Impact of Eye Fixated Regions in Visual Action Recognition

Mentor : Dr. Amitabha Mukerjee

Deepak Pathak
deepakp@

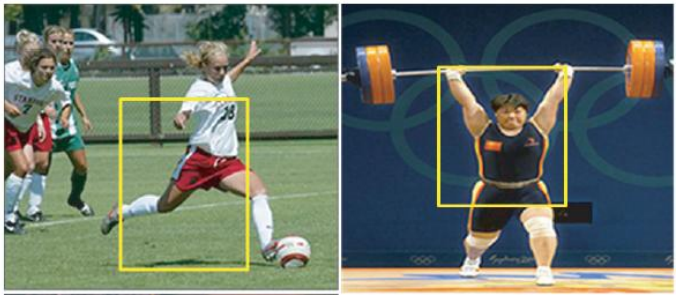
Dept. of Computer Science & Eng. ,
IIT Kanpur

Introduction

- Human Action Recognition (**What** ?)
- Human actions are major event in movies , news etc.
- Why is it useful ?
  - Annotating the videos
  - Content Based Browsing
  - Video Surveillance
  - Patient Monitoring
  - Analyzing sports videos
  - ..

YouTube Broadcast Yourself

150,000 uploads every day !



[UCF Sports Action Dataset]

[Live Snapshot – earthcam.com]

Motivation

- Issues in human action recognition –
  - Diversity - in actions (sitting , running , jumping etc)
    - interaction (hugging , shaking hands, fighting , killing etc)
  - Occlusions , noise, reflection, shadow etc
- Computer vision techniques still lag significantly behind human performance on similar tasks.
- Aim :
  - Study human gaze patterns in videos and utilize them
    - In activity recognition task.
    - Human visual saliency prediction

Human and Computer Vision

[Poggio 2007]

- Feature descriptor inspired from human visual cortex. Suggested hierarchical model with simplex features which are in coherence with the ventral stream of visual cortex.

[Mathe 2012]

- Provided large human eye tracking dataset recorded in context of dynamic visual action recognition tasks.
- Proposed saliency detector and visual action recognition pipeline

Experiment

- Recorded Human fixation for Hollywood-2 and UCF Sports Action Dataset
- 16 Subjects (Both M/F)
  : Free viewing – 4 subjects
  : Action Recognition – 12 subjects
- 92 subject-video hours, 500Hz sample rate.
- Dataset – coordinates of fixation and saccadic movement of eyes.

Experimental Setup [Mathe 2012]

Hollywood-2 Dataset

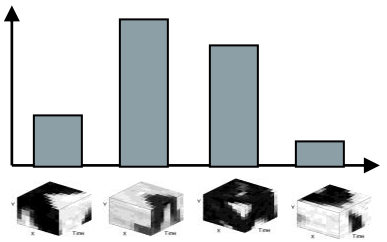
* Realistic human actions in unconstrained video clips of hollywood movies.
* 12 Action Classes
* 823 Training Video clips
  884 Test Video clips

Our Approach

Eye "fixation" points as Interest Points

Get HoG3D descriptor centered at these interest points

K-means clustering to map it to Visual Vocabulary



Histogram of Visual words

Train Classifier over the feature histograms

Done !!

Target

Through this, we would like to explore:

How useful is the foveated area formed by eye-fixated regions of entire video in the task of action classification ?

- Will be determined by comparing the result of our approach with other state of the art performances.

Implementation Details (Our approach)

- Interest points – Eye gaze ('F'-fixation) coordinates of one subject with 12 frame overlap. (computational reasons)
- Hog3D [Klaser 2008] descriptor for (823+884) videos.
- K-means clustering :
  - mapping of 6 lac descriptors to 4000 word vocabulary (dimension=300)
  - each video : normalized histogram of 4000 bins
- Learn SVM (Support Vector Machines) over 823 training videos feature histogram.
  Test over 884 test videos.

Intermediate Results

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

Action – FightPerson

Video Sample

Embedded

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).

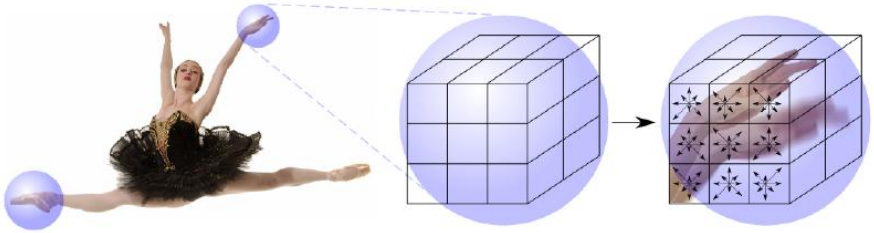
- Laptev, Ivan. "On space-time interest points." International Journal of Computer Vision 64.2 (2005): 107-123. – Hollywood-2 Dataset

Thank You

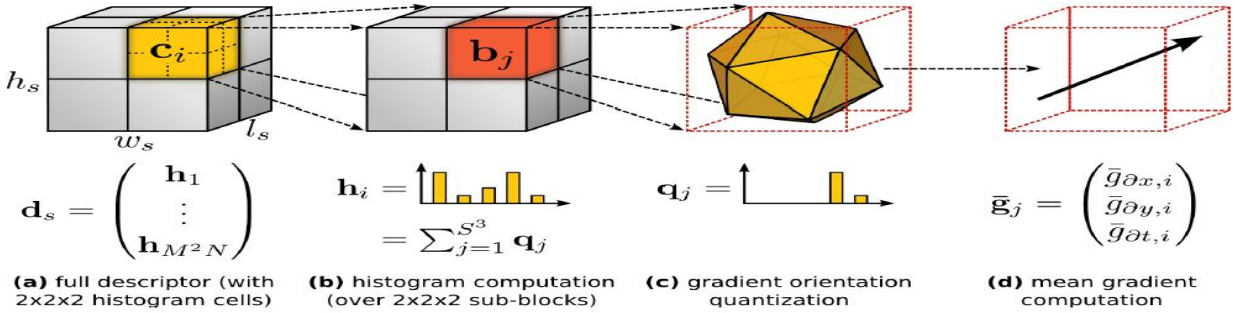
Questions ?

HoG3D – Feature 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]