Learning video saliency from human gaze using candidate selection

Rudoy, Goldman, Shechtman, Zelnik-Manor

CVPR 2013

Paper presentation by Ashish Bora

Outline

- What is saliency?
- Image vs video
- Candidates : Motivation
- Candidate extraction
- Gaze Dynamics : model and learning
- Evaluation
- Discussion

What is saliency?

- Captures where people look
- Distribution over all the pixels in the image or video frame
- Color, high contrast and human subjects are known factors

Images credit : http://www.businessinsider.com/eye-tracking-heatmaps-2014-7
http://www.celebrityendorsementads.com

Image vs video saliency

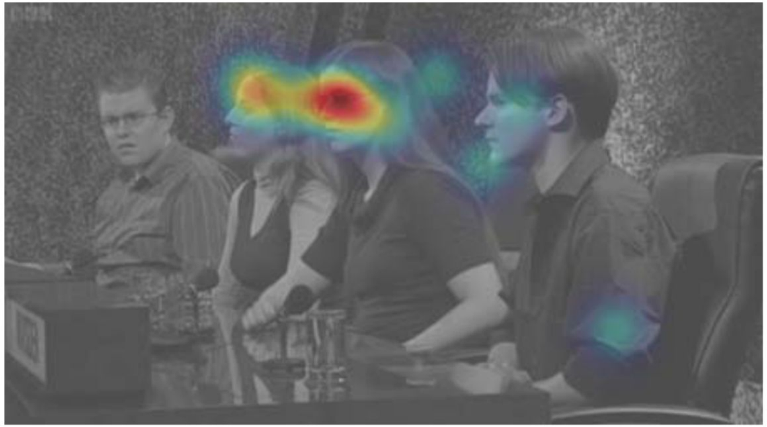


Image (3 sec)

Video

- Shorter time - typically single most salient point (sparsity)
- Continuity across frames
- Motion cues

Image credit : Rudoy et al

How to use this?

- Sparse saliency in video
  - Redundant to computer saliency at all pixels
  - Solution : inspect a few promising candidates
- Continuity in gaze
  - Use preceding frames to model gaze transitions

Candidate requirements

- **Salient**
- **Diffused :** Salient area rather than a point
  - Represented as a gaussian blob (mean, covariance matrix)
- **Versatile :** incorporate broad range of factors that cause saliency
  - Static : local contrast or uniqueness
  - Motion : inter-frame dependence
  - Semantic : arise from what is important for humans
- **Sparse :** few per frame

Candidate extraction pipeline : Static

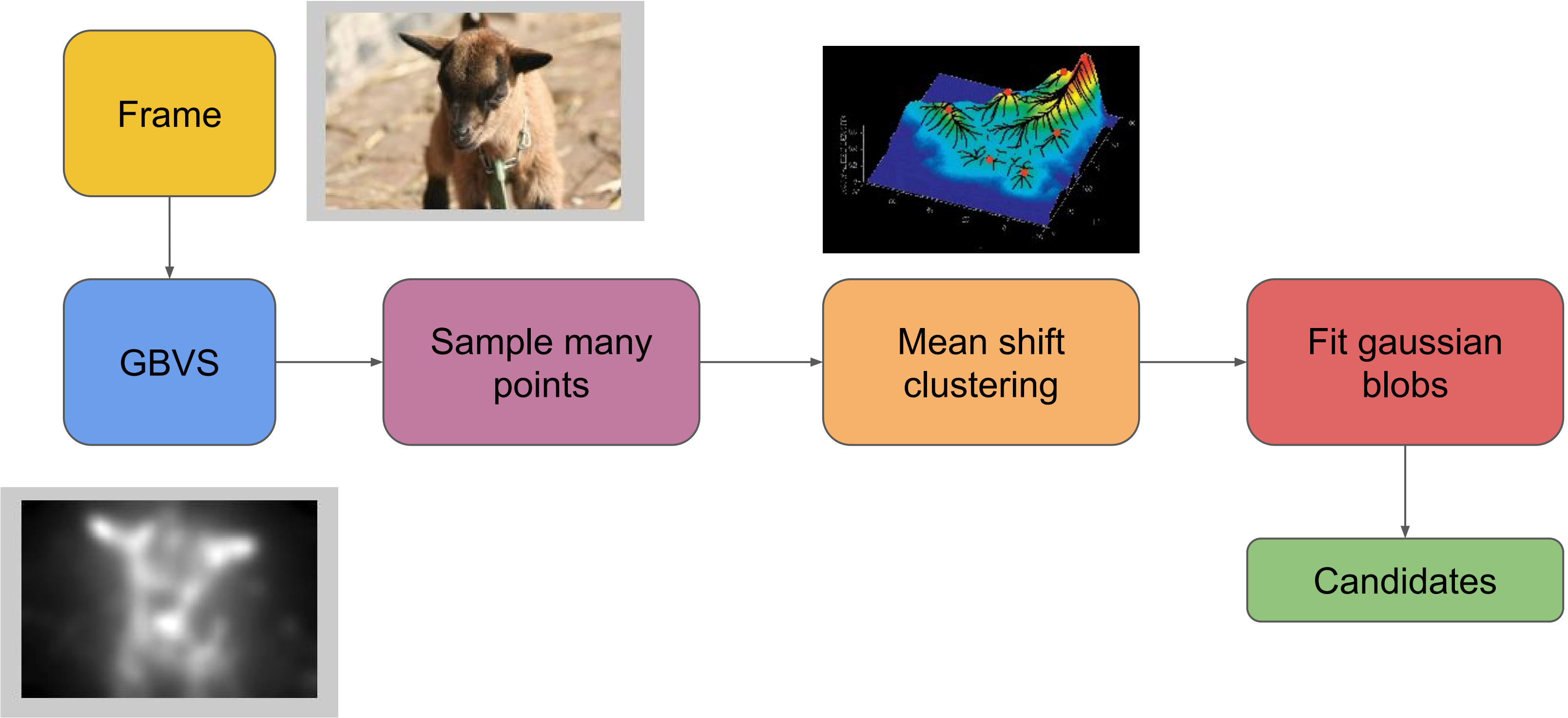


Image credit : http://www.fast-lab.org/resources/meanshift-blk-sm.png
http://www.vision.caltech.edu/~harel/share/gbvs.php

Static candidates : example

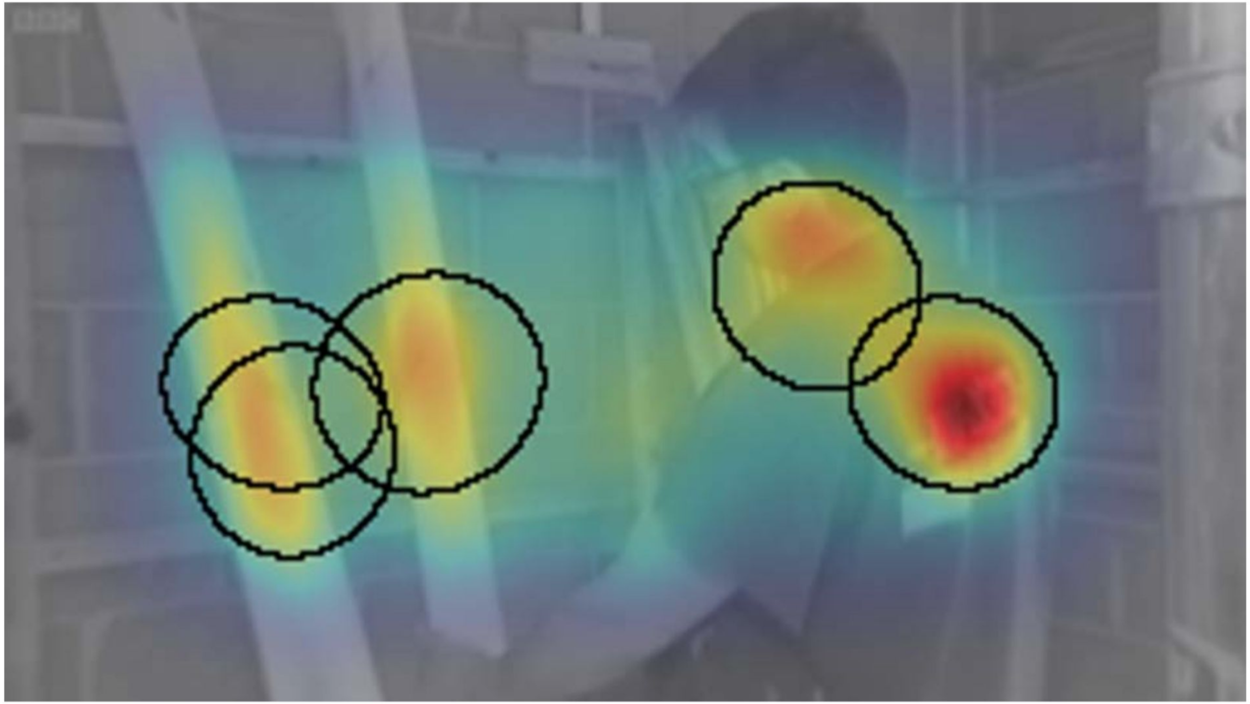


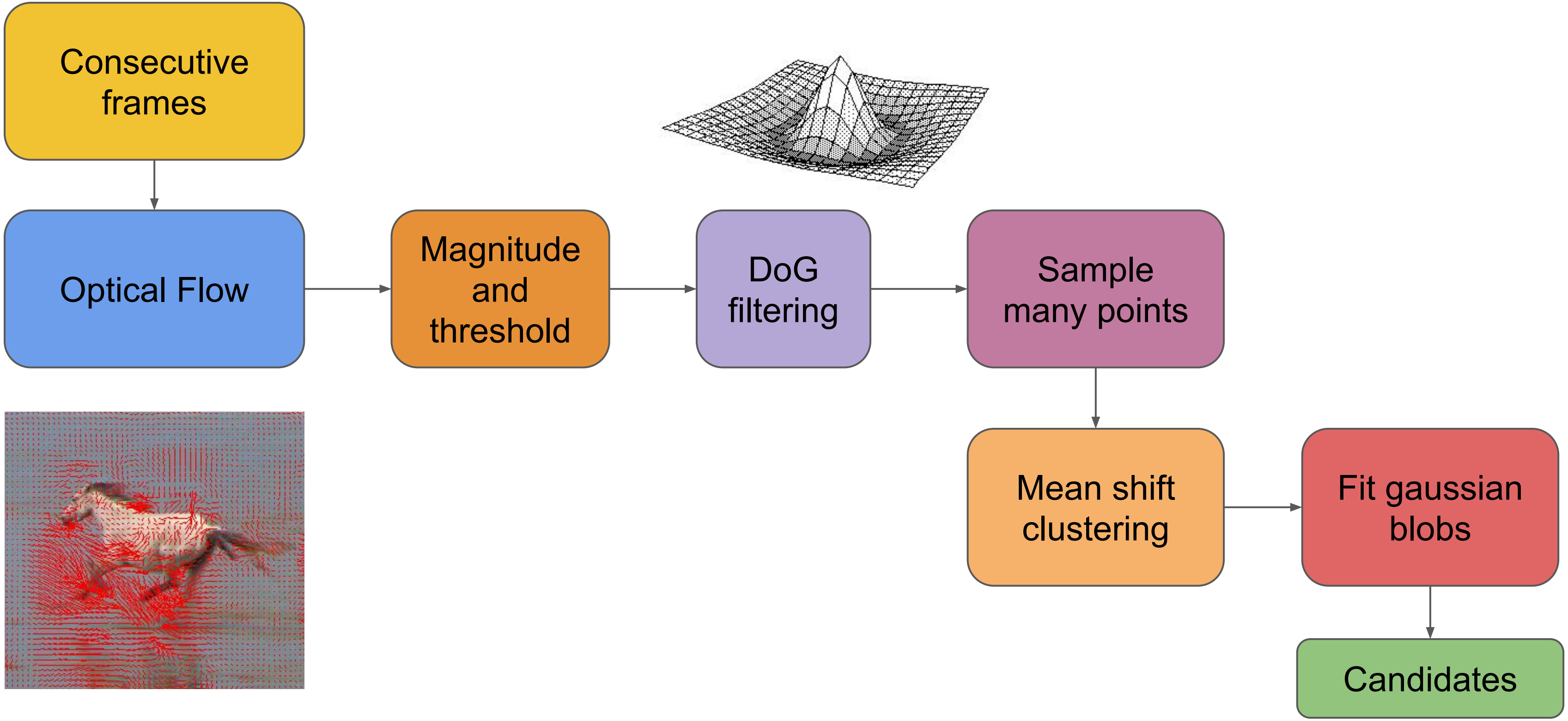
Image credit : Rudoy et al

Discussion

Why not fit a mixture of gaussians directly?

- Rationale in paper : Sampling followed by mean shift fitting gives more importance to capturing the peaks
- Is this because more points are sampled near the peaks and we weigh each point equally?

Candidate extraction pipeline : Motion

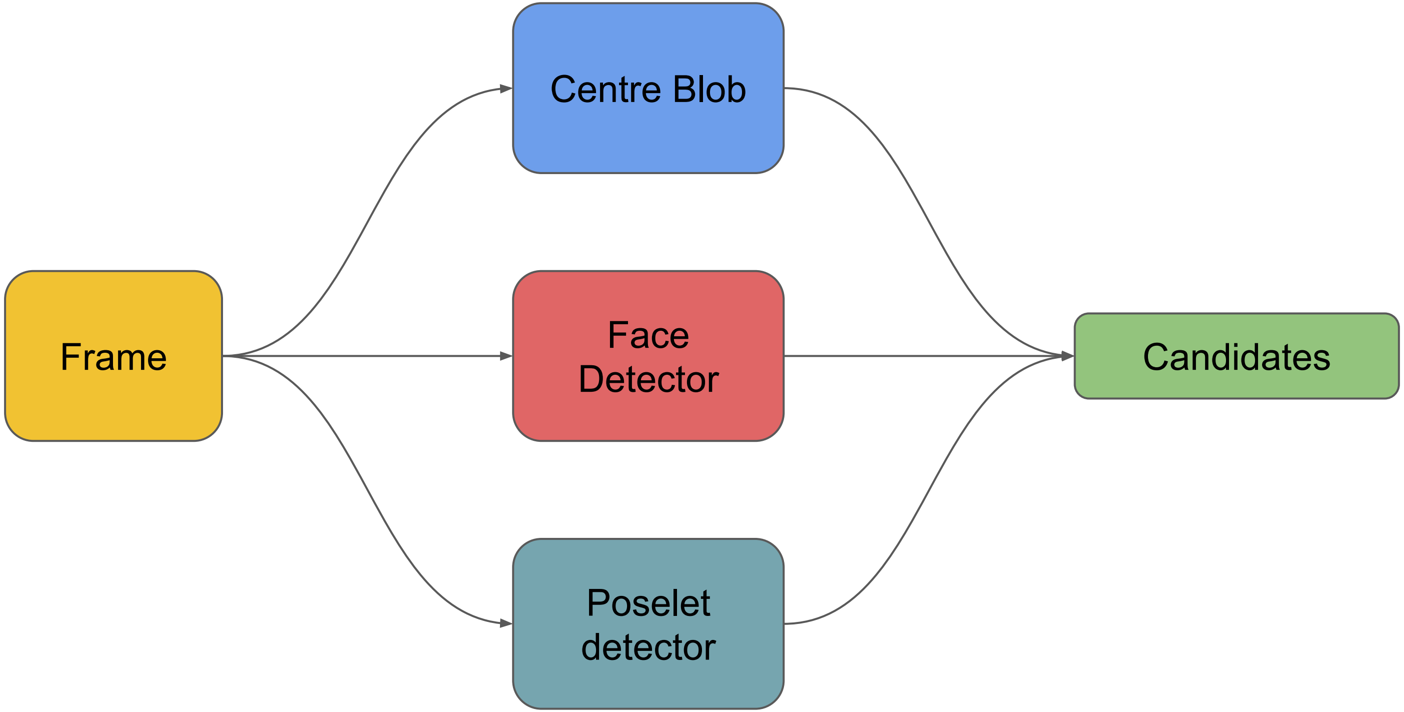


Images cropped from : http://cs.brown.edu/courses/csci1290/2011/results/final/psastras/images/sequence0/save_0.png
http://www.liden.cc/Visionary/Images/DIFFERENCE_OF_GAUSSIANS.GIF

Motion candidates : example

Image credit : Rudoy et al

Candidate extraction pipeline : Semantic



Semantic candidates : example

Image credit : Rudoy et al

Modeling gaze dynamics

- s_i = source location
- d = destination candidate
- Learn transition probability $P(d|s_i)$

Image credit : Rudoy et al

Modeling gaze dynamics

- $$P(s_i) = \frac{Sal(s_i)}{\sum_{i \in S} Sal(s_i)}$$

- Use $P(s_i)$ as a prior to get P(d)

$$P(d) = \sum_{i \in S} P(d|s_i) \cdot P(s_i)$$

- Combine destination gaussians with P(d)

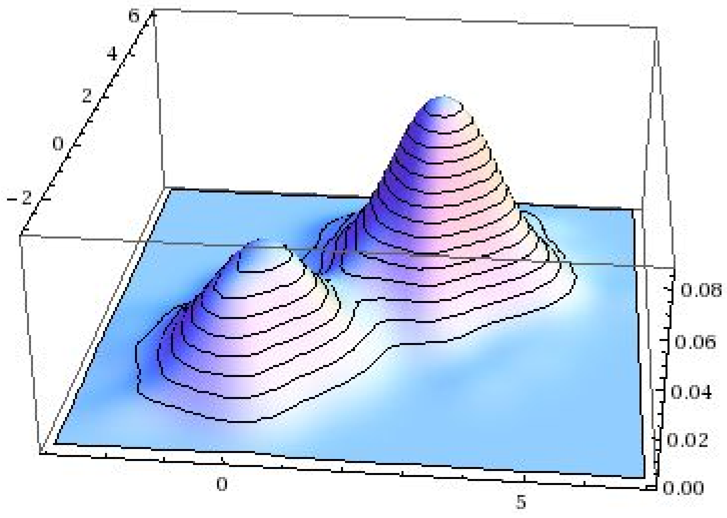


Image credit : http://i.stack.imgur.com/tYVJD.png
Equation credit : Rudoy et al

Learning $P(d|s_i)$: Features

Only destination and interframe features are used

- Local neighborhood contrast

$$C_l = \frac{I_n^{max} - I_n^{min}}{(I_n^{max} + I_n^{min}) \cdot C_g}$$

where

$$C_g = \frac{I^{max} - I^{min}}{I^{max} + I^{min}}$$

Equation credit : Rudoy et al

Learning $P(d|s_i)$: Features (contd)

Only destination and interframe features are used

- Mean GBVS of the candidate neighborhood
- Mean of Difference-of-Gaussians (DoG) of
  - Vertical component of the optical flow
  - Horizontal components of the optical flow
  - Magnitude of the optical flow

  in local neighborhood of the destination candidate
- Face and person detection scores
- Discrete labels : motion, saliency (?) , face, body, center, and the size (?)
- Euclidean distance from the location of d to the center of the frame

Discussion : unclear points

- It seems no feature depend on source location.
  In that case $P(d|s_i)$ will be independent of s_i .
  That would mean $P(d)$ is independent of $P(s_i)$
  This is like modeling each frame independently with optical flow features

- Discrete labels for saliency and size

Discussion

- Non-human semantic candidates?
  - not handled

- Extra features that can be useful
  - General : Color and depth, SIFT, HOG, CNN features
  - Task specific
    - non-human semantic candidates (for example text, animals)
    - activity based candidates
    - memorability of image regions

Learning $P(d|s_i)$: Dataset

- DIEM (Dynamic Images and Eye Movements) dataset [1]
- 84 videos with gaze tracks of about 50 participants per video

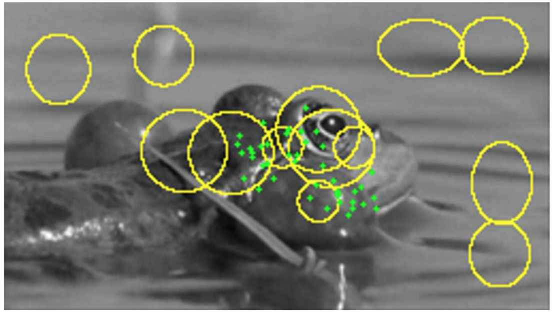
[1] https://thediemproject.wordpress.com/

Learning $P(d|s_i)$: Get relevant frames

- (Potentially) positive samples
  - Find all the scene cuts
  - Source frame is the frame just before the cut
  - Destination is 15 frames later

- Negative samples
  - Pairs of frames from the middle of every scene 15 frames apart

Learning $P(d|s_i)$: Get source locations



Ground truth human fixations

Smoothing

Thresholding (keep top 3%)

Find Centres (foci)

Source locations

Image credit : Rudoy et al

Learning $P(d|s_i)$

- Take all pairs of source locations and destination candidates for training set
- Positive labels:
  - Pairs with centre of d “near” a focus of the destination frame
- Negative labels:
  - If centre of d is “far” from every focus of destination frame
- Training
  - Random Forest classifier

Labeling : example

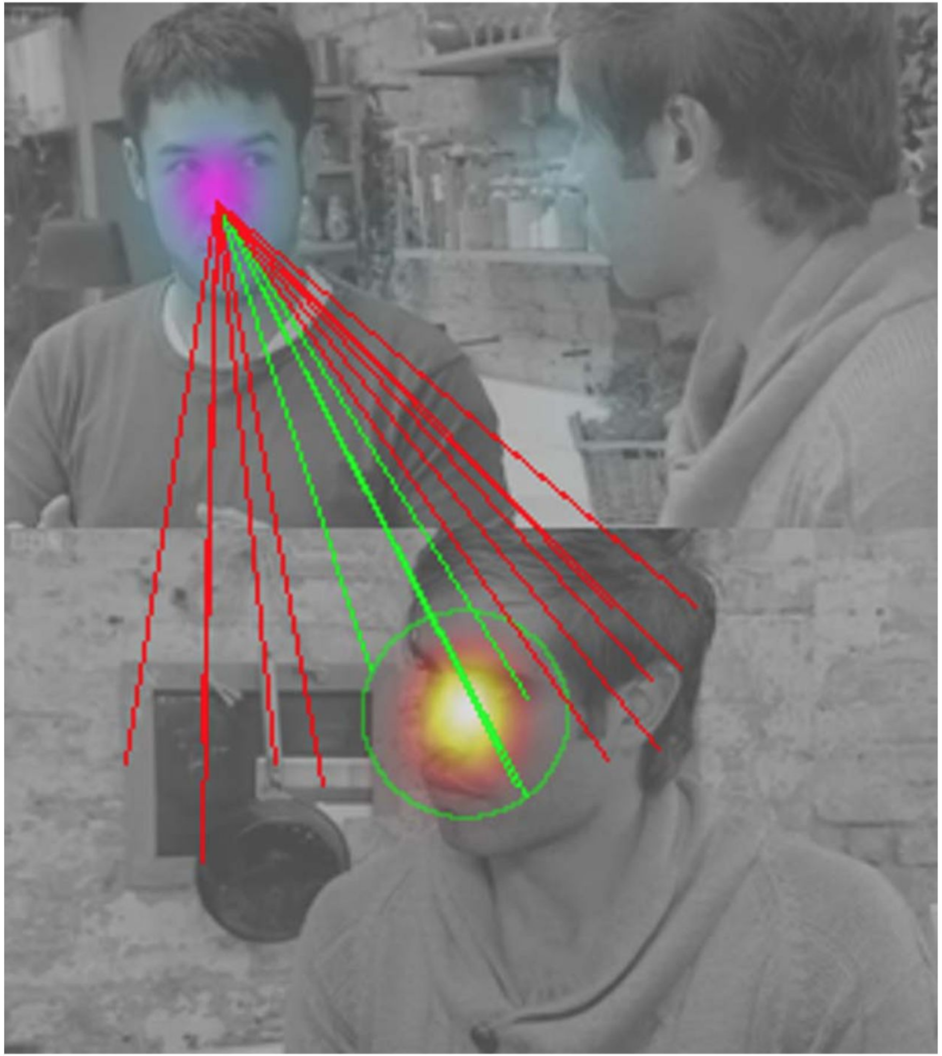


Image credit : Rudoy et al

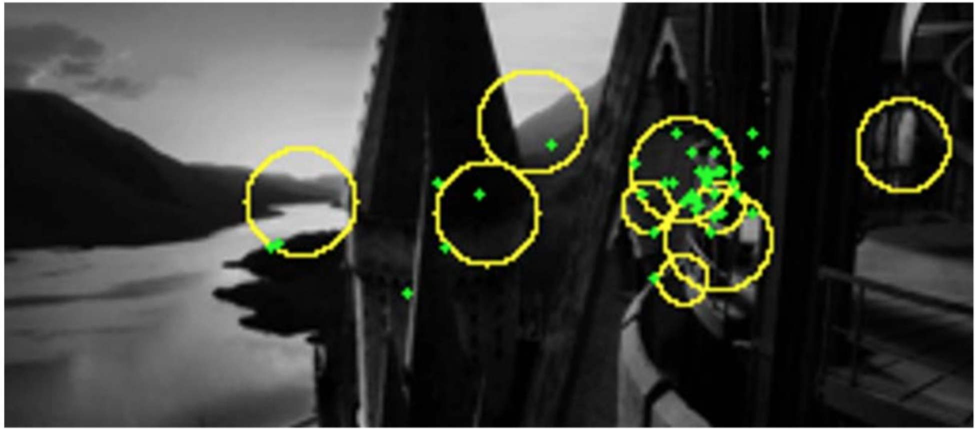
Discussion

- Why Random Forest?
  - No discussion in paper
  - Other classifiers/models that can be used
    - XGBoost
    - LSTM to model long term dependencies

Results : video

Experiments : How good are the candidates?

Candidates cover most human fixations



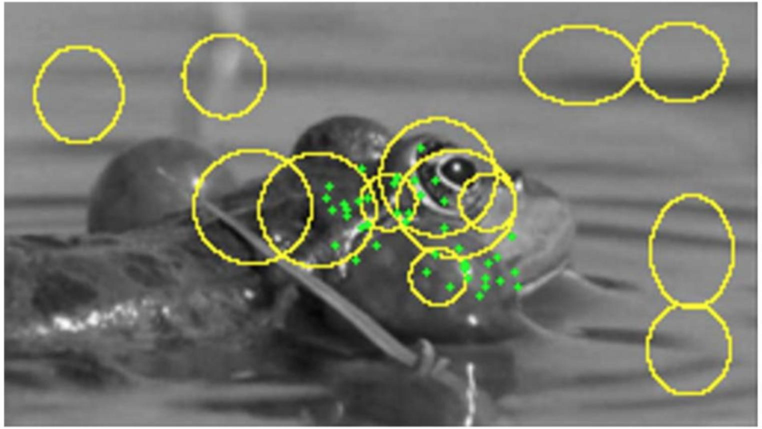


Image credit : Rudoy et al

Experiments : How good are the candidates?

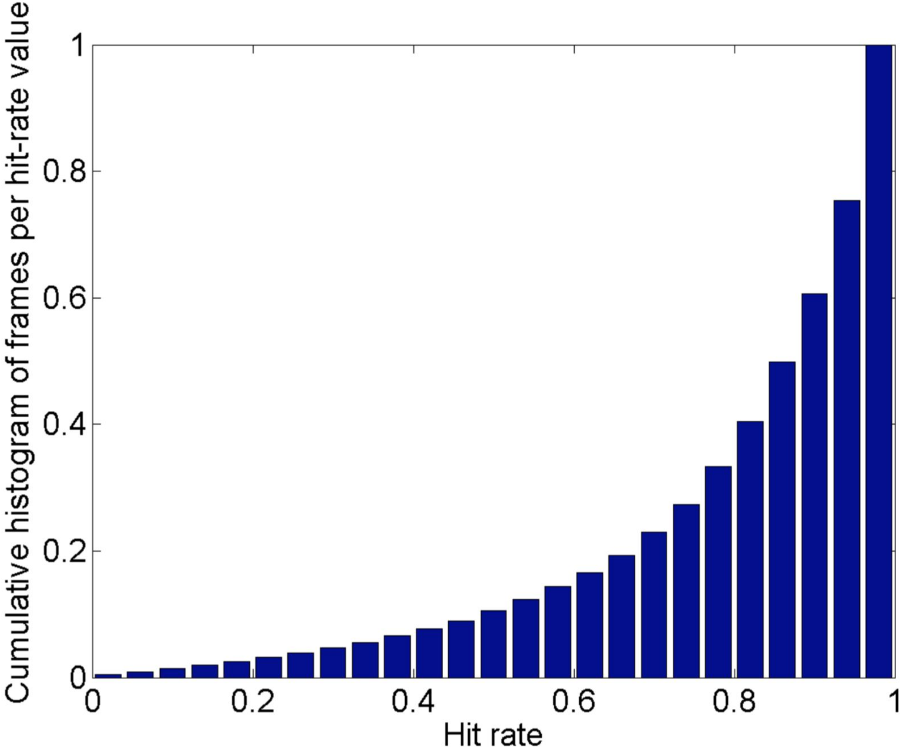
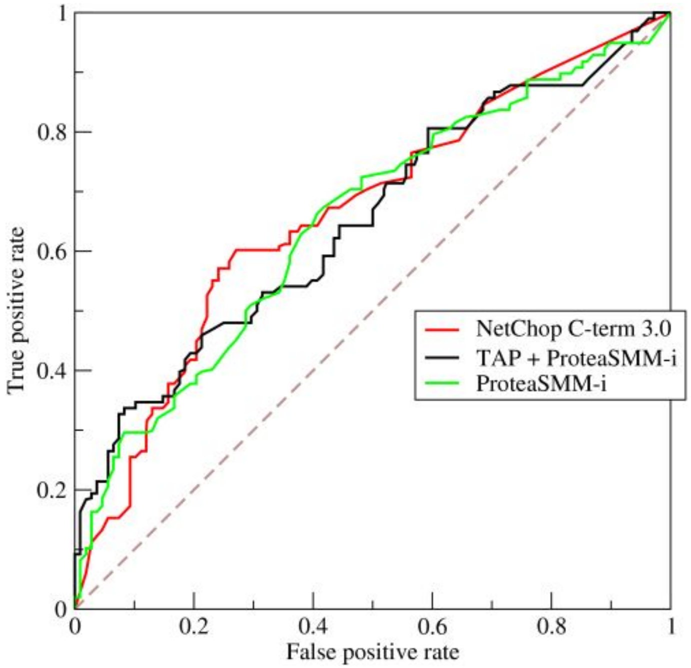


Image credit : Rudoy et al

Experiments : Saliency metrics

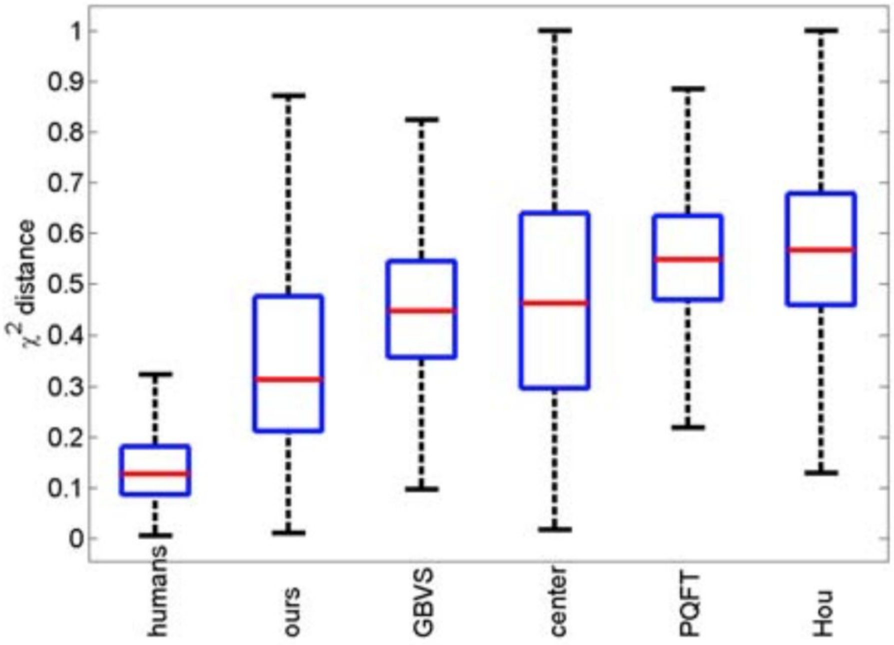
- AUC ROC to compute the similarity between human fixations and the predicted saliency map
- Chi-squared distance between histograms

$$d(x,y) = \frac{1}{2} \sum_i \frac{(x_i - y_i)^2}{x_i + y_i}$$



Equation credit : http://mathoverflow.net/questions/103115/distance-metric-between-two-sample-distributions-histograms
Image credit : https://upload.wikimedia.org/wikipedia/commons/6/6b/Roccurves.png

Results



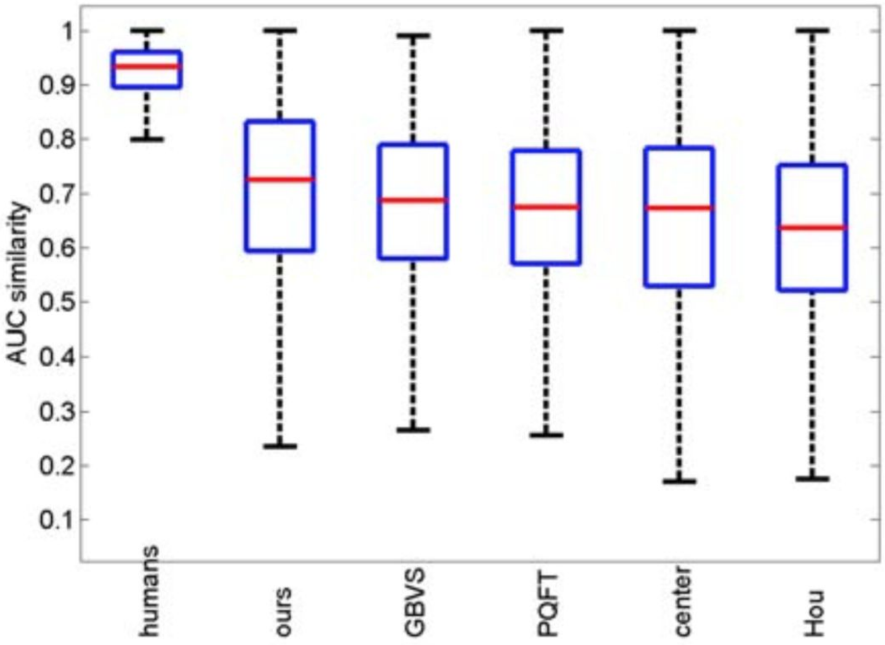


Image credit : Rudoy et al

Discussion

- In paper authors mention that AUC considers the saliency results only at the locations of the ground truth fixation points.
- This will only give true positives and false negatives
- AUC ROC needs true negative and false positives as well. How is AUC computed without them?

Ablation results

	All cues	No motion cues	No inter-frame cues	No semantic cues	No static cues
χ^2	0.313	0.322	0.326	0.347	0.385

- Dropping static or semantic cues results in big drop

Results snapshot : Rudoy et al

More discussion points

- Why 15 frames?
  This parameter is based on typical time taken by human subjects to adjust gaze on a new image.
- Across scene-cuts, the content can change arbitrarily. Use in-shot transitions?
- The model needs dataset with human gaze and video to train
- Why does dense estimation (without candidate selection) give lower accuracy?
  Not clearly mentioned in the paper. Possible reason : candidate based model is able to model the transition probabilities better. The dense model gets confused due to large number of candidates.

More discussion points

- How can we capture gaze transitions within a shot?
- Relation between saliency and memorability
  We can reasonably expect saliency and memorability to be correlated.
- What is the breakdown between failure cases for this model?
- Besides DIEM and CRCNS, are there other datasets that could be used to experiment video saliency
  - http://saliency.mit.edu/datasets.html
- Saliency to evaluate memorability?