Impact of eye fixated regions in visual action recognition

Deepak Pathak
Dept. of Computer Science and Engineering
IIT Kanpur

Dr. Amitabha Mukerjee
IIT Kanpur

{deepakp,amit}@iitk.ac.in

April 18, 2013

Abstract

Computer vision techniques using machine learning have been quite successful in the task of action recognition, object classification and scene segmentation, but still way behind the human performance on equivalent tasks. Most of these techniques computationally model interest points in the input data which have less correlation with the visual system in living beings. In this project, we study the features derived out of "fixation and saccadic motions" of human gaze in the action classification task on Hollywood-2 dataset, thus bridging the gap between computer and human vision. We implement two visual recognition pipelines. First pipeline is based on standard bag of words model while in other we propose a novel approach based on array of words to capture global temporality. Results show that entire fixated region is not so useful in classifying general actions thus verifying the theory of "covert attention" [Land 2006] in humans. We further discuss the advantages of the suggested array of words approach.

1 Introduction

With the advancement in machine learning techniques, computer vision community has made a lot of progress in classification and segmentation tasks in still images as well as sequences. Most of the approaches utilised in these tasks have been orthogonal or unrelated to visual processing system in humans. In this paper, we study human gaze patterns in the video sequence and determine how useful they are in determining the action present in it.

There have been several attempts to model vision techniques similar to biological visual processing systems. Inspired from human visual system, [Poggio 2005] discusses hierarchical model for feature extraction based on simplex descriptors, which is in coherence with the ventral stream of visual cortex, which were then trained over discriminative classifier for object recognition tasks. But such approaches have not been able to outperform current standard techniques in computer vision for interest point detection, feature descriptor extraction and classification.[Laptev 2005][Klaser 2008]

[Mathe 2012] provides large dataset of human eye fixation and saccade coordinates on couple of large and useful datasets : Hollywood-2 and UCF Sports action dataset. In this project, we use these fixation coordinates over Hollywood-2 dataset to get gaze inspired interest points in the video sequences, thus utilising them in classifying the action classes in the dataset.

1.1 Overview of the paper

In this report, we initially describe the dataset that we have used over the project. In further section, we discuss our methodology to tackle the problem of action recognition and then provide detailed description of the approach. We discuss both, the standard approach and our proposed approach for implementing action recognition pipeline.

Then we discuss data specific implementation details and the results produced. Finally we conclude by discussing the different aspects of human attention and the achievements of our proposed approach.

2 Dataset

[Laptev 2009] provides the huge collection of clips from Hollywood movies broadly classifiable into 12 action classes namely : { Kiss, SitDown, HandShake, Eat, Run, SitUp, FightPerson, HugPerson, StandUp, DriveCar, GetOutCar, AnswerPhone }. This dataset is one of the naturalistic action dataset containing unconstrained clips. The dataset in all contains 1707 videos of 12 classes divided into 823 training samples while 884 test video clips.

[Mathe 2012] provides human gaze fixation data for Hollywood-2 and UCF Sports Action Dataset. The dataset contains 16 subjects in all (including both male and females). Out of these, 4 were asked to freely view the video were other 12 were given the task to identify action in the video. This dataset was recorded for 92hrs per subject with 500Hz sampling rate for gaze tracking. The gaze coordinates contain three types of data points "fixation", "saccade" and "blink".

Figure: Some action classes of Hollywood-2 Action Dataset. [Laptev 2009]

3 Methodology Used

Our aim is to use this gaze fixation data and get some useful feature for action classification. Our approach is to initially take these fixation coordinates as the interest points in video sequences. Then we compute temporal HoG feature descriptor around these interest points. HoG3D is extension of Histogram of Oriented Gradients to spatio-temporal space in three dimensions. They capture local temporality around the interest points along spatial locality. [Klaser 2008]

After this, we cluster the computed descriptors using K-means clustering algorithm. We then treat these obtained k clusters as visual words and then try to represent the video in terms of those. After this step we design two different pipelines for classification task. These pipelines are described as follows -

3.1 Pipeline 1 : Bag Of Words

This is the standard approach used in computer vision community for representing the videos for performing action recognition task. Here we represent all the descriptors of each video as a histogram over the k clusters. Then, we normalise this histogram using L1

normalisation technique. Finally, we have each video represented as **feature histogram** over the k visual words.

Then we train these videos using discriminative classifier: Support Vector Machine (SVM). For this purpose we have used one is to all SVM in 10-fold cross validation manner over 1707 videos.

3.2 Pipeline 2: Array of Words

This is the novel technique which we present in this project. Through this technique we try to preserve the global temporality in video sequences, which is inherent in the feature obtained by gaze fixation coordinates. Each video descriptor is mapped to the nearest visual word. Thus we represent each video as an array of visual words put in proper sequence. Now this array captures high-level temporality while HoG3D already captures local temporality.

Now each video is represented as array of visual words of different length over which we now cannot train normal classifiers. Since this feature set involves sequencing and variable length input, the best choice is to use Hidden Markov Model with first order Markovian Assumption (i.e. the output of next state depends only on current state). In the arrar, since each element is index of visual word then the dimension of each observation is 1 and its domain is between 1 to number of visual words in vocabulary.

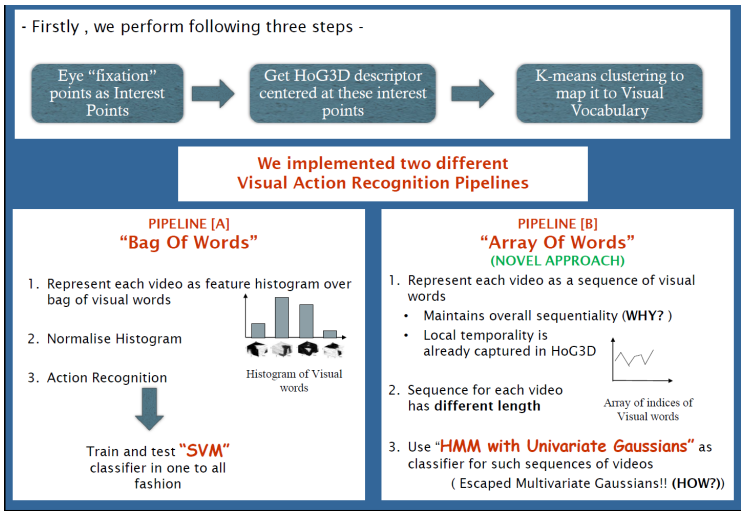


Figure: Summary of Approach

4 Implementation Details

Interest points are the eye gaze (F-fixation) coordinates of one subject with 12 frame overlap. Each video contains 1500-2000 descriptors each of dimension 300. K-means clustering involved mapping of 6 lac descriptors to 4000 word vocabulary (dimension=300). Each video :

- Feature histogram over 4000 bins [pipeline 1].
- Array of visual words [pipeline 2].

Learn classifier over the 1707 videos with 10-fold cross validation

Hurdles :

1. Coding from Scratch [Implementation Source was not available]
2. Computational Limitations: K-means clustering on such a large dataset was taking too much. So we used the unconverged clusters finally.

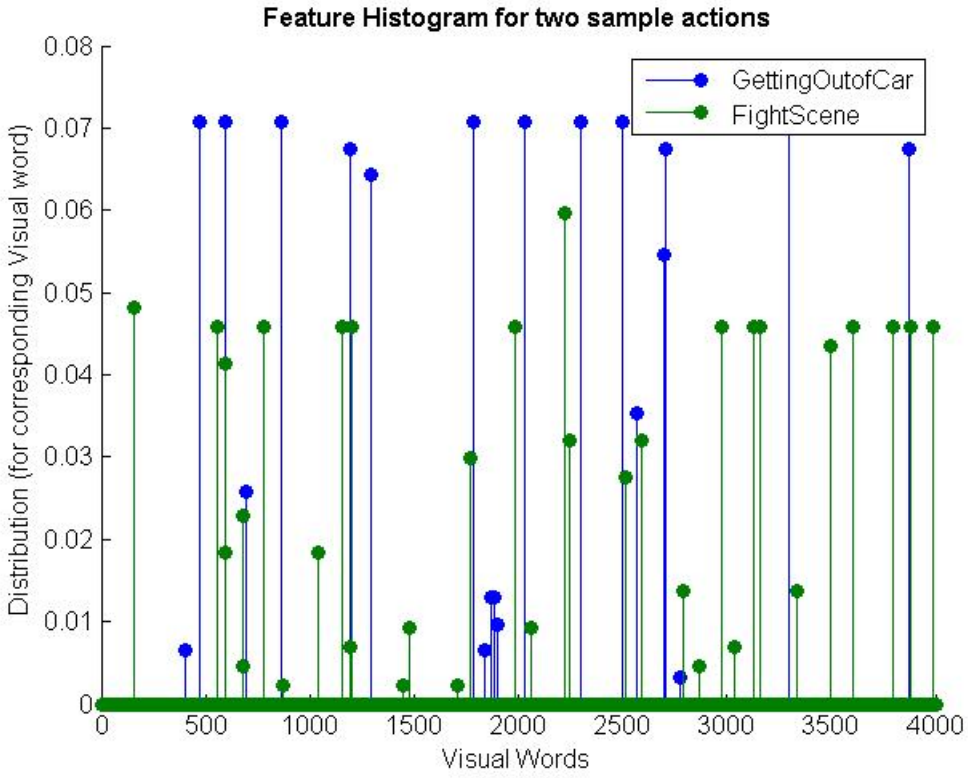
5 Results

- Figures showing HoG descriptor around eye gaze fixated centers.

Figure: Action is GetOutCar

Figure: Action is FightPerson

- Results from Pipeline-1 : Bag of Words



Feature Histogram representation of two sample sequences

One-all SVM
10 fold Cross Validation
Results ->

Dataset	Hollywood-2
Correct	384
Incorrect	1323
Accuracy	22.5 %
Harris Corners	49 % [Mathe]

Accuracy in percentage

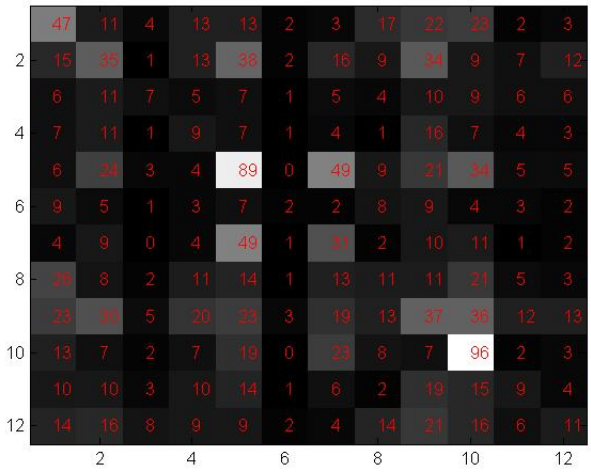
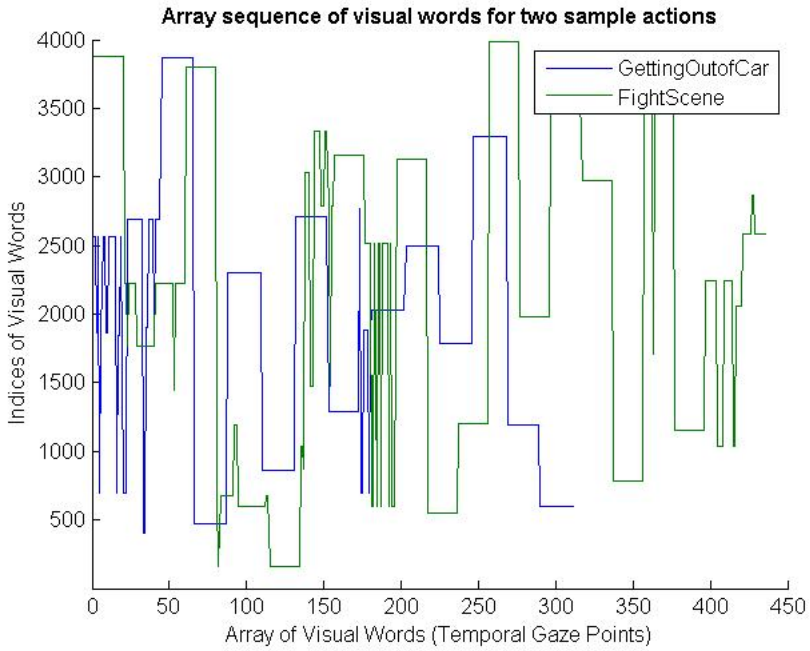


Figure: Confusion Matrix for classes in order: { Kiss, SitDown, HandShake, Eat, Run, SitUp, FightPerson, HugPerson, StandUp, DriveCar, GetOutCar, AnswerPhone }

- Results from Pipeline-2 : Array of Words



Basic HMM 6 states each,
10 fold Cross Validation
Results ->

Dataset	Hollywood-2
Correct	316
Incorrect	1391
Accuracy	18.5 %

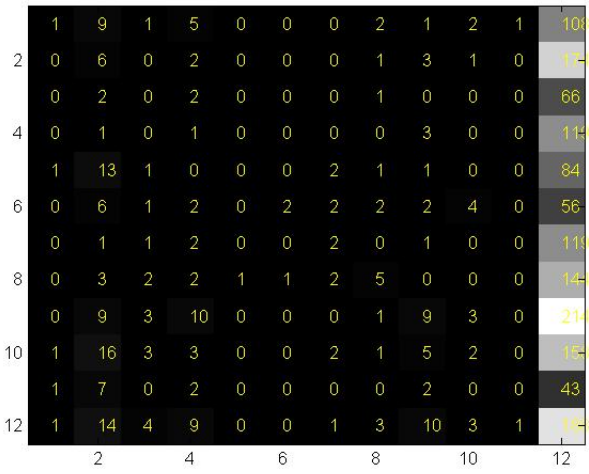


Figure: Confusion Matrix for classes in order: { Kiss, SitDown, HandShake, Eat, Run, SitUp, FightPerson, HugPerson, StandUp, DriveCar, GetOutCar, AnswerPhone }

6 Conclusion

Results show that features based on the eye gaze track patterns did not perform upto the expectations in both the approaches. These are way better than random but still much below the benchmarks.

Through this we conclude that the entire foveated region formed out of fixated points is not so useful, though its sub-region have potential interest points. This conclusion is in coherence with the theory of covert attention in humans according to which they can mentally focus on subset of the vast sensory input information in the form of different stimuli. Immense parallelism in brain makes covert attention possible quite easily. Conversely in Yarbus experiment, the attention was "overt" and the fixations are pretty useful in determining the context of objects in images. Moreover, determining general actions like running, walking etc. is intuitively simple and does not rely that much on visual fixations.

6.1 Achievements

We proposed novel "Array of Words" approach which has the following advantages -

- Capture high-level temporality in a given video sequence, and local temporality is maintained by the feature descriptor.
- Reduce usage of multivariate HMM to univariate basic HMM, as the dimension of each observation becomes 1 (index of corresponding visual word).
- It reduces the continuous real domain of observations to discrete value ranging from 1 to number of visual words in vocabulary.

7 Future Work

We can extend this approach to design Human Visual Saliency Predictor. Here we propose a small architecture for that :
By training binary classifier over feature descriptor.
Input: HoG3D feature detector around each pixel of the video data.
Output: Yes or No (being salient)
One of the problem that this approach can face is that its computationally intensive.

References

[1] Mathe, Stefan, and Cristian Sminchisescu. "Dynamic eye movement datasets and learnt saliency models for visual action recognition." Computer Vision-ECCV 2012. Springer Berlin Heidelberg, 2012. 842-856.

[2] Mathe, Stefan, and Cristian Sminchisescu." Actions in the eye: dynamic gaze datasets and learnt saliency models for visual recognition." Technical report, Institute of Mathematics of the Romanian Academy and University of Bonn (February 2012), 2012.

[3] Laptev, Ivan. "On space-time interest points." International Journal of Computer Vision 64.2 (2005): 107-123.

[4] Klaser, Alexander, and Marcin Marszalek. "A spatio-temporal descriptor based on 3D-gradients." (2008).

[5] Land, Michael F. "Eye movements and the control of actions in everyday life." Progress in retinal and eye research 25.3 (2006): 296-324.

[6] Marszalek, Marcin, Ivan Laptev, and Cordelia Schmid. "Actions in context." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.

[7] Rodriguez, M.D., Ahmed, J., Shah, M., "Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition," Computer Vision and Pattern Recognition, 2008.