# Human Action Recognition In Videos using Bag of Words (BoW) Model

Deogratias N. LUKAMBA[1], Sadiq MACAULAY O.[1], Abdul Salam Rasmi ASRAF ALI[1], Dro-Desire SIDIBE[2]

***Abstract*— In this project, we tried to implement the recognition of human action in video sequences using the Bag of Words paradigm. We described how efficiently the algorithms: 3D SIFT, K-Means and Support Vector Machine(SVM) are implemented to replace previously used description methods. We tried to improve the classification of some similar and correlated actions like jogging-running and clapping-waving hands.**

***Keywords: Bag of Words(BoW), 3D-SIFT, K-Means, SVM***

## I. Introduction

Human Action Recognition (HAR) aims to classify human actions performed in a scene and track them throughout the video sequence. In computer vision, Human Activity Recognition is one of the most integral parts of scene understanding. From time being, it is still one of the most challenging research fields of Computer Vision because of it is wide range of importance in video surveillance, health care, human-computer interaction and many other.

We made a review of the state-of-art approaches, especially in activity representation and classification. We intend to design a system that is more efficient for real-time human activity recognition.

In this project, We train and test our algorithm with the KTH[1] human motion dataset. This data set contains different kinds of human actions like walking, jogging, running, boxing, waving and clapping. Our implementation is in MATLAB.

## II. Background

### A. Bag of Words

In Recent days Bag of Words(BoW) (also called Bag of Features or Histogram of Features), has been widely used by the Computer Vision Community specially for its accuracy. We acquire a distribution which resembles a histogram in which we count the number of occurrence of a particular feature. This method has a property of "order-less meaning", i.e. the order is not preserved whenever we find the histogram of features.

### B. Harris Corner Detector

The Harris Corner Detector is a mathematical operator that finds features in an image. It is simple to compute, and is fast enough to work on computers. Also, it is popular because it is rotation, scale and illumination variation independent[3]. Mathematically, Harris Detector is given by,

$$E(u,v) = \begin{bmatrix} u & v \end{bmatrix} \left( \sum_{(x,y)\in W} \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \right) \begin{bmatrix} u \\ v \end{bmatrix}$$

where,

- $E$ is the difference between the original and the moved window $W$.
- $u$ is the window's displacement in the x direction
- $v$ is the window's displacement in the y direction
- $I_x$ is the $1^{st}$ order derivative of intensity value $I$ w.r.t. $x$
- $I_y$ is the $1^{st}$ order derivative of intensity value $I$ w.r.t. $y$
- $I_x^2$ is the $2^{nd}$ order derivative of $I$ w.r.t. $x$
- $I_y^2$ is the $2^{nd}$ order derivative of $I$ w.r.t. $y$

### C. 3D SIFT Descriptor

The 3D SIFT descriptor encodes the information in both space and time in a manner which allows for robustness to orientations and noise. In addition, after describing the videos as a bag of spatio-temporal words using the proposed SIFT descriptor, we discover relationships between words to form spatio-temporal word groupings.

The 3D SIFT descriptor is able to robustly describe the 3D nature of the data in a way that vectorization of a 3D volume can not. Using sub-histograms to encode local time and space information allows 3D SIFT to better generalize the spatio-temporal information than features used in previous works[4].

The Gradient magnitude $m_{3D}(x,y,t)$ and the Orientation $\phi(x,y,t)$ of the SIFT (in 3D) is given by,

$$m_{3D}(x,y,t) = \sqrt{L_x^2 + L_y^2 + L_t^2}$$

$$\phi(x,y,t) = \tan^{-1}\left(\frac{L_t}{\sqrt{L_x^2 + L_y^2}}\right)$$

where, $(L_x, L_y, L_t)$ are the spatio-temporal gradient in 3D $(x,y,t)$ [4]. A representation of how SIFT algorithm is adapted for Video Frames has been illustrated in figure 1 which shows the 3D SIFT descriptor with its 3D sub-volumes, each sub-volume is accumulated into its own sub-histogram. These histograms together makes up the final descriptor.

[1]Masters in Computer Vision, University of Burgundy, 71200 - Le Creusot, France

[2]Associate Professor - University of Burgundy, VIBOT ERL CNRS 6000/ ImViA EA 7535, Le Creusot, France

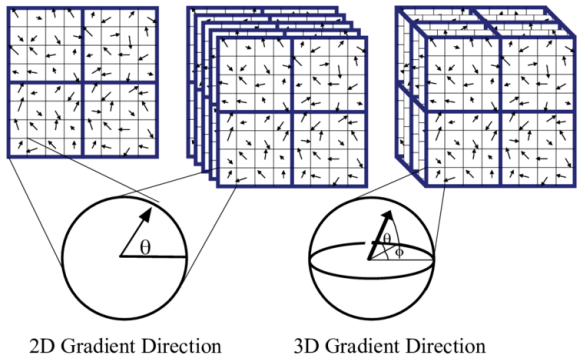


Fig. 1. 3D SIFT-An Extension of SIFT Descriptor for Video Frames

### D. K-means Clustering

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

- The centroids of the K clusters, which can be used to label new data
- Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically[5].

### E. Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimentional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side[6].

## III. Methodology

In this project, we propose to achieve a better result of classification when comparing to the co-occurrence nature of clusters words method. The idea is to use the k-means clustering technique to create a spatio-temporal word vocabulary (clusters). Signature(or Patterns) of histogram representation of videos is created using 3-D SIFT descriptors to match with each Word vocabulary. Finally, the histograms are grouped based on the co-occurrence of Word vocabulary which is a simple matrix showing how many times a particular word occurs. This approach is implemented with the Bag of Words Model [9] [10] [11] [12].

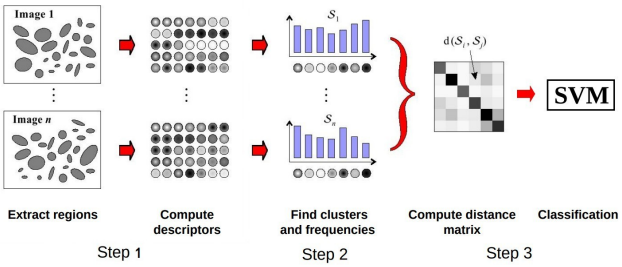


Fig. 2. Process Flow

A co-occurrence matrix is created to signify the number of occurrence of a particular word among all other words. Row vectors of this matrix represent contextual distribution of a word in terms of other words of the vocabulary. Similar contextual distribution vector of any two words signifies that these two words frequently occurs together. The measure of similarity is computed by correlation of vectors, i.e. if the correlation of two vectors is above a certain threshold, then those two words are joined and their corresponding frequency from initial histograms are added, thus creating new histograms for videos[13].

The process flow of the BoW model for Action Recognition is shown in fig. 2. The Video input is converted into a matrix using a MATLAB function video2mat.m. In the next step interest points are extracted followed by computing the descriptor as explained before. This process is repeated for several videos, $v_1, v_2, \cdots, v_n$. Finally, we train a classifier using SVM to recognize whether the action is **Boxing, Clapping, Jogging, Running, Waiving, Walking**.

The Methodology can be grouped into three major tasks.

### A. Feature Extraction

To extract features for classification we have implemented a Harris Corner Detector which extract the Interest points followed by 3D-SIFT to compute the descriptors for the Interest points, so that we will have one descriptor for each Interest points.

### B. Quantization

The computed descriptors will be in a high dimensional space (with multiple classes), and so we cluster them to create a word for each class which forms our vocabulary. The clustering is achieved using the most simple yet effective method, K-Means with $K = 6$. Here the value of $K$ denotes the number of actions to be classified. Once clustering is done, each cluster will become a word for which we will build frequency histogram (bag) of each input video with respect to those words.

### C. Classification

Once we have the Bags of Features, we train a classifier implemented using SVM, to classify the different actions in videos. In the training phase, Bag of features, represented as one Vector for each video, is provided to the classifier with a Label for the action happening in that particular video. In the

testing phase the classifier detects the label (Human action) for the given input video.

## IV. Results and Discussions

Our implementation was mainly based on the actions in the videos of KTH dataset[1] which employed 600 videos of people performing 6 different actions like: **Boxing, Clapping, Jogging, Running, Waiving, Walking**. A sample output is shown in fig. 3.


Fig. 3. Boxing action recognition shown in GUI

Some of the actions in this database are very closely related like jogging-running, and waiving-clapping. These are the actions that are often misclassified affecting the accuracy of the model. But in our case we achieved results with greater accuracy which is represented in the form of a confusion matrix shown in the table below.

TABLE I
Confusion Matrix

| | Boxing | Clapping | Jogging | Running | Waving | Walking |
|---|---|---|---|---|---|---|
| Boxing | 1.00 | | | | | |
| Clapping | | 0.90 | | | 0.10 | |
| Jogging | | | 0.90 | 0.10 | | |
| Running | | | 0.10 | 0.90 | | |
| Waving | | | | | 0.90 | 0.10 |
| Walking | | | | | | 1.00 |

## V. Conclusion

The accuracy of the BoW model mainly depends on two factors, one is the number of descriptors considered per video and the other is the co-occurrence nature of the cluster words. We used 100 descriptors per video which increased the accuracy upto 93% at the expense of time complexity. This algorithm can be further improved by implementing Convolution Neural Networks since can have a higher probability of detecting co-occurrence nature of words.

## References


[1] Schldt, C. and Laptev, I. and Caputo, B., Recognizing Human Actions: a Local SVM Approach,Proc. Int. Conf. Pattern Recognition (ICPR'04), Cambridge, U.K, 2004.

[2] Scovanner, Paul and Ali, Saad and Shah, Mubarak, A 3-Dimensional SIFT Descriptor and Its Application to action Recognition, Proceedings of the 15th ACM international conference on Multimedia, 2007, pg. 357 - 360.

[3] Fundamentals of Features and Corners: Harris Corner Detector - AI Shack, URL - http://aishack.in/tutorials/harris-corner-detector/

[4] Scovanner, Paul and Ali, Saad and Shah, Mubarak, A 3-dimensional sift descriptor and its application to action recognition, Proceedings of the 15th ACM international conference on Multimedia, 2007.

[5] Introduction to K-means Clustering, URL - https://www.datascience.com/blog/k-means-clustering

[6] Savan Patel, Chapter 2 : SVM (Support Vector Machine) Theory Machine Learning 101 Medium, URL - https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

[7] Bhatti, Naeem A and Hanbury, Allan, Co-occurrence bag of words for object recognition, Proceedings of the 15th Computer Vision Winter Workshop, Citeseer, 2010, pg. 21 - 28.

[8] Zhang, Yimeng and Liu, Xiaoming and Chang, Ming-Ching and Ge, Weina and Chen, Tsuhan, Spatio-temporal phrases for activity recognition, European Conference on Computer Vision, Springer, 2012, pg. 707 - 721.

[9] Cózar, Julián Ramos and González-Linares, José María and Guil, Nicolás and Hernández, R and Heredia, Y, Visual words selection for human action classification, 2012 International Conference on High Performance Computing & Simulation (HPCS), IEEE, 2012, pg. 188 - 194.

[10] Niebles, Juan Carlos and Wang, Hongcheng and Fei-Fei, Li, nsupervised learning of human action categories using spatial-temporal words, International journal of computer vision, Springer, 2008, vol. 79-3, pg. 299-318.

[11] Yuan, Chunfeng and Li, Xi and Hu, Weiming and Wang, Hanzi, Human action recognition using pyramid vocabulary tree, Asian conference on computer vision, Springer, 2009, pg. 527 - 537.

[12] Nguyen, Thanh Phuong and Manzanera, Antoine, Action recognition using bag of features extracted from a beam of trajectories, 2013 IEEE International Conference on Image Processing, IEEE, 2013, pg. 4354 - 4357.

[13] Bag, Suvam and Kulhare, Sourabh, Human Activity Recognition in Videos, Department of Computer Engineering, Rochester Institute of Technology.