Activity Recognition using Text Mining and Object Recognition

Niranjan A. Viladkar
Under the guidance from

Dr. Subhashis Banerjee and Dr. Parag Singla
Department of Computer Science
IIT Delhi

M.Tech Minor Project Presentation – May 13, 2013

Problem Statement

- What is Video Activity Recognition?

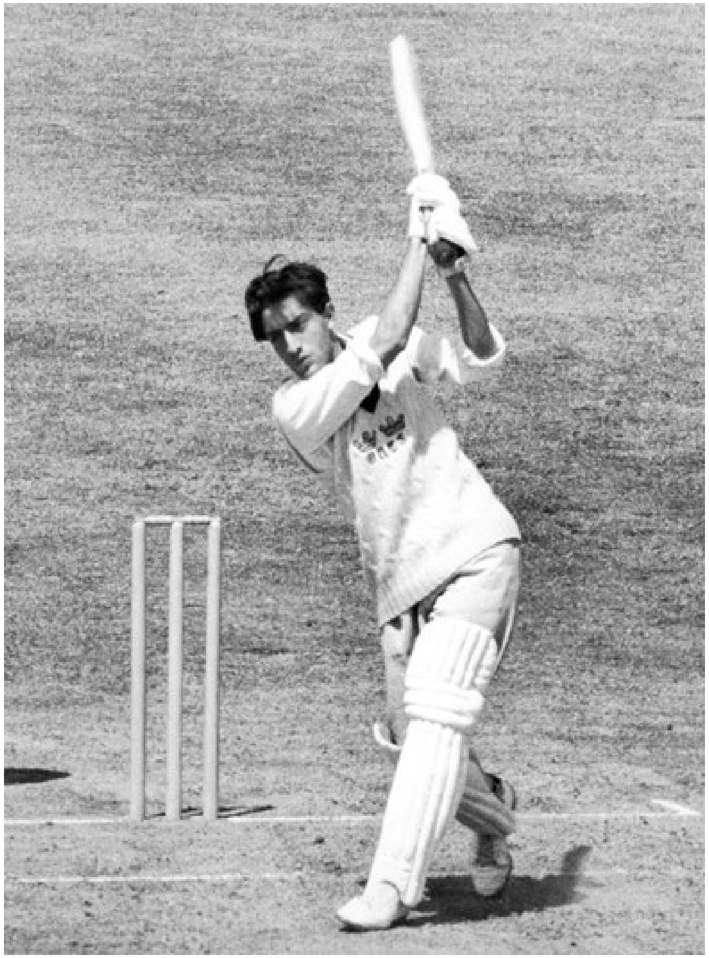
Problem Statement

- What is Video Activity Recognition?

Dance

Problem Statement

- What is Video Activity Recognition?



Play

Approach

Improving Video Activity Recognition using Object Recognition and Text Mining

by Tanvi Motwani and Raymond J. Mooney, ECAI-2012

Approach – Only Text Mining

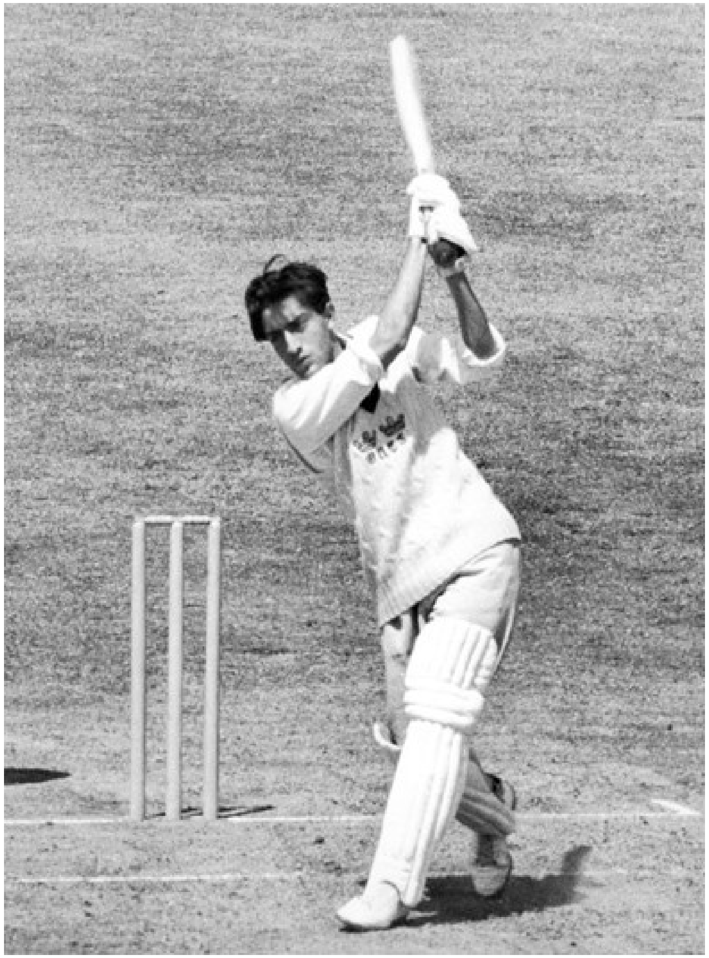
- Extract Labels - Use Natural Language descriptions of video clips.

- Extract STIP features – Represent a clip in HoG and HoF features.

- Train a model

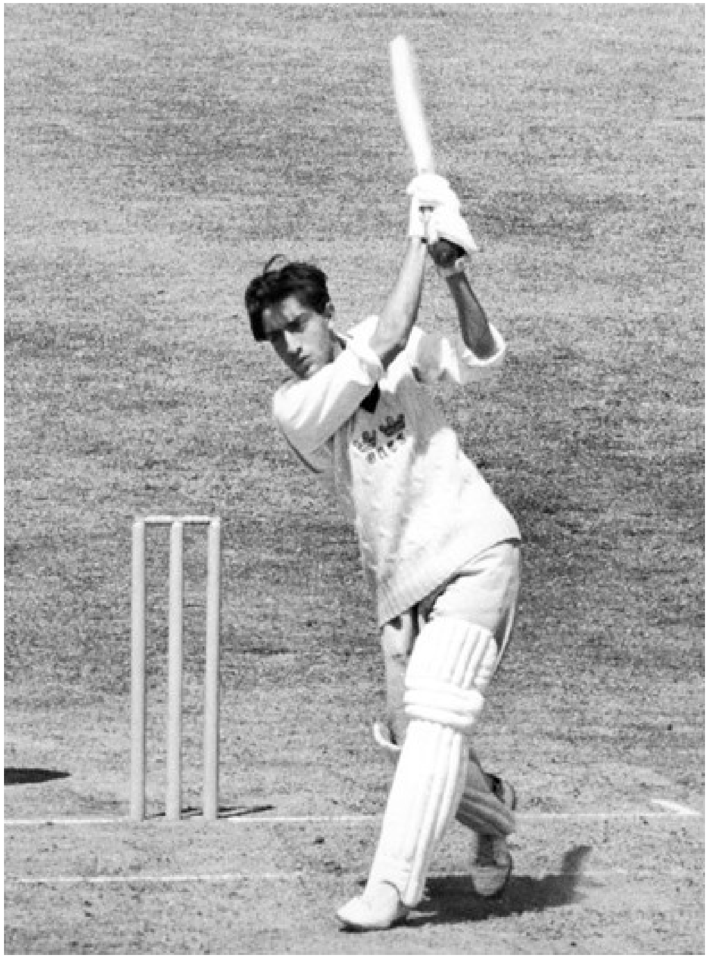
Approach – Only Text Mining

- Natural Language Description of a video



Approach – Only Text Mining

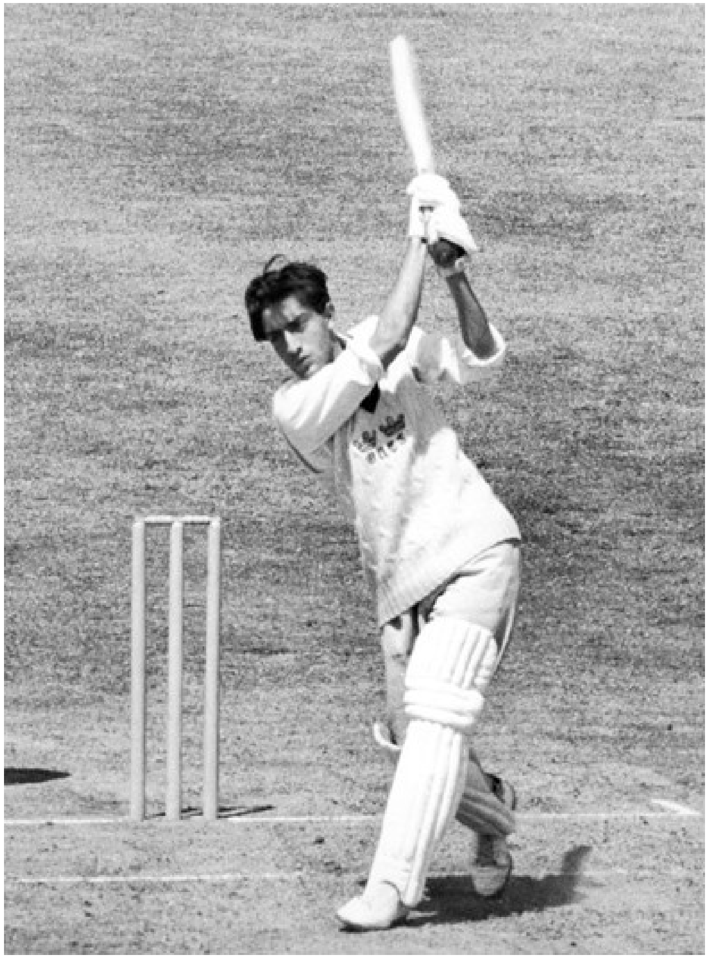
- Natural Language Description of a video



- Batsman is playing cricket

Approach – Only Text Mining

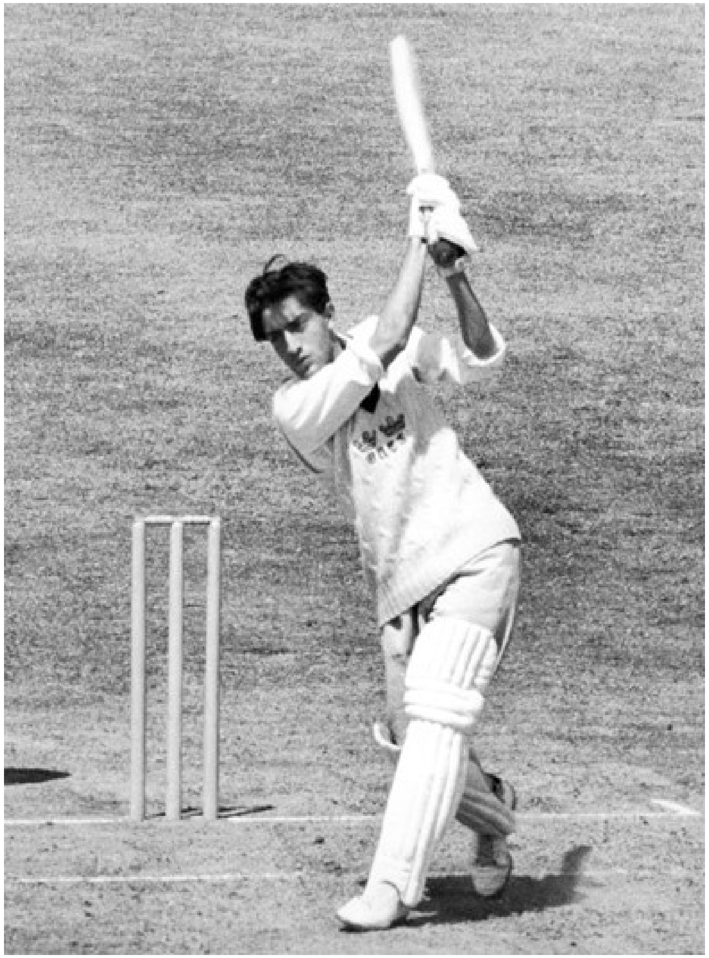
- Natural Language Description of a video



- Batsman is playing cricket
- Mansoor Ali Khan Pataudi is playing cricket.

Approach – Only Text Mining

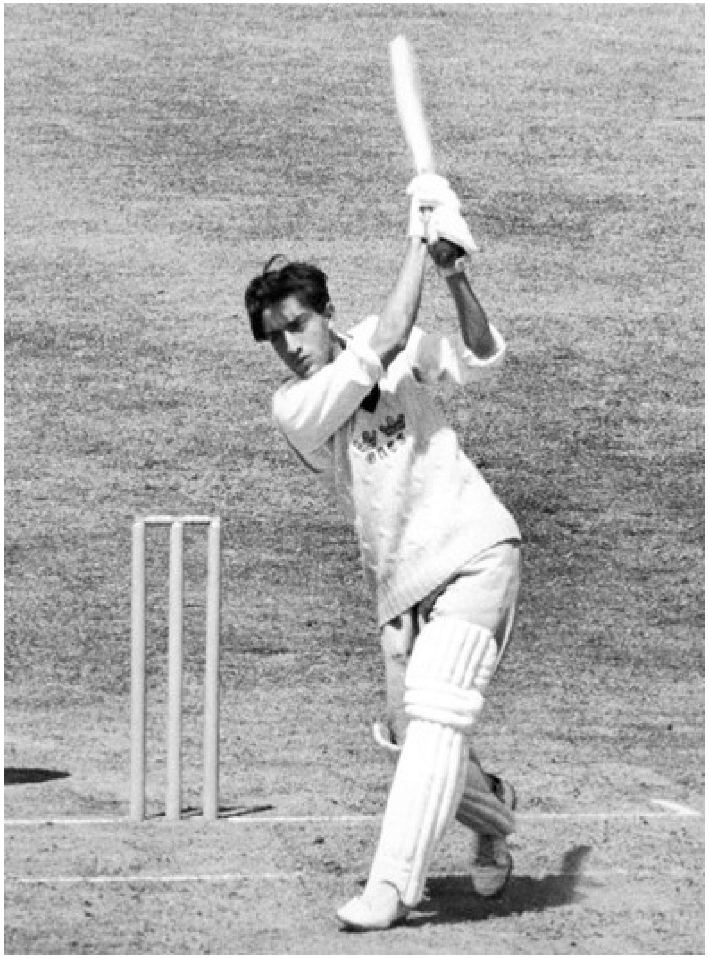
- Natural Language Description of a video



- Batsman is playing cricket
- Mansoor Ali Khan Pataudi is playing cricket.
- Players are playing cricket.

Approach – Only Text Mining

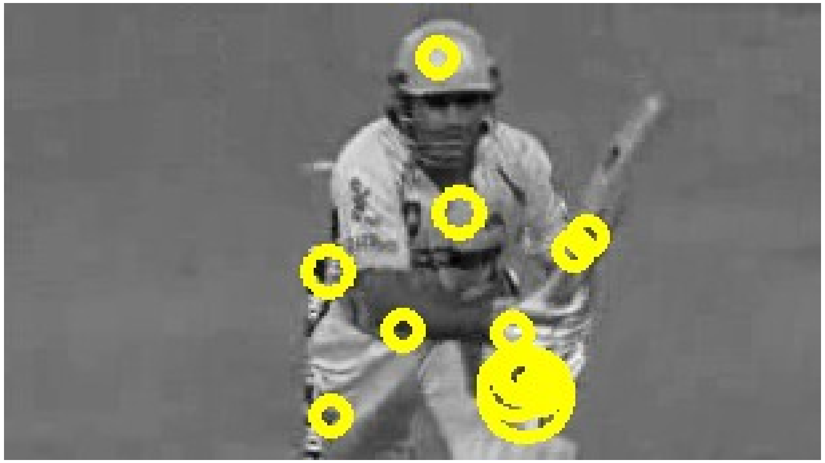
- Extracting Verbs from Description



- Batsman is playing cricket : PLAY
- Mansoor Ali Khan Pataudi is playing cricket. : PLAY
- Players are playing cricket. : PLAY

Approach – Only Text Mining

- Extracting STIP features



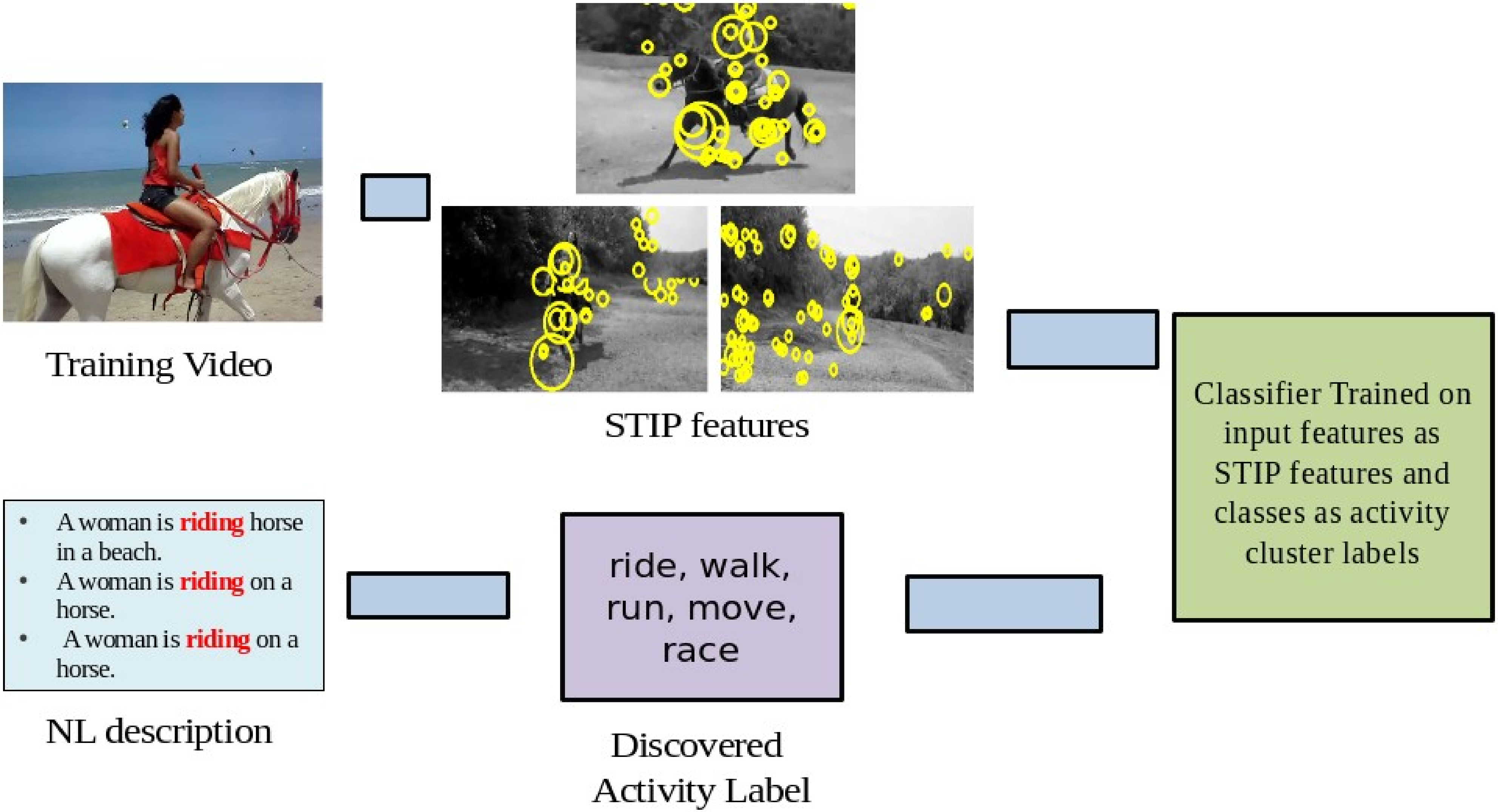
STIP HoG and HoF feature vector :

```
0.627496   0.0892087      0.0293946   0.253901   0.668772
0.160494   0.000758835  0.169975     0.414533   0.508073 …
……
……
```

Approach – Only Text Mining

- Take random samples of STIP feature descriptors
- Clustering K-Means
- Describe a video clip in terms of these clusters

Activity Recognizer using Video Features



Training Video

STIP features

Classifier Trained on input features as STIP features and classes as activity cluster labels

- A woman is **riding** horse in a beach.
- A woman is **riding** on a horse.
- A woman is **riding** on a horse.

NL description

ride, walk, run, move, race

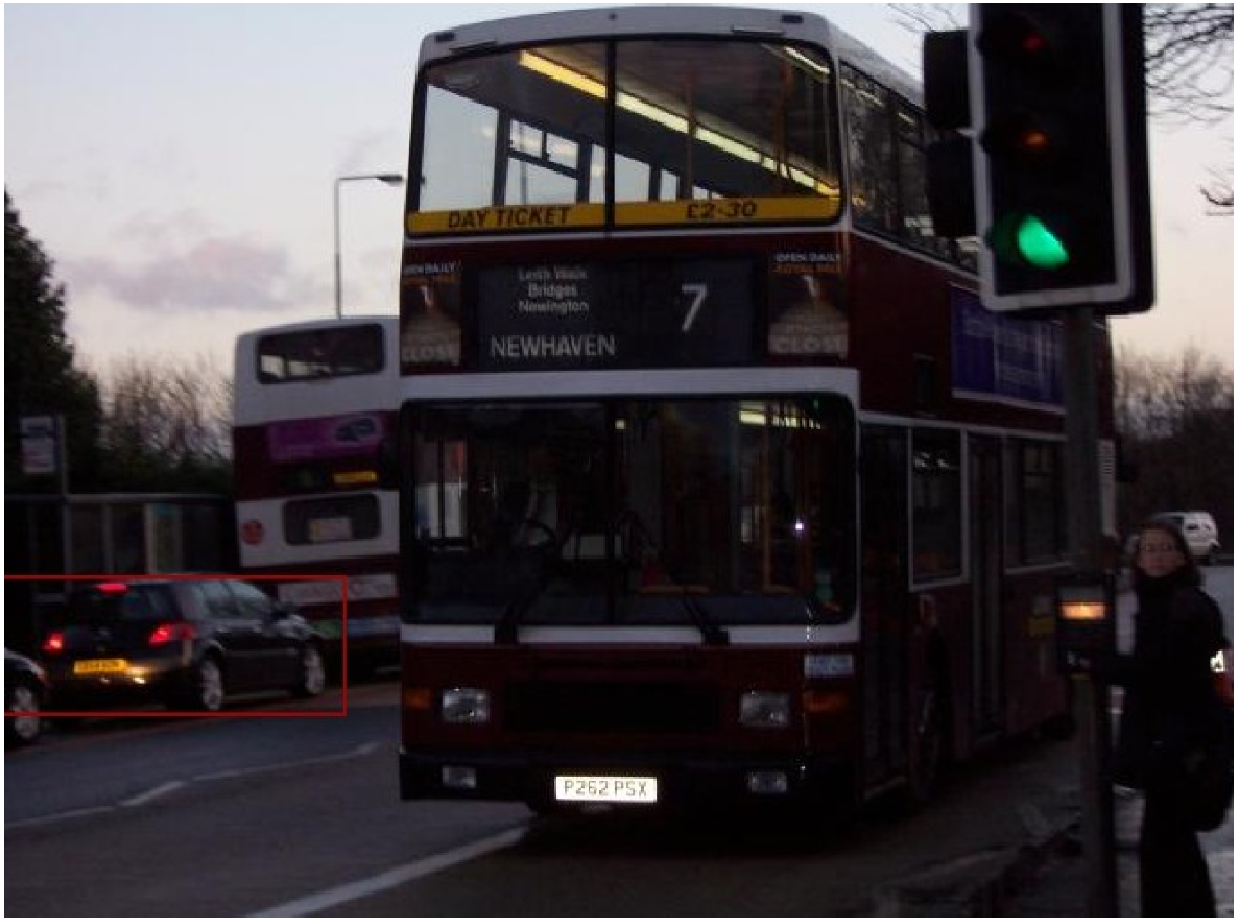
Discovered Activity Label

Object Detection

- Using Discriminatively Trained Deformable Part Models
  - Pre-trained object detector for 19 objects

Object Detection

Object Detection



Relation between Activity and Objects

- English Gigaword Corpus – 15 GB of raw text
- **Occurrence counts:**
  - of an activity A_i: occurrence of the verbs
  - of an object O_j: occurrence of object noun O_j or its synonym.
- **Co-occurrence of an Activity and an Object:**
  - ***POS Tagging***
    - Using Stanford tagger.
    - Occurrence of the object (tagged as noun) within a window of *w* or fewer words of an occurrence of the activity (tagged as verb).

Relation between Activity and Objects

Probability of each activity given each object

$$P(A_i|O_j) = (Count(A_i, O_j) + 1)/(Count(O_j) + |A|)$$

Integrated Activity Recogniser

- $P(A_i \mid F_v)$ – Calculated in 1st part.

Integrated Activity Recogniser

- $P(A_i \mid F_v)$ – Calculated in 1st part.
- $P(A_i \mid F_0)$ -

$$P(A_i|F_o) = \sum_{j=1}^{|O|} \underbrace{P(A_i|O_j)}_{\text{Gigaword Corpus}} * \underbrace{P(O_j|F_o)}_{\text{Object Detector}}$$

Integrated Activity Recogniser

- $P(A_i \mid F_V)$ – Calculated in 1st part.
- $P(A_i \mid F_0)$ -

$$P(A_i|F_o) = \sum_{j=1}^{|O|} P(A_i|O_j) * P(O_j|F_o)$$

- Consider only $P (A_i \mid F_V)$ when no object is detected and $P (A_i \mid F_0 , F_V)$ when objects are recognized

Work Done

- Verbs Extraction from Natural language description of clips done.

Clip Name	Natural Language Description
_0nX-El-ySo_83_93	A man is cutting a piece of paper.
_0nX-El-ySo_83_93	A man is cutting a paper by scissor.
_0nX-El-ySo_83_93	A man is cutting paper.
_0nX-El-ySo_83_93	A man is cutting a piece of paper.

Work Done

- Verbs Extraction from Natural language description of clips done.

Clip Name	Most frequent verb identified
_0nX-El-ySo_83_93	cut
_1vy2HIN60A_32_40	jump
_6OTzzK7t9Y_158_170	play
_6OTzzK7t9Y_73_78	crash

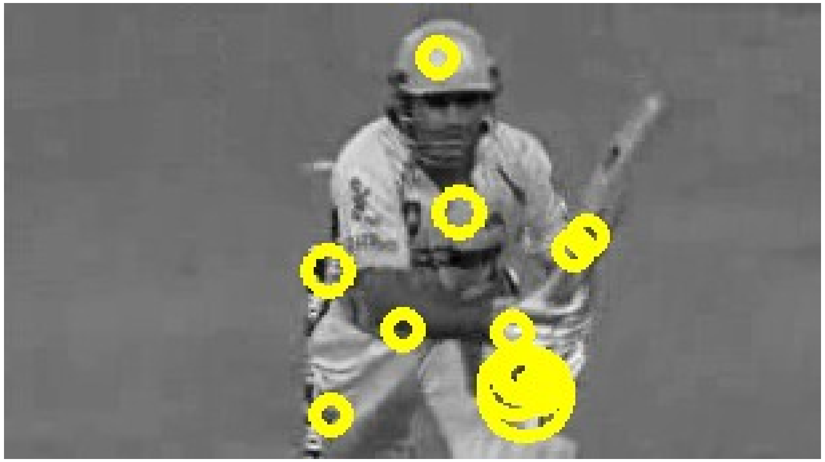
Work Done

- **Classes** Extraction from Natural language description of clips done.

Clip Name	Most frequent verb identified
_0nX-El-ySo_83_93	cut, slice
_O9kWD8nuRU_70_76	peel, remove
_JVxurtGIhI_32_42	sing, talk, bark
_WRC7HXBJpU_414_425	pour, stir, put

Work Done

- Classes Extraction from Natural language description of clips done.
- STIP features extraction done.



Work Done

- Classes Extraction from Natural language description of clips done.
- STIP features extraction done.
- Clustering done.

Work Done

- Classes Extraction from Natural language description of clips done.
- STIP features extraction done.
- Clustering done.

Work To be Done

- Representation of each clip
- Learning a model

Work Done

- Classes Extraction from Natural language description of clips done.
- STIP features extraction done.
- Clustering done.

Work To be Done

- Representation of each clip
- Learning a model
- Object Detection
- Learning Gigaword Corpus

Novel Idea

- Approach by Motwani et al. is only in forward direction.
- We plan to introduce notion of feedback
  - To improve accuracy of weak object detector and activity recogniser

References

- *Improving Video Activity Recognition using Object Recognition and Text Mining* by Tanvi Motwani and Raymond J. Mooney, ECAI-2012

- WordNet – 3.0 from Princeton University

- MIT Java Wordnet Interface from MIT

- WordNet Similarity from Sussex university