# PARSING AND RECOGNITION OF ANIMALS IN VIDEOS

Mentor : Dr. Amitabh Mukerjee

Atul Kumar Sinha

Sanchit Gupta

April, 2013

# Abstract

Object detection in images is an active field of research. Much of work has been done in the past on detection in static images. We extend this work to object detection in videos. This work explains and analyses the implementation of Bag of Words Model using SIFT features to detect the animals present in the video and classify them into their species. We also use the technique of Bootstrapping for robust classification. Saliency and motion maps have also been used to extract out the effective regions of interest. Testing is done on some IIT Kanpur zoo videos as well as on static images from image-net.

# Contents


1. Introduction and Motivation
2. Challenges
3. Previous Work
4. Image Features – SIFT
5. Bag of Words
6. Sliding window
7. Boot-strapping
8. Saliency and Motion Maps
9. Implementation
10. Results and Performance
11. Tools and Dataset
12. References

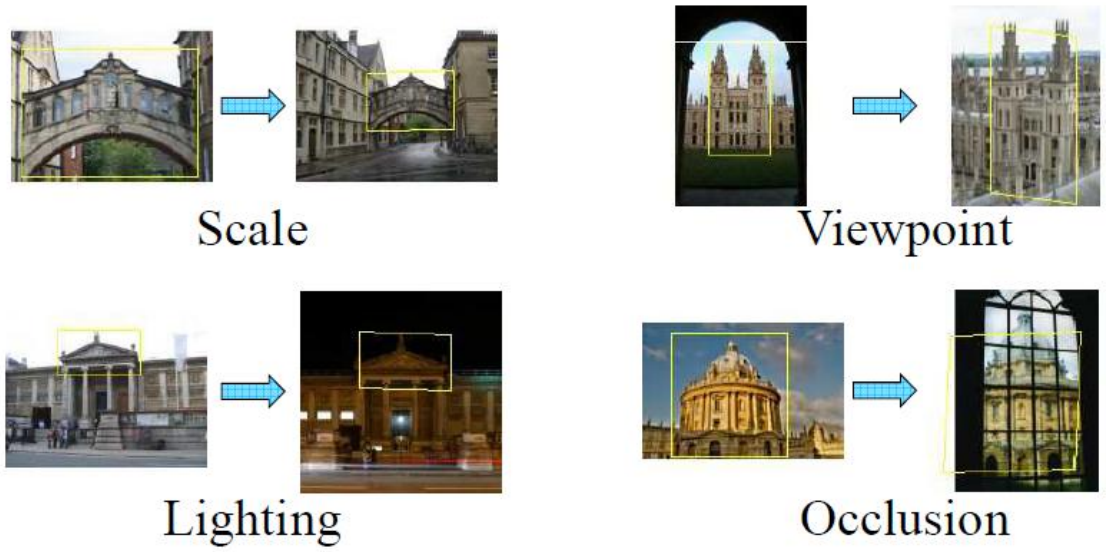
# 1. Introduction and Motivation

One of the important problems is in Computer Vision is Object Detection. Works on the same started quite early. It is important that we consider the problem of Object Detection for a sequence of images/videos because humans see things as a sequence of images and not as an isolated image. Most of the real world applications like robotic vision, medical imagery surveillance etc. would have a stream of images as an input as opposed to static image. Apart from this biologists are interested in studying the action and behavior patterns of animals in wildlife.

# 2. Challenges

Image Formation is a lossy process in the sense that a 3D scene is mapped to 2D image plane. Further, the difficulty is increased due to following aspects :

[from Summer School, Grenoble 2012]



# 3. Previous Work

**In [1],** Robust Principal Component analysis is used to separate out the foreground from the background followed by entropy analysis and application of Optical Flow algorithm for detection. The limitation of this work is that it works on videos with low frame rates.

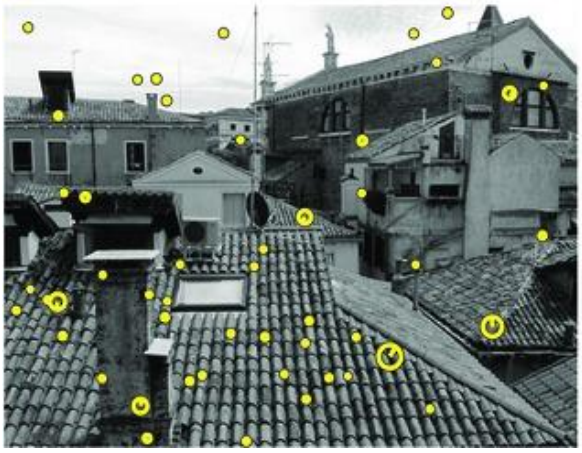
**In [2],** face detection techniques are used to detect animals in videos and thus it fails to work properly if face of animal is not visible.

# 4. Image Features – SIFT

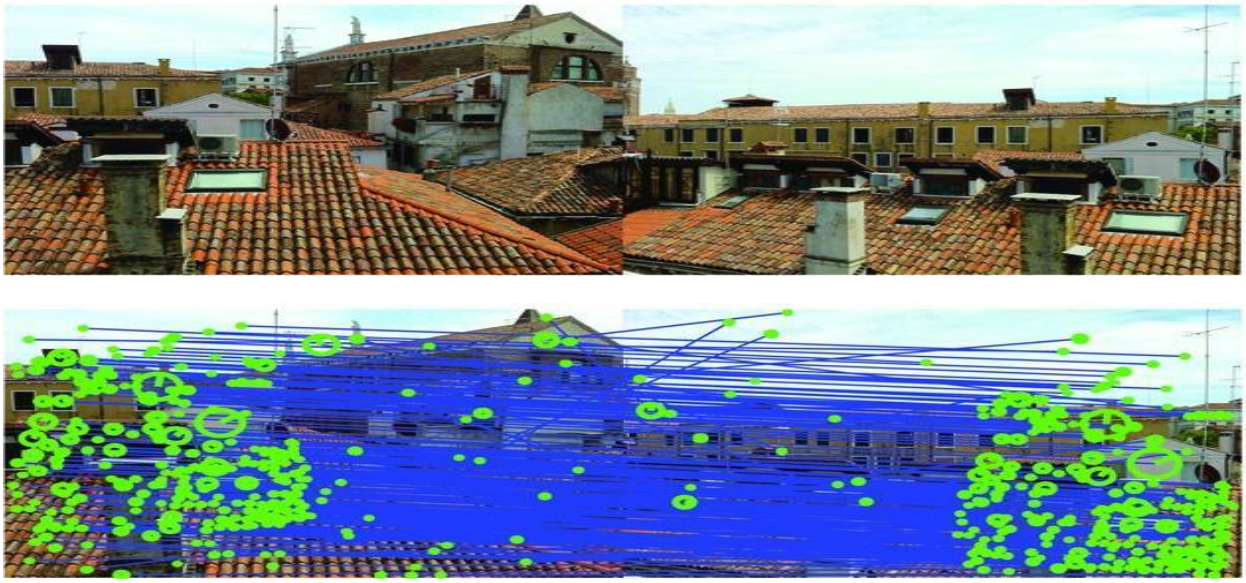
The Scale-Invariant Feature Transform (SIFT) bundles a feature detector and a feature descriptor. The detector extracts from an image a number of interesting points in a way which is consistent with variations of the illumination, viewpoint and other viewing conditions. The descriptor maps the regions with a signature through which their appearance can be identified robustly and compactly.

[all images from http://www.vlfeat.org/overview/sift.html]

**Input Image**



**Some SIFT features**

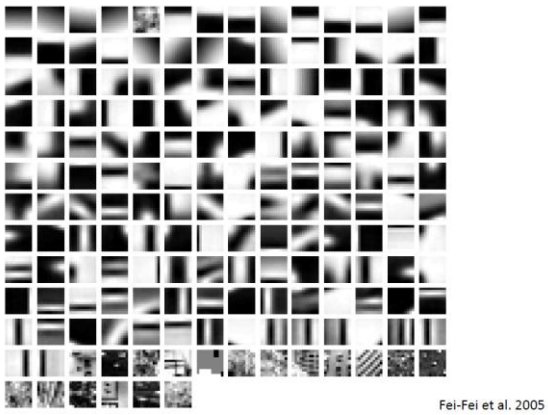


**Top: A pair of images of the same scene.**
**Bottom: Matching of SIFT descriptors with vlfeat library**

# 5. Bag of Words

In bag of words method every image is represented as a histogram of frequencies of visual words. The mapping from image features to visual words is done using Nearest Neighbour approach. The information about the location of visual words i.e. spatial sense is lost but still this method turns out to be quite effective.

[from Summer School, Grenoble 2012]




**Visual Words Vocabulary**

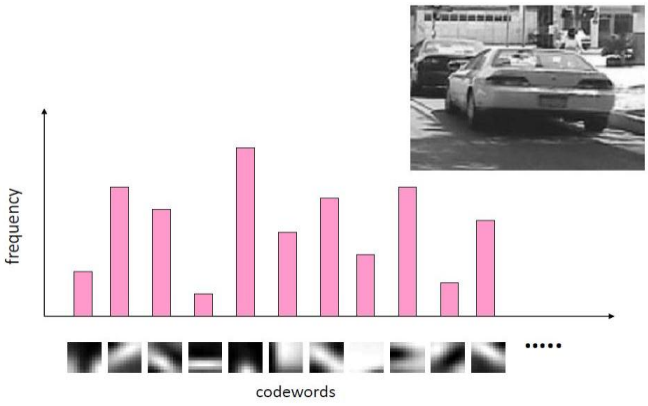
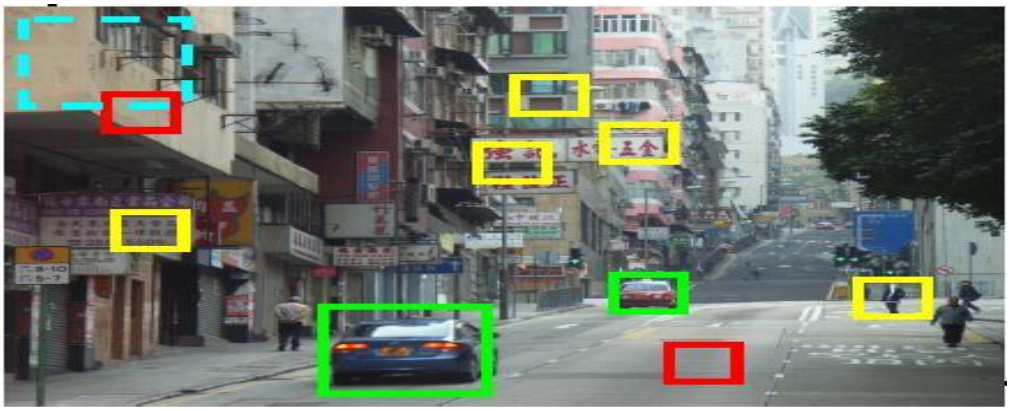



**Image as a histogram of visual words**

# 6. Sliding window

Sliding window is the state of the art in object detection area in computer vision. In this method a classifier is applied at all positions, scales and in some cases, orientations of an image. However, testing all points in the search space can turn out to be slow. For n x n image , we need to test $O(n^4)$ windows.

[from Summer School, Grenoble 2012]



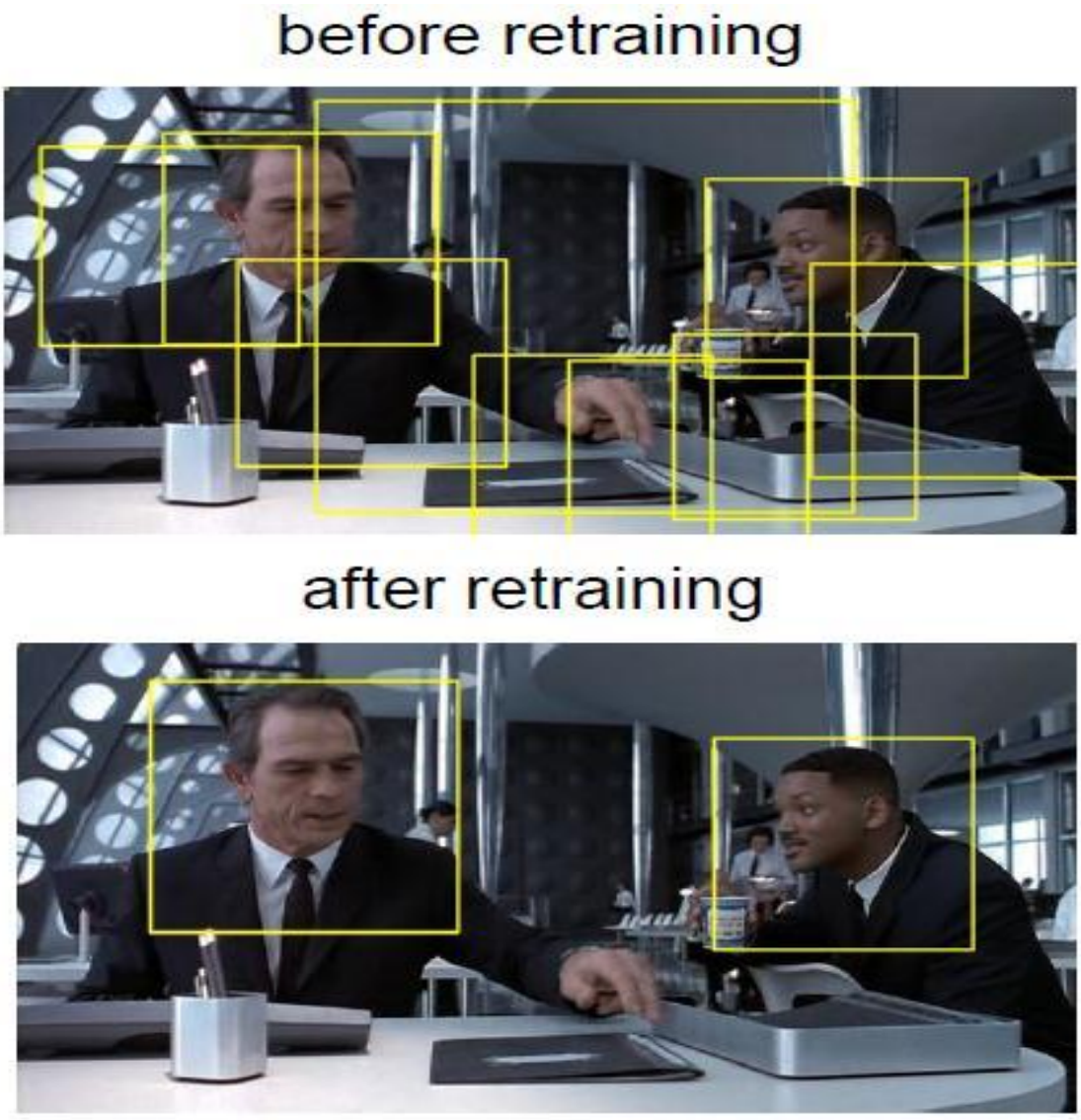
**Bounding boxes over the image using Sliding Window**

# 7. Boot-strapping

Boot-strapping is used to improve the robustness of the classifier, so as to eliminate the false positives. In bootstrapping we test the classifier on the training data itself. The hard negative data i.e. the data which gets high negative confidence score is extracted out and the classifier is retrained iteratively with these hard negative instances as negative images.

**Algorithm :**

- Negative data picked at random and classifier is trained
- Run on training data and extract images with high negative confidence points
- Add these false positives to training set and train a new classifier**.**
- Repeat for 2 to 3 times.




[from Summer School, Grenoble 2012]

# 8. Saliency and Motion Maps

Both saliency and motion maps are bottom-up cues. Bottom up cues are independent of the target class that is being searched for in the video. In this work, we exploit these techniques for pruning our search space.

**Saliency Maps:** The saliency maps are used to localize the salient points in the image based on contrast, orientation etc. The features that stand out from the surroundings can be extracted using saliency maps. Higher the saliency more attention the region grabs.

**Motion Maps:** The motion maps are used to localize the points in the image which have motion associated with them. We need to focus attention on regions where motion is detected since the probability of animal being present in that area is high.

# 9. Implementation

Our approach can be summarized in the following points in sequence:

## Training

- Take large set of images( +ve and –ve)
- Compute SIFT features
- Map features to visual words
- Compute histograms of visual words frequency
- Train all Vs. one SVM's on these histograms
- Bootstrapping

## Testing

- Take the video as input
- Extract the frames from the video
- Apply saliency and motion maps to determine effective candidate regions
- Sliding window iteration over the candidate regions
- Run SVM classifier over each iteration
- Retain bounding box for regions with confidence score above the threshold value

# 10. Results and Performance

We trained the model on about 700 images from 5 classes of animals : lion, peacock,  zebra, hyena and bison.

We tested the model on

- An input video of peacock
- About 120 images of 5 classes of animals : lion, peacock,  zebra, hyena, and bison.

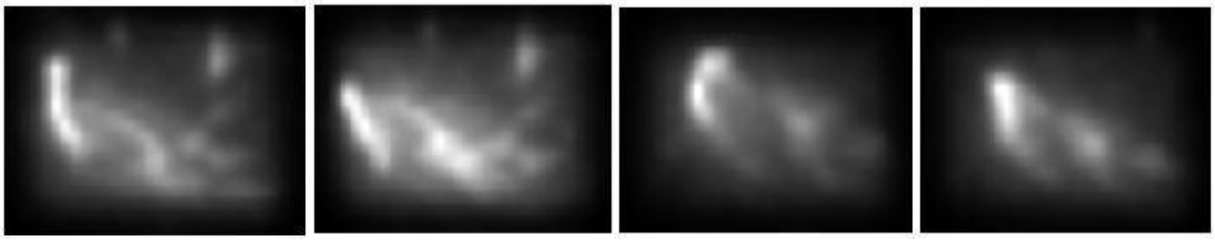
**The results for test on images for the various classes of animals are as follows:**

| Sr No. | Class of animal tested | No. of Test Images | No. of images correctly classified | Performance |
|---|---|---|---|---|
| 1 | Peacock | 120 | 45 | **37.5%** |
| 2 | Lion | 120 | 41 | **31.6%** |
| 3 | Zebra | 120 | 34 | **28.3%** |
| 4 | Hyena | 100 | 28 | **28%** |
| 5 | Bison | 100 | 23 | **23%** |

**The results for test on peacock video are as follows:**

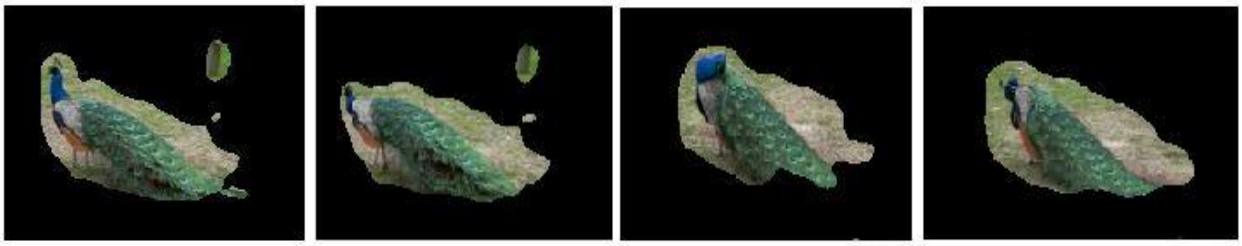
**Input Video Frames:**

**Saliency Maps :**

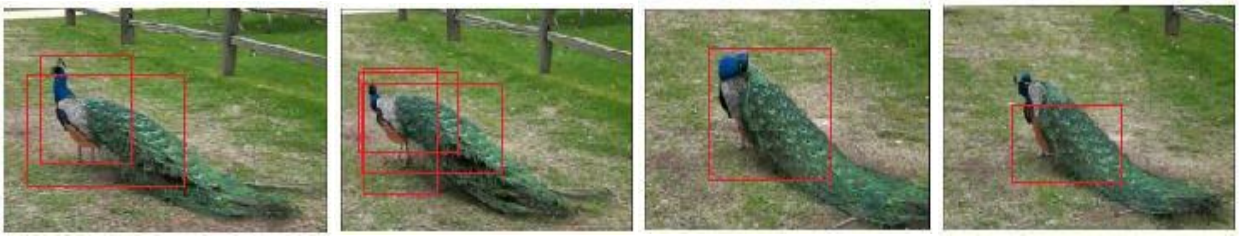


**Motion Maps :**

**Region of Interest :**



**Output :**



# 11. Tools and Dataset

**TOOLS :**

- VLFeat : computes SIFT features
- Saliency : Matlab implementation of Itti-Koch saliency algorithm
- LibSVM : Multi-class SVM

**DATASET :**

- ImageNet : +ve and –ve Image samples
- Oxford's Visual vocabulary
- Self-made IIT Kanpur animal video

# References


- P. Khorrami, J. Wang, and T. Huang : Multiple Animal Species Detection Using Robust Principal Component Analysis and Large Displacement Optical Flow. [2012]
- T. Burghardt and J. Calic. Analysing animal behavior in wildlife videos using face detection and tracking. Vision, Image and Signal Processing, IEEE Proceedings, 153(3) 305-312 [2006]
- A. Vedaldi and B. Fulkerson. VLFeat : A Portable Library of Computer Vision Algorithms. [2008]
- Visual Recognition and Machine Learning. Summer School, Grenoble. [2012]
- www.image-net.org