Predictability of eye movements analyzed in a machine learning framework

Eleonora Vig, Michael Dorr, and Erhardt Barth
Institute for Neuro- and Bioinformatics, University of Lübeck, Germany
{vig|dorr|barth}@inb.uni-luebeck.de, http://www.inb.uni-luebeck.de

Introduction

Motivation

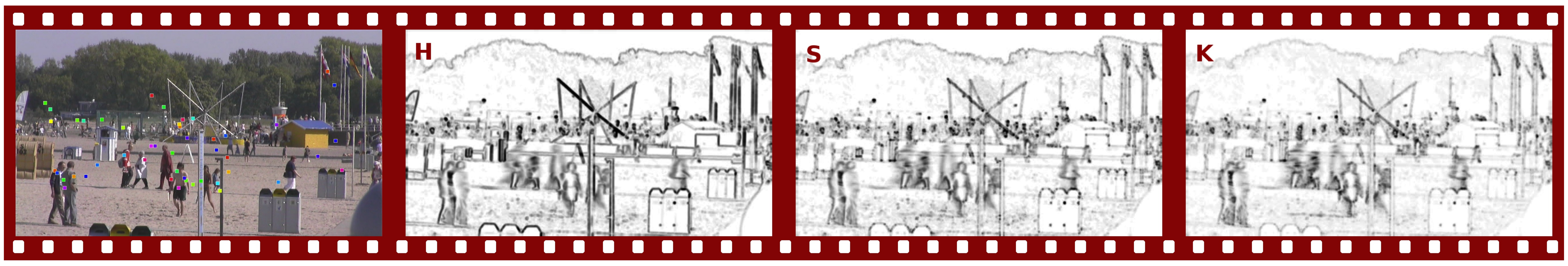
We investigate the extent to which a simple model of low-level saliency based on local spectral energy computed on different visual representations can predict saccade targets in natural dynamic scenes. Our objective is to learn transformations that alter the saliency distribution of the scene in real time, thus implementing gaze guidance [1].

Experimental setup

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

Representations

It is well known that changes in space and in time of the visual input often attract attention. The intrinsic dimension (ID) is a simple and unbiased way to encode the spatio-temporal signal change of the visual input. It denotes the number of degrees of freedom that are necessary to represent the signal. A movie can be locally $i0D$, $i1D$, $i2D$, or $i3D$.



Eye Movement Prediction

Estimating the ID

We use the structure tensor defined in terms of the spatio-temporal gradient of the image intensity function (f_x, f_y, f_t are partial derivatives):

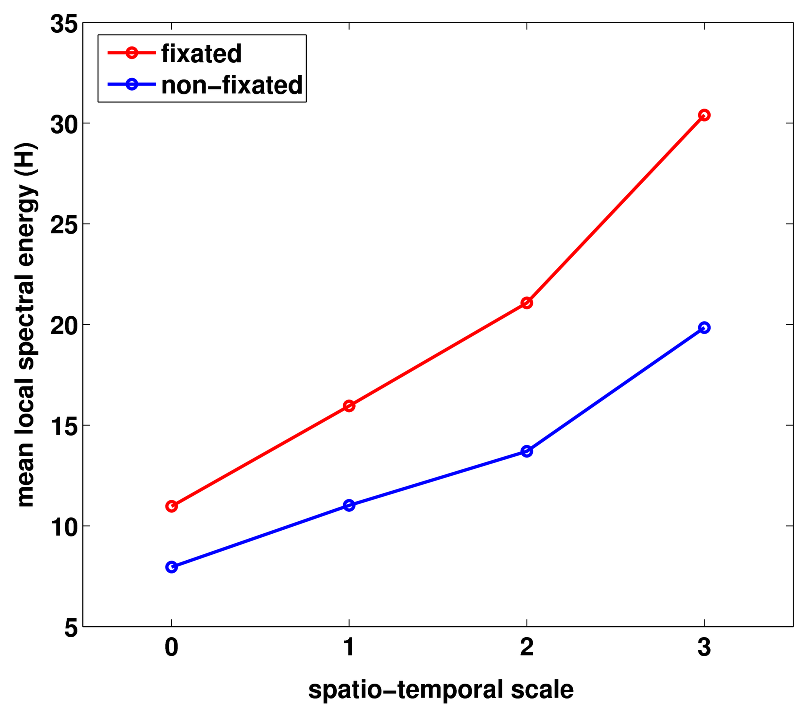
$$J = \int_\Omega \begin{bmatrix} f_x^2 & f_x f_y & f_x f_t \\ f_x f_y & f_y^2 & f_y f_t \\ f_x f_t & f_y f_t & f_t^2 \end{bmatrix} d\Omega$$

Our saliency measures (I) are its symmetric invariants (M_{11}, M_{22}, M_{33} are minors of J):

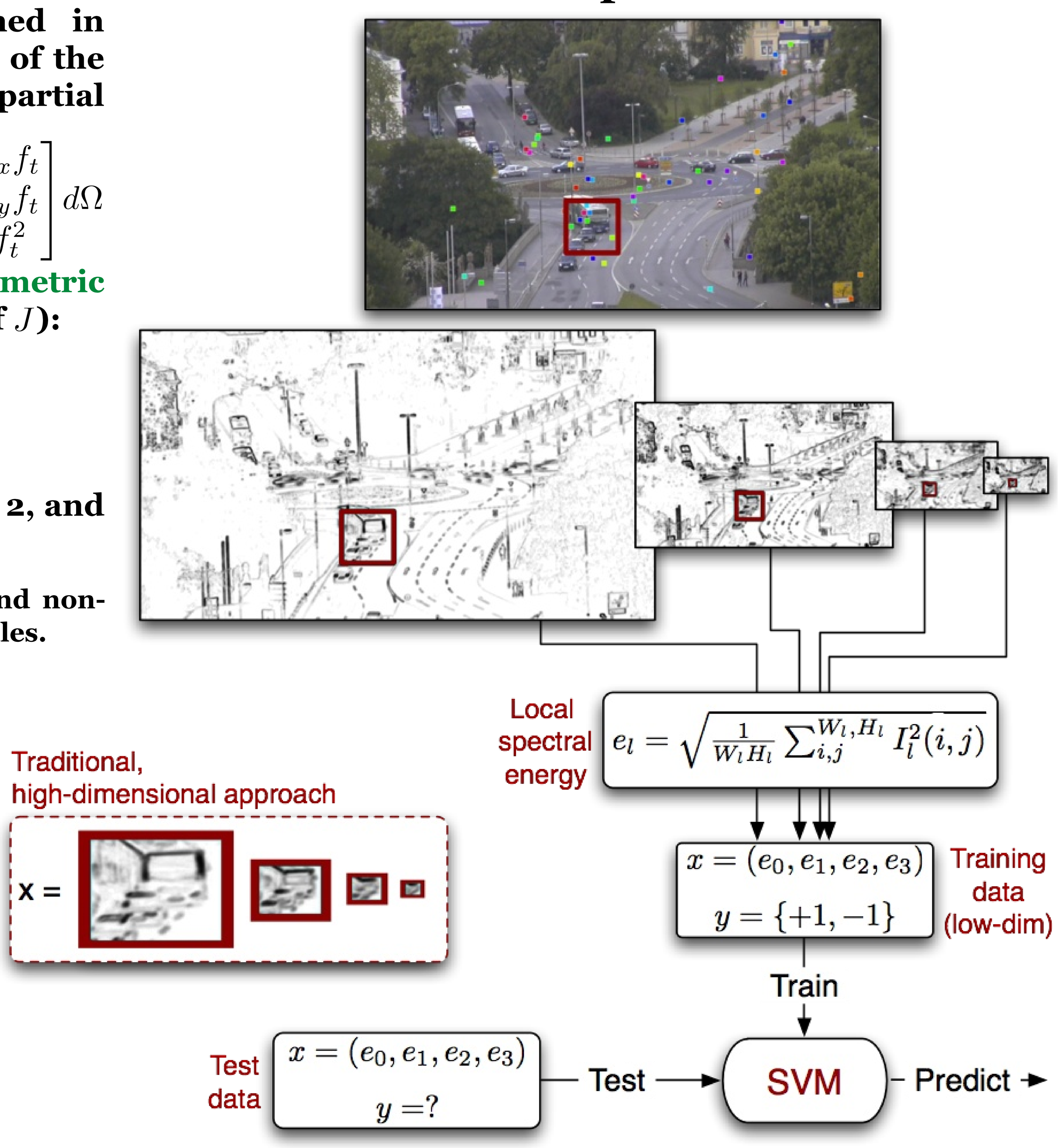
$$\begin{aligned} H &= 1/3\ \mathrm{trace}(J) \\ S &= |M_{11}| + |M_{22}| + |M_{33}| \\ K &= |J| \end{aligned}$$

If $K \neq 0$, the ID is 3, if $S \neq 0$ it is at least 2, and if $H \neq 0$ it is at least 1.

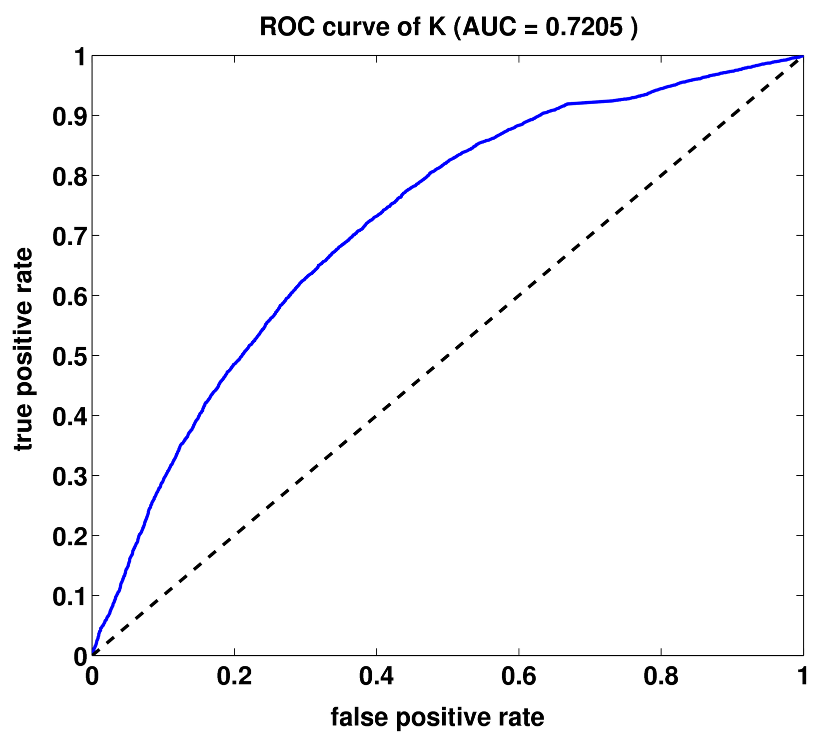
Average local spectral energy at attended and non-attended locations over 4 spatio-temporal scales.



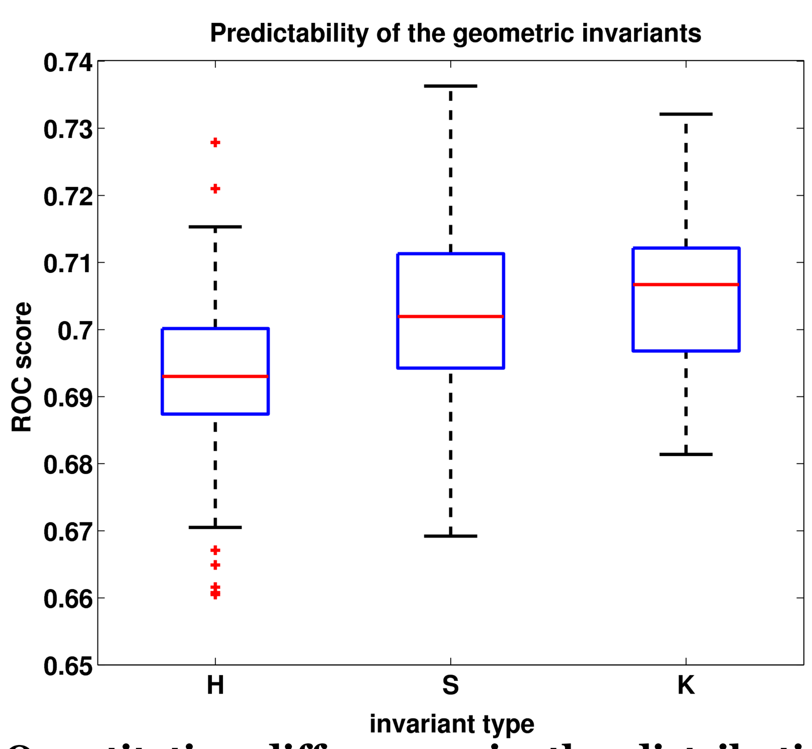
Prediction procedure:



1. The geometric invariants are computed on multiple scales of an isotropic spatio-temporal Gaussian pyramid.
2. To obtain equally sparse representations, H, S, and K are adaptively thresholded.
3. Local spectral energy is extracted in the neighborhood of a certain size of each location on these scales.
4. We obtain for each location a feature vector of the same dimensionality as the number of spatio-temporal pyramid levels.
5. To quantify the extent to which eye movements can be predicted, we use a support vector machine whose parameters are found by cross-validation.
6. ROC analysis is used to test the prediction performance on unlabeled test data.



Results & Summary



Quantitative differences in the distribution of prediction rates for invariants H, S, and K (window size of ~5 deg).

The predictability of eye movements

- is higher (AUC of 0.71) than previously reported results on both static and dynamic scenes.
- is high although little information is used: only one feature (local spectral energy) per movie patch and per scale.
- correlates with the intrinsic dimension: the higher the intrinsic dimension (S and K) the higher the predictive power.
- increases to 0.8 AUC with a moderate increase in the number of dimensions (e.g. anisotropic Gaussian pyramid).

References

[1] E. Barth, M. Dorr, M. Böhme, K. R. Gegenfurtner, and T. Martinetz. Guiding the mind's eye: improving communication and vision by external control of the scanpath. In: Human Vision and Electronic Imaging, volume 6057 of Proc. SPIE. B. E. Rogowitz, T. N. Pappas, and S. J. Daly (Eds.) (2006).

[2] 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).

[3] W. Kienzle, B. Schölkopf, F. A. Wichmann, and M. O. Franz. How to find interesting locations in video: a spatiotemporal interest point detector learned from human eye movements. In Proceedings of the 29th Annual Symposium of the German Association for Pattern Recognition (DAGM 2007), 405-414. Springer Verlag (2007).

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.