Impact of Eye Fixated Regions in Visual Action Recognition

SE367: Introduction to Cognitive Science

By: Deepak Pathak
deepakp@iitk.ac.in

Guide: Dr. Amitabha Mukerjee
amit@iitk.ac.in

Introduction

Why useful :

- Content Based Browsing
- Annotating the videos
- Video Surveillance
- Patient Monitoring etc..

Issues

- Diversity in actions like sitting, running, jogging, walking etc.
- Occlusions
- Reflection
- Shadow , Background Clutter

Motivation

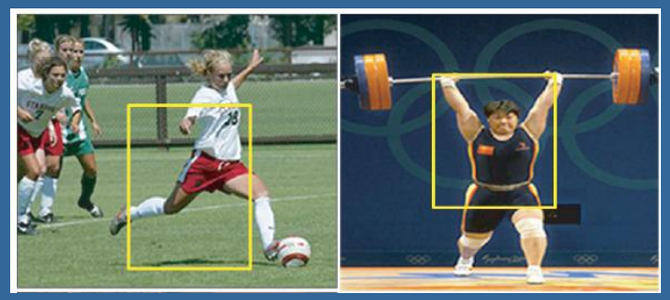
"Computer vision techniques still lag significantly behind human performance on similar tasks"

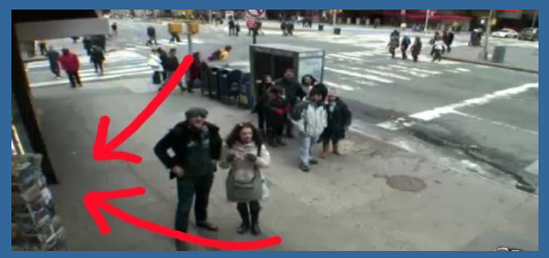
- bridging the gap..!!

YouTube Broadcast Yourself

150,000 uploads every day !

[UCF Sports Action Dataset]





[Live Snapshot – earthcam.com]

Objective of the project ..

Study human gaze patterns in videos and utilize them

- In activity recognition task.
- Human visual saliency prediction (**Next Phase !!**)

Targeted question ..

How **useful / interesting** are the eye-fixated points in a visual sequence in determining the action present in it ?

Overview of Project

Expected Result ..

The points {(x,y,t) – coordinates} where humans fixate their eyes should **not** be so useful (when taken as interest points) in the task of action recognition in natural scenes. **Why ?**

Datasets ..

[Mathe 2012]

- Provided large human eye tracking dataset recorded in context of dynamic visual action recognition tasks.
- 16 Subjects : Free viewing – 4 subjects ; Action Recognition – 12 subjects
- 92 subject-video hours, 500Hz sample rate.

Hollywood-2 Dataset [Laptev]

- Realistic human actions in unconstrained video clips from Hollywood movies.
- Naturalistic Action Dataset
- 12 Action Classes, 1707 video clips

Experimental Setup [Mathe 2012]

[Hollywood-2 Dataset]

Our Approach

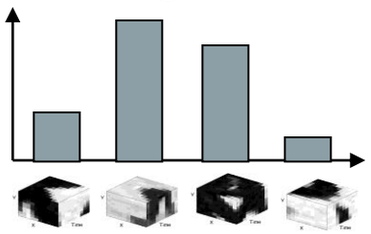
- Firstly , we perform following three steps -

Eye “fixation” points as Interest Points → Get HoG3D descriptor centered at these interest points → K-means clustering to map it to Visual Vocabulary

We implemented two different Visual Action Recognition Pipelines

PIPELINE [A]
“Bag Of Words”

1. Represent each video as feature histogram over bag of visual words
2. Normalise Histogram
3. Action Recognition

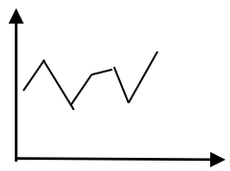


Histogram of Visual words

Train and test **“SVM”** classifier in one to all fashion

PIPELINE [B]
“Array Of Words”
(NOVEL APPROACH)

1. Represent each video as a sequence of visual words
   - Maintains overall sequentiality (**WHY?**)
   - Local temporality is already captured in HoG3D
2. Sequence for each video has **different length**



Array of indices of Visual words

3. Use “**HMM with Univariate Gaussians**” as classifier for such sequences of videos

(Escaped Multivariate Gaussians!! **(HOW?)**)

Implementation Details

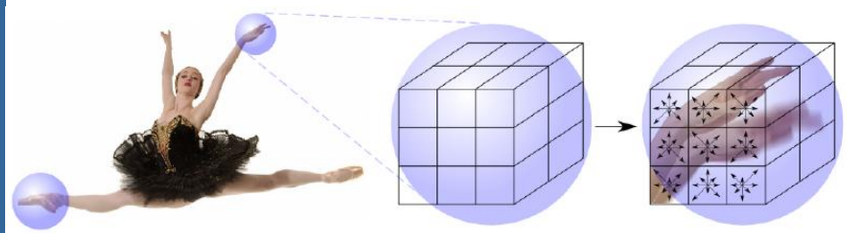
- Interest points – Eye gaze ('F'-fixation) coordinates of one subject with 12 frame overlap. (computational reasons)
- K-means clustering :
  - mapping of **6 lac descriptors to 4000 word** vocabulary (dimension=300)
  - each video :
    > normalized histogram of 4000 bins (pipeline 1)
    > array of indices of visual words (pipeline 2)
- Learn **Classifiers** over 823 training videos feature histogram. Test over 884 test videos. (Tool used – Weka)

Hurdles !!

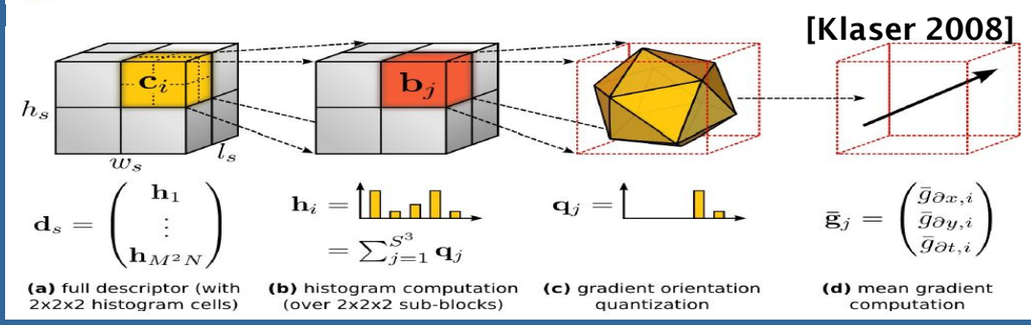
- Coding from Scratch [Implementation Source was **not** available]
- Computational Limitations –
  - K-means clustering on such a large dataset was taking too much.
    So we used **un-converged** clusters finally.

HoG3D Descriptor :

This involves Gradient computation and Orientation binning. Gradient computation requires filtering the image with the kernels [-1,0,1] and [-1,0,1]'



[CVPR '08]



[Klaser 2008]

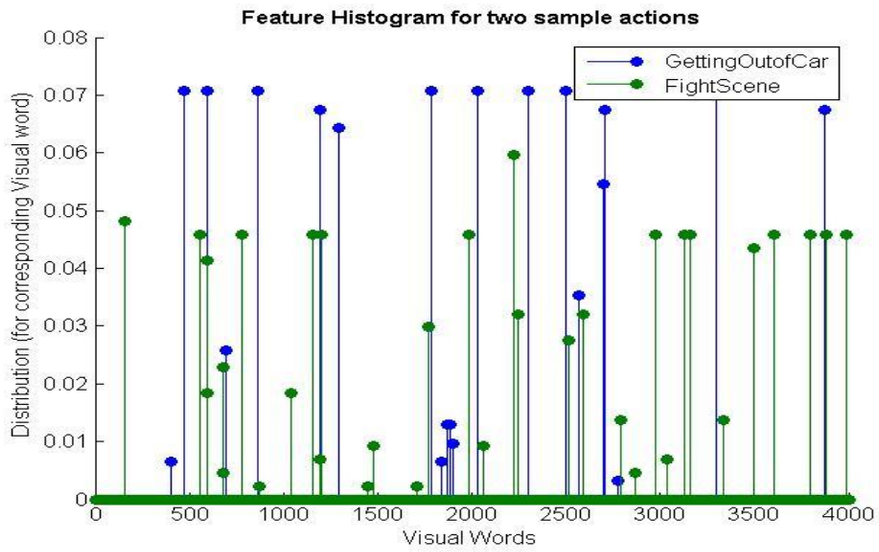
Results

Frame showing Hog descriptors around Interest point (From Eye Gaze data) Action – GetOutCar

Action – FightPerson

Results

Results - PIPELINE [A]
"Bag Of Words"

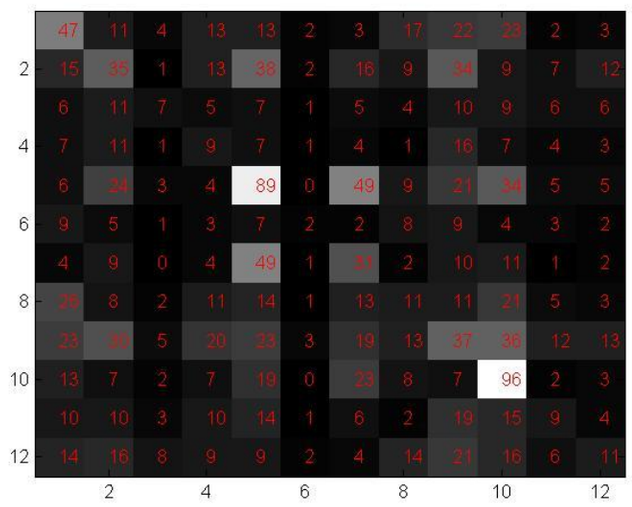


One-all SVM
10 fold Cross Validation
Results ->

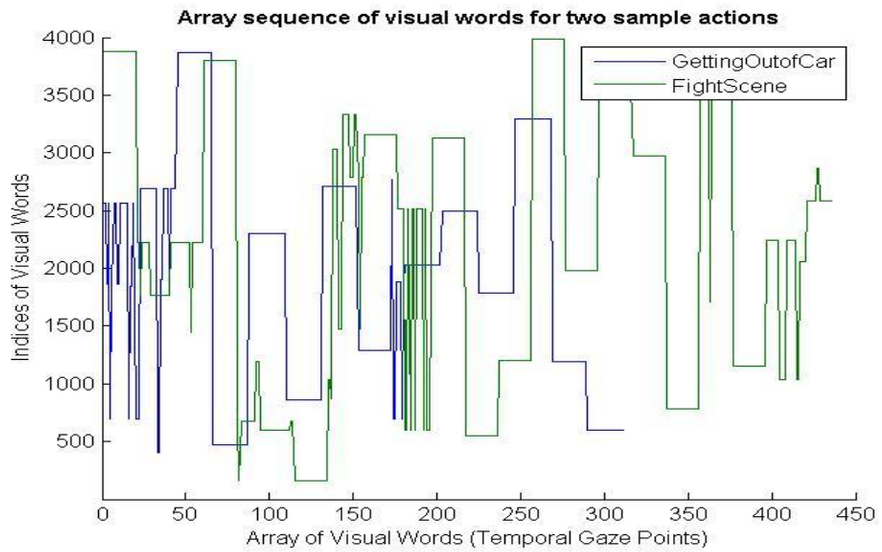
Dataset	Hollywood-2
Correct	384
Incorrect	1323
Accuracy	22.5 %
Harris Corners	49 % [Mathe]

Confusion Matrix

Actions in order : [Kiss, SitDown, HandShake, Eat, Run, SitUp, FightPerson, HugPerson, StandUp, DriveCar, GetOutCar, AnswerPhone]



Results - PIPELINE [B]
"Array Of Words"

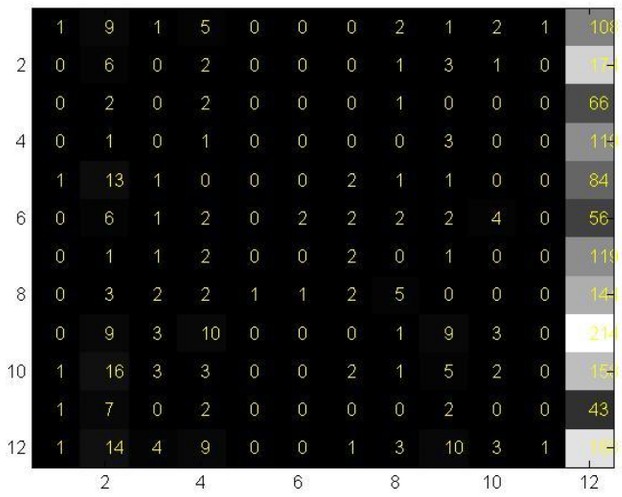


Basic HMM 6 states each,
10 fold Cross Validation
Results ->

Dataset	Hollywood-2
Correct	316
Incorrect	1391
Accuracy	18.5 %

Confusion Matrix

Actions in order : [Kiss, SitDown, HandShake, Eat, Run, SitUp, FightPerson, HugPerson, StandUp, DriveCar, GetOutCar, AnswerPhone]



Conclusion

- Low accuracy in action recognition task.
- All the eye fixated regions in videos are not so much relevant for action recognition.
- Unlike **Yarbus' experiment**, action recognition is high level task which is largely intuitive and just does not rely on fixated visual input for common actions.
- **"Covert attention"** and immense **parallelism** in human brain.

Achievements

We proposed novel **"Array of Words"** approach with following advantages –

1. Capture **high-level temporality** in a given video sequence
2. Reduce usage of multivariate HMM to univariate basic HMM.
3. It reduces the continuous real domain of observations to discrete value from 1to #(Words)

Further Work

Can we extend this approach to design Human Visual Saliency Predictor ?

- Yes ! By training binary classifier over feature descriptor.
  Input: HoG3D feature detector around each pixel of the video data.
  Output: Yes or No (being salient)
- Problem – Might be computationally intensive.

REFERENCES

- 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.
- 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.
- Klaser, Alexander, and Marcin Marszalek. "A spatio-temporal descriptor based on 3D-gradients." (2008).
- Laptev, Ivan. "On space-time interest points." International Journal of Computer Vision 64.2 (2005): 107-123. – Hollywood-2 Dataset