Space-Variant Descriptor Sampling for Action Recognition Based on Saliency and Eye Movements

Eleonora Vig[1], Michael Dorr[2], and David Cox[1]

[1]The Rowland Institute at Harvard, Harvard University, [2]Schepens Eye Research Institute, Harvard Medical School

Action Recognition in Videos – State of the Art

- Bag of Features video representations have become popular
- descriptor sampling is either:
  - interest point based: e.g. SIFT, SURF; few stable descriptors only
  - dense grid-based: provides robust recognition → preferred
- drawbacks of dense sampling: must process a large number of (often irrelevant) features → huge computational load

Motivation – Biological Inspiration

- space-variant processing of the human visual system:
  - first, potentially relevant (i.e. salient) regions are detected
  - further detailed processing happens only at salient locations
- **goal**: use saliency either to emphasize the most informative parts of the visual scene or to limit the processing to them

Saliency Masks

- **central mask**: filmmakers place the subject of interest in the center → "distance of pixel to center" as simplest saliency measure
- **analytical saliency mask**: saliency model built on the structure tensor and its geometric invariants
  - for a video $f(x, y, t)$ the structure tensor $\mathbf{J}$ is:

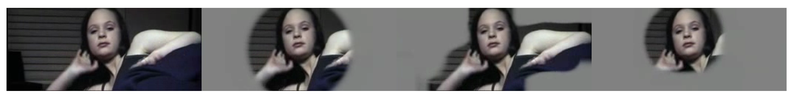
$$\mathbf{J} = \int_\Omega \nabla f \otimes \nabla f \, \mathrm{d}\Omega = \int_\Omega \begin{bmatrix} f_x^2 & f_x f_y & f_x f_t \\ f_x f_y & f_y^2 & f_y f_t \\ f_x f_t & f_y f_t & f_t^2 \end{bmatrix} \mathrm{d}\Omega$$

  - $\mathbf{J}$'s geometric invariants characterize typical video structures (uniform regions, edges, corners, transient corners):

$$H = 1/3 \ \mathrm{trace}(\mathbf{J})$$
$$S = M_{11} + M_{22} + M_{33}$$
$$K = |\mathbf{J}|$$

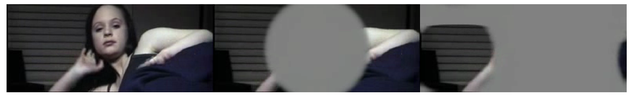
S

  - predict well eye movements on naturalistic videos (Vig et al. PAMI'12)
- **empirical saliency mask**: a "ground truth" saliency map determined by measuring where humans actually look in the video → fixation density map: superposition of Gaussians centered at each gaze sample



Control Masks (non-biological)

- **peripheral mask**: central mask inverted
- **random uniform sampling** from a dense grid
- **randomly-offset empirical saliency mask**: simulates random gaze patterns, while disrupting their relationship with the movie content



Video Representations

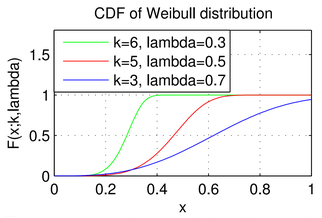
- **HOGHOF with dense sampling** (Laptev et al., CVPR'08)
- **Dense Trajectories** (Wang et al., CVPR'11): densely sampled points tracked using an optical flow field; 4 descriptors: trajectory shape, HOG, HOF, and Motion Boundary Histograms
- **Stacked Convolutional Independent Subspace Analysis**: features learnt through deep learning techniques (Le et al., CVPR'11)

Descriptor Pruning and Feature Combination

- prune the descriptor set based on a saliency mask of the video
  - use the cumulative distribution function of the Weibull distribution

$$F(x; k, \lambda) = 1 - e^{-(x/\lambda)^k}$$

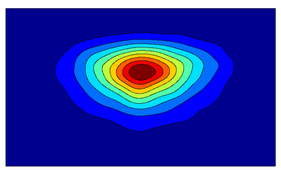
with $k > 0$ shape and $\lambda > 0$ scale parameters, to sample descriptors with certain probability



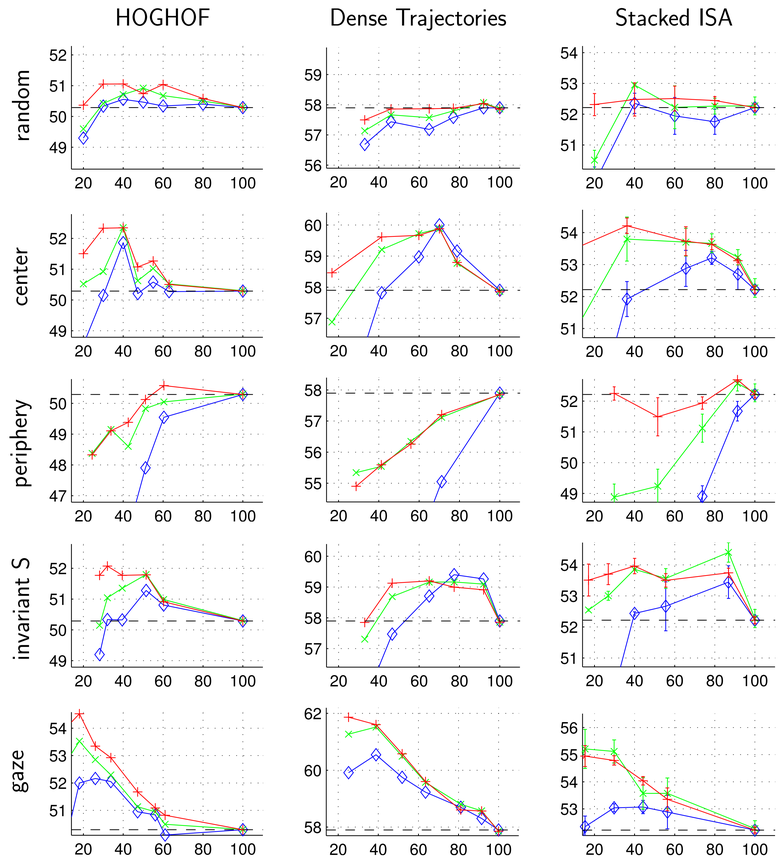
- complement the densely extracted BoF representations with saliency-based descriptor sets:
  - concatenate the two codebook-histograms
  - adaptive feature combination with Multiple Kernel Learning (MKL)

Eye Movement Data Set for Hollywood2

- > 5 hours Hollywood clips: 823 training and 884 test videos; 12 action classes
- binocular eye movements recorded at 1000 Hz from 3 subjects whose task was to identify the action(s) (gaze set is publicly available)



Results



- x-axis: % descriptors kept, y-axis: mean Average Precision
- Legend: prune, concatenate, MKL, dashed line – baseline (no pruning)
- adopted the BoF processing pipeline of Wang et al., BMVC'09

Conclusions

- many descriptors are unnecessary: discarding up to 70% has no effect
- mimicking visual attention improves action recognition: Dense Trajectories enhanced with saliency-based descriptor sampling achieves best mAP (60%) on Hollywood2 to date
- collected eye movements to probe the limits of saliency-based pruning (62%)
- feature combination is beneficial: separate representations for coarse scene gist and detailed foveal view of the scene
- strong center bias in man-made videos

vig@fas.harvard.edu