Analyzing bottom-up saliency in natural movies

Eleonora Vig[1], Michael Dorr[2], and Erhardt Barth[1]

[1]Institute for Neuro- and Bioinformatics, University of Lübeck, Germany, http://www.inb.uni-luebeck.de
[2]Schepens Eye Research Institute, Dept. of Ophthalmology, Harvard Medical School, USA
{vig|barth}@inb.uni-luebeck.de, michael.dorr@schepens.harvard.edu

Data & State-of-the-art

Motivation

We investigate the contributions of local spatio-temporal variations of image intensity to saliency. To measure different types of variations, we use invariants of the structure tensor. Considering a video to be represented in spatial axes (x,y), and temporal axis t, the n-dimensional structure tensor (nD-ST) can be evaluated for different combinations of axes (2D- and 3D-ST) and also for the (degenerate) case of only one axis (1D-ST).

Experimental setup

A large dataset of ~40,000 saccades was obtained from 54 human subjects free-viewing 18 high-resolution movie clips of real-world outdoor scenes of ~20 sec durations each (1280x720 pixels, 29.97 fps, subtending 48x27 deg of visual angle). The saccade landing points were used to label image regions as attended. For the non-attended class, we shuffled the movies and their scanpaths, thus eliminating the central fixation bias.

State-of-the-art

The performance of two standard models of bottom-up saliency on this dataset:

1. Itti & Koch (Itti et al. 1998): 0.644 ROC score. The most well known model of bottom-up attention; inspired by the Feature Integration Theory.
2. SUNDAy (Zhang et al. 2009): 0.635 ROC score. Uses a Bayesian framework. Novelty is defined as the self-information of the visual features; natural statistics are learned from previous examples, not only on the current video.

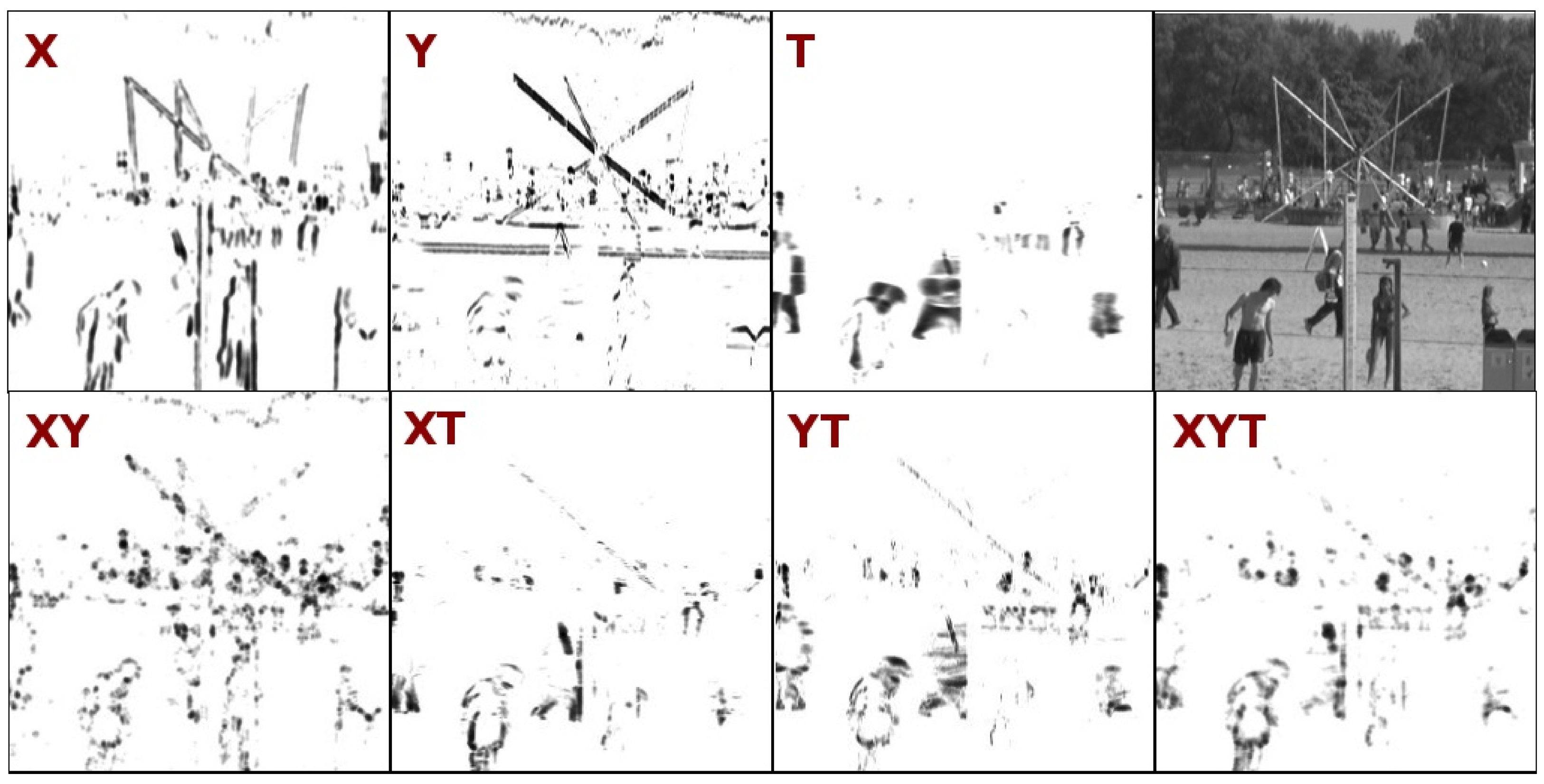
Saliency prediction

Representations

For saliency prediction, we evaluate the intrinsic dimension (iD) of the visual input. The iD denotes the number of degrees of freedom necessary to describe locally a signal. To estimate the iD, we use the structure tensor (ST) which captures statistics of the spatial and/or temporal derivatives at each pixel in the video. The intrinsic dimension of a movie region corresponds to the rank of the ST and can be obtained from ST's symmetric invariants. The invariants correspond to the minimum iD of the region. The scale on which the ST is evaluated depends on the bandwidth of the derivative operators (and the filter kernel omega), therefore computations are performed on a spatio-temporal multiresolution pyramid (see Fig. on the right).

Below: ω is a (spatial and/or temporal) filter kernel, $f_x = \delta f/\delta x$ denote partial derivatives, and the λ_i are the eigenvalues of the structure tensor J.

n	nD-Structure Tensor	Invariants (eigendecomposition of J_{nD})		Dimensions & ROC scores
1	$J_{1D} = \omega * f_x^2$	$H = \lambda_1$	$iD = 1$	$x \rightarrow 0.621$ $y \rightarrow 0.617$ $t \rightarrow 0.623$
2	$J_{2D} = \omega * \begin{pmatrix} f_x^2 & f_x f_t \\ f_x f_t & f_t^2 \end{pmatrix}$	$H = \lambda_1 + \lambda_2$ $\boxed{K = \lambda_1 \lambda_2}$	$iD \geq 1$ $iD = 2$	$xy \rightarrow 0.639$ $xt \rightarrow 0.637$ $yt \rightarrow 0.656$
3	$J_{3D} = \omega * \begin{pmatrix} f_x^2 & f_x f_y & f_x f_t \\ f_x f_y & f_y^2 & f_y f_t \\ f_x f_t & f_y f_t & f_t^2 \end{pmatrix}$	$H = \lambda_1 + \lambda_2 + \lambda_3$ $S = \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3$ $\boxed{K = \lambda_1\lambda_2\lambda_3}$	$iD \geq 1$ $iD \geq 2$ $iD = 3$	$xyt \rightarrow 0.673$

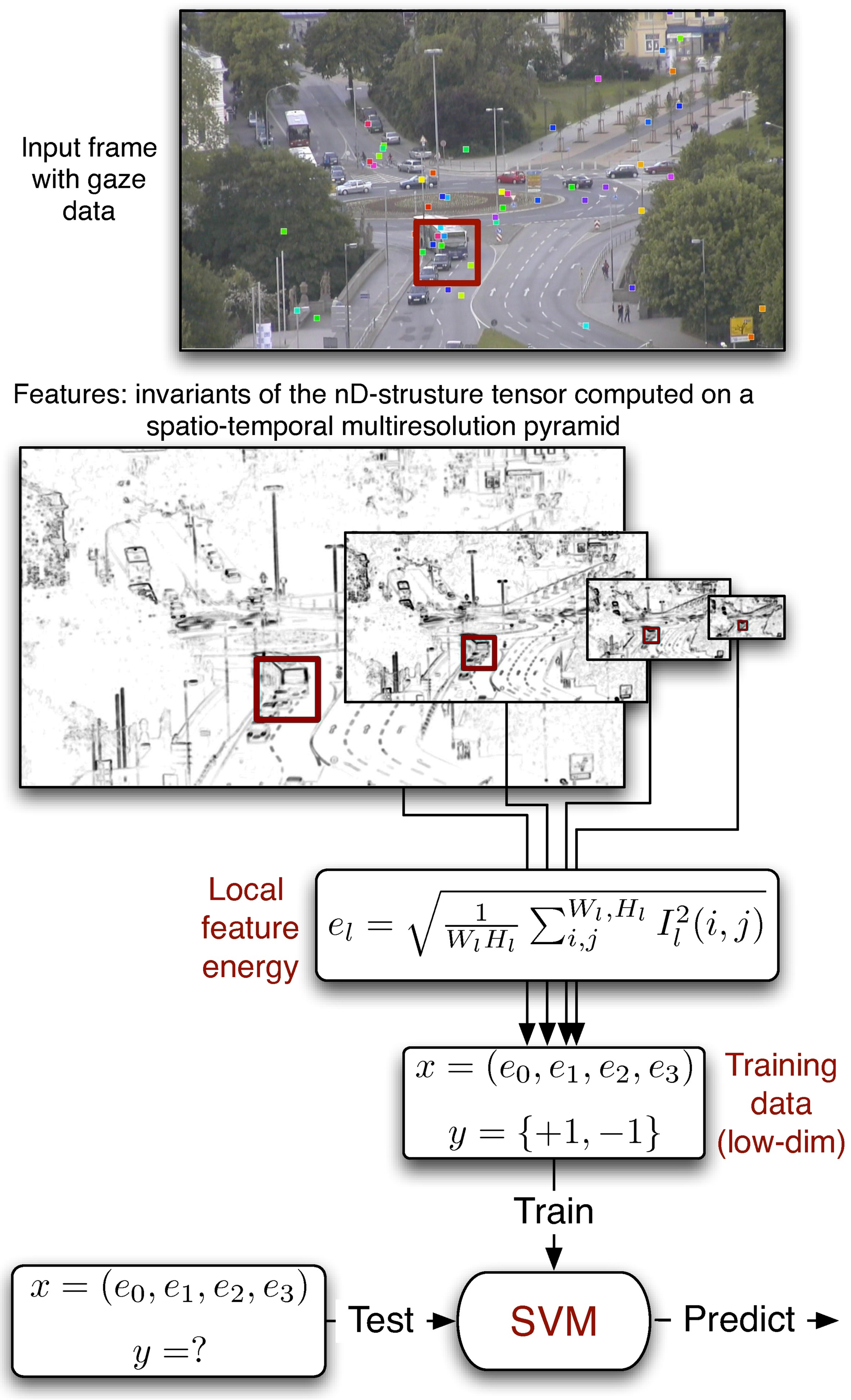


Top row: invariant H of the 1D structure tensor computed along the individual dimensions x, y, t; original frame also shown.

Bottom row (from left): invariant K of a 2D-ST computed along the axes (x,y), (x,t), and (y,t); below the original image: invariant K of a 3D-ST along all three axes.

The classifier

Using eye movements, movie regions are labelled as attended and non-attended. Image features (invariants) are extracted on multiple scales. For a neighbourhood around each location, the average feature energy is computed on each pyramid scale. An SVM is trained on the energy vectors and is then used to predict the saliency of an unseen video region.



Discussion & Summary

- We show that the 3D-ST is optimal (average ROC score of 0.673), i.e. the most predictive regions of a movie are those where intensity varies along all spatial and temporal directions.
- Analyzing two-dimensional variations, the 2D-ST evaluated on the axes (y,t) gave the best score (0.656), followed by (x,y) (0.639), and (x,t) (0.637).
- Bottom-up saliency is therefore determined by spatio-temporal variations of image intensity rather than spatial or temporal variations.
- The proposed model (3D-ST) demonstrates significant improvement over the selected baseline models with ROC scores 0.644 (Itti and Koch) and 0.635 (SUNDAy).

References

[1] E. Vig, M. Dorr, T. Martinetz, and E. Barth. A learned saliency predictor for dynamic natural scenes. In: Proceedings of the 20th Int. Conference on Artificial Neural Networks, Greece, 2010.

[2] E. Vig, M. Dorr, and E. Barth. Efficient visual coding and the predictability of eye movements on natural movies. Spatial Vision 22(5):397-408, 2009.

[3] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11):1254-1259, 1998.

[4] L. Zhang, M.H. Tong, G.W. Cottrell. SUNDAy: Saliency using natural statistics for dynamic analysis of scenes. In: Proceedings of the 31st Annual Cognitive Science Conference, Netherlands, 2009.

Our research has received funding from the European Commission within the GazeCom project (IST-C-033816) of the FP6.
All views herein are those of the authors alone; the European Community is not liable for any use made of the information.