters is the dictionary size .Thus, each patch in an image is mapped to a certain codeword through the clustering process and the image can be represented by the histogram of the codewords.These histogram image is fed to svm classifier as input.

C. *Action Classification:*

The Support Vector Machine is a theoretically superior machine learning methodology with great results in classification of high dimensional datasets.SVMs have often been found to provide better classification results that other widely used pattern recognition methods, such as the maximum likelihood and neural network classifiers.Most of the practical applications involve multiclass classification, especially in object classification. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.Originally, SVMs were dev-+eloped to perform binary classification. However, applications of binary classification are very limited especially in object classification where most of the classification problems involve more than two classes. A number of methods to generate multiclass SVMs from binary SVMs.

III. RESULTS AND DISCUSSIONS

Testing activity recognition algorithm is essential as it providesqualitative and quantitative performance analysis. silvanip camera to capture the video sequence.The images used in this work are 720 ×480 pixels in size, obtained at a rate of 22 frames/sec.The dataset is created for three activities(Handshake,Kicking,Hugging)and its input video sequences are as follows :

(a) Video Sequence showing Handshaking

(b) Video Sequence showing Kicking

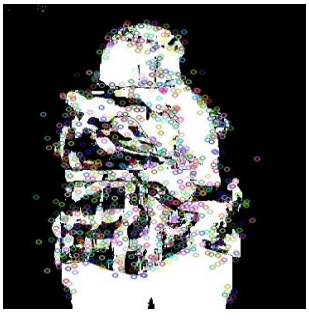
Fig 2: Sample Video Sequence from the Database

This input video sequence is fed to background subtraction process for background elimination and foreground detection. The results for this process are as follows :

Fig 3: Background Subtracted Output

In the feature extraction technique, the important characteristics of foreground image frames are extracted and key points extracted for one the image is shown below:

(a) Keypoints from Kicking



(b) Keypoints from Hugging

(c)Keypoints from Handshaking

Fig 4: Keypoints Extracted from the Video Sequence

The training dataset consists of 2 videos for each action.i.e.2 videos for handshake,2 videos for hugging,2 videos for kicking and the actions were performed by different persons.Each video consists 15 frames.
The testing dataset consist of a single video for each action and the persons in the video are different from persons in the training dataset.Each video consists of 10 frames.

Table 1: Recognition Rate of the Proposed Human Interaction Detection Algorithm

Action	No. of Samples used		Classification Rate (%)
	For Training	For Testing	
Handshake	2 videos (15 frames each)	1 video-10 frames	97
Kicking	2 videos (15 frames each)	1video-10 frames	98
Hugging	2 videos (15 frames each)	1video-10 frames	98

So far we have classified activities hugging, handshake and kicking with less error rate. It is also possible to recognize other interaction like kissing, punching, etc.

IV. Conclusion

This model gives a method for automatic recognition of three activities handshake, kicking and hugging. The proposed algorithm for one-to-one human interaction detection involves the saliency based approach, where high recognition rate is obtained but the compuptional complexity is high. The same method can be extended to detect several other actions and interactions. The system can be trained with more number of video sequences and more number of interactions can be involved.

References

[1]Horprasert, T.; Harwood, D.; Davis, L.S.A statistical approach for real-time robust background subtraction and shadow detection.*IEEE ICCV* 1999, *99*, 1–19.

[2]Brendel, W.; Todorovic, S. Video Object Segmentation by Tracking Regions. In proceedings of IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 833–840

[3] David G. Lowe "Object Recognition from LocalScale-Invariant Features," The Proceedings of the Seventh IEEE International Conference on ComputerVision,Vol. 2,pp.1150-1157,1999.

[4]Maya Dawood,CindyCappelle,Maan E.EI Najjar , Mohamad Khalil, Denis Pmorski, "Harris, SIFT, and SURF
Features Comparison for Vehicle Localization based on Virtual 3D Model and Camera," 3'd International Conference on Image Processing Theory, Tools, and Applications (IPTA),pp. 307;312,October 2012.

[5]Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 886–893.

[6] Lin, C.; Hsu, F.; Lin, W. Recognizing human actions using NWFE-based histogram vectors, Eurasip Journal on Advances in Signal Processing - EURASIP J ADV SIGNAL PROCESS , vol. 2010, pp. 1-16, 2010.

[7] Lucas, B.D.; Kanade, T. An Iterative Image Registration Technique with An Application to Stereo Vision. In Proceedings of the 7th International Joint Conference on Artificial Intelligence, Vancouver, B.C., Canada, 24–28 August 1981 .

[8] Shi, J.; Tomasi, C. Good Features to Track. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 21–23 June 1994; pp. 593–600 ,*EURASIP J. Adv. Signal Process.* 2010.

[9] Stauffer, C. and Grimson, W.E.L,Adaptive Background Mixture Models for Real-Time Tracking, Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, Vol. 2 (06 August 1999), pp. 2246-252 Vol. 2.

[10]Teng Li, Tao Mei; In-So Kweon ; Xian-Sheng Hua, "Contextual Bag-of-Words for Visual Categorization"Published in Circuits and Systems for Video Technology, IEEE Transactions on (Volume:21 , Issue: 4)

[11]Herbert Bay, Andreas Ess, TinneTuytelaars, Luc Van Gool, "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346--359, 2008.