Contribution of Spatio-Temporal Intensity Variation to Bottom-up Saliency

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

[1] Institute for Neuro- and Bioinformatics, University of Lübeck,
Ratzeburger Allee 160, 23538 Lübeck, Germany
{vig,barth}@inb.uni-luebeck.de
[2] Schepens Eye Research Institute, Dept. of Ophthalmology, Harvard Medical School,
20 Staniford Street, Boston, MA 02114, USA
michael.dorr@schepens.harvard.edu

Abstract. We investigate the contribution of local spatio-temporal variation of image intensity to saliency. To measure different types of variation, we use the geometrical invariants of the structure tensor. With a video represented in spatial axes x and y and temporal axis t, the n-dimensional structure tensor can be evaluated for different combinations of axes (2D and 3D) and also for the (degenerate) case of only one axis. The resulting features are evaluated on several spatio-temporal scales in terms of how well they can predict eye movements on complex videos. We find that a 3D structure tensor is optimal: the most predictive regions of a movie are those where intensity changes along all spatial and temporal directions. Among two-dimensional variations, the axis pair yt, which is sensitive to horizontal translation, outperforms xy and xt by a large margin, and is even superior in prediction to two baseline models of bottom-up saliency.

Key words: video saliency, eye movements, intrinsic dimension, structure tensor, natural dynamic scenes

1 Introduction

Visual attention, the selective processing of visual information, is an important component of biologically-inspired machine vision systems. Computational models of attention have proven to be invaluable in identifying points of interest within a scene and e.g., through that, enabling the otherwise time- and resource-consuming image processing to focus only on these potentially relevant scene locations.

In human vision, the extent to which a certain scene region captures the viewers' attention, i.e. its level of *salience*, is determined by two different kinds of mechanisms. On the one hand, basic visual properties, such as motion, contrast, and colour influence where we direct our gaze. On the other hand, top-down knowledge, i.e. our goals and interests, also modulate attentional selection. The relative contribution of the two mechanisms is still under debate; however, due to the involuntary and computationally more tractable nature of stimulus-driven

attention, much work has focused on the bottom-up factors that drive eye movements.

Of major importance was the recognition that scene statistics at the centre of fixation differ significantly from those at random, control locations. Studies have shown that attended regions have high luminance-contrast [6, 7], and found regularities in the higher-order statistics (e.g. high edge density [5]). Knowledge about such distinct image properties has been then used to build models of saliency that successfully predict human fixations in natural scenes, e.g. [4, 2, 9].

Another key finding is related to the region's degree of spatial (and temporal) variance. It shows for images that intrinsically two-dimensional scene structures (i.e. of higher spatial variance), such as edges and curved lines, have a higher probability to be fixated [5]. In previous work, we could demonstrate for videos that features that change over space and in time also tend to be more salient. We found that the predictability of eye movements correlates with the intrinsic dimension: the higher the intrinsic dimension the higher the predictive power [8, 1].

In the present study, we extend our previous analysis by quantifying the contribution of local spatio-temporal variation of image intensity to saliency. To measure different kinds of variation, we compute, for a set of natural outdoor videos, invariants of the n-dimensional structure tensor ($1 \leq n \leq 3$). Considering a video to be represented in spatial axes (x, y) and temporal axis t, the nD structure tensor is evaluated for different combinations of axes (2D and 3D) and also for the (degenerate) case of only one axis. To obtain a simple measure of bottom-up saliency, we use the symmetric invariants of the structure tensors, which we compute on several spatio-temporal scales. Finally, the resulting simple representations are evaluated and compared with two prototypical saliency models of dynamic scenes in terms of how well they can predict eye movements on videos.

2 Invariants of the n-dimensional Structure Tensor

It has been previously shown that eye movement predictability correlates with the *intrinsic dimension* (iD), i.e. with the number of spatio-temporal directions in which the video changes locally. A classical method to estimate the intrinsic dimension is to consider the rank of the *structure tensor*. Given a grayscale video $f : \mathbb{R}^3 \to \mathbb{R}$, the structure tensor captures signal variations based on the spatial and temporal derivatives at each pixel. For three-dimensional data, i.e. the spatio-temporal volume of the video, usually a three-dimensional structure tensor is defined. However, on subspaces of the video volume (e.g. combinations of two axes, or even considering the degenerate case of a single axis only) 1D or 2D structure tensors can be constructed.

Here, we formalize the problem for the two-dimensional structure tensor $\mathbf{J_2}$, considering only the vertical spatial dimension y and the temporal dimension t. The generalization of the formulas for the n-dimensional case ($1 \leq n \leq 3$) is given in Table 1. For the axis combination yt $\mathbf{J_2}$ is defined as

$$\mathbf{J}_2 = \omega(y,t) * \begin{pmatrix} f_y^2 & f_y f_t \\ f_y f_t & f_t^2 \end{pmatrix}, \tag{1}$$

where $\omega(y,t)$ is a Gaussian smoothing function and f_y and f_t stand for the first order partial derivatives $\delta f/\delta y$ and $\delta f/\delta t$. The scale on which the structure tensor is evaluated depends on the bandwidth of the filter kernel $\omega(y,t)$ and the derivative operators. Therefore, computations are performed on a spatio-temporal multiresolution pyramid.

The intrinsic dimension is, in practice, obtained from the symmetric invariants of the structure tensor:

$$\begin{aligned} H &= 1/2 \text{ trace}(\mathbf{J_2}) = \lambda_1 + \lambda_2 \\ K &= |\mathbf{J_2}| \qquad\qquad\quad = \lambda_1 \lambda_2 \end{aligned} \tag{2}$$

where λ_i denote the eigenvalues of $\mathbf{J_2}$. Regions where $H > 0$ are at least intrinsically one-dimensional ($iD \geq 1$), e.g. non-vertical stationary edges, vertically translating edges, and uniform regions that change in time, whereas $K > 0$ indicates an $i2D$ feature such as yt corners (changing motion) and structures that appear or disappear in yt, which correspond to non-vertical translation.

Table 1. n-dimensional structure tensors and their invariants, which correspond to the minimum intrinsic dimension (iD) of a region. Invariants that encode features of higher iD are in general better predictors of eye movements; therefore, they are used for further analysis (these are marked with a box).

n	nD Structure Tensor	Invariants (eigendecomposition of $\mathbf{J_n}$)	
1	$\mathbf{J_1} = \omega(u) * f_u^2$ $u \in \{x,y,t\}$	$\boxed{H = \lambda_1}$	$(iD = 1)$
2	$\mathbf{J_2} = \omega(u,v) * \begin{pmatrix} f_u^2 & f_u f_v \\ f_u f_v & f_v^2 \end{pmatrix}$ $u,v \in \{x,y,t\}, u \neq v$	$H = \lambda_1 + \lambda_2$ $\boxed{K = \lambda_1 \lambda_2}$	$(iD \geq 1)$ $(iD = 2)$
3	$\mathbf{J_3} = \omega(x,y,t) * \begin{pmatrix} f_x^2 & f_x f_y & f_x f_t \\ f_x f_y & f_y^2 & f_y f_t \\ f_x f_t & f_y f_t & f_t^2 \end{pmatrix}$	$H = \lambda_1 + \lambda_2 + \lambda_3$ $S = \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3$ $\boxed{K = \lambda_1\lambda_2\lambda_3}$	$(iD \geq 1)$ $(iD \geq 2)$ $(iD = 3)$

3 Prediction of Eye Movements with Tensor-based Approaches

Having reviewed simple tensor-based video representations that characterize different types of spatio-temporal changes, we now quantitatively compare their power in predicting eye movements on complex natural videos.

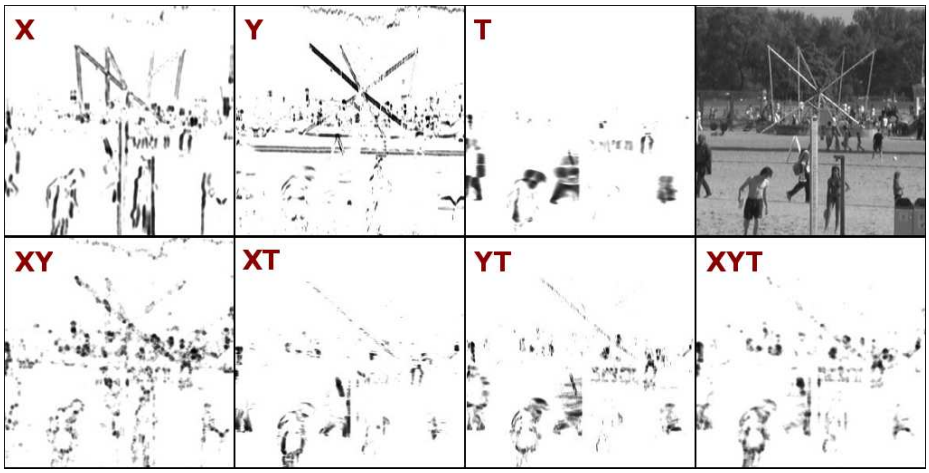

Fig. 1. Top row (from left): H of $\mathbf{J_1}$ computed along the individual axes x, y, and t; original frame also shown. Bottom row (from left): K of $\mathbf{J_2}$ computed along the axes xy, xt, and yt; below the original image: K of $\mathbf{J_3}$ along all three axes.

For our evaluation, we used a public data set[1] [3] that consists of 18 high-resolution movie clips (1280 by 720 pixels, 29.97 fps, about 20 s duration each) of natural outdoor scenes, and the gaze data of 54 human subjects freely viewing these videos. From the raw gaze data, collected with an Eye Link II eye tracker at 250 Hz, about 40,000 saccades were extracted. All movies were cropped to the same size along the spatial axes (preserving the central 600 by 600 pixels), to make the resulting space-time cubes rotation-invariant with regard to size (because movies had 600 frames). The total number of saccades that remained after the cropping was 24,370.

Invariants that encode features of higher intrinsic dimensionality were shown to be better predictors of eye movements; therefore, here only these were considered (see Table 1). For each video, we computed the invariants of the tensors $\mathbf{J_1}$, $\mathbf{J_2}$, and $\mathbf{J_3}$ along all possible dimensions/combinations of dimensions. See Figure 1 for still shots from a movie and the corresponding invariants. The above invariants were computed on each scale of an anisotropic spatio-temporal multiresolution pyramid with $S = 2$ spatial and $T = 2$ temporal scales, in which each spatial pyramid was decomposed further into its temporal bands.

To determine how well the different representations can predict the saliency level of video regions, next, we labelled areas in the videos as salient and non-salient. The class of salient locations is well defined by the human fixations (more precisely by the saccade landing points). To obtain the non-salient class, a number of biases need to be addressed (e.g. the central fixation bias, the tendency of observers to fixate more in the centre of the display). A common approach in the human vision literature, which we also follow here, is to shuffle

[1] `http://www.inb.uni-luebeck.de/tools-demos/gaze/`

scanpaths: the non-attended locations of a movie are chosen using randomly selected scanpaths from the other movies.

For eye movement prediction, instead of using directly the feature response at these locations, one must consider a spatio-temporal neighbourhood centred around fixations. This is partly accounted for by the image pyramid; however, we further consider a spatial window (of 32 pixels, i.e. about 1.2 deg, on the highest pyramid level), as uncertainty in the measurements is higher in the spatial domain. On each pyramid level, we compute the window's *energy*, i.e. the root-mean-square of the feature values (i.e. invariants) in the window. Thus, we obtain for each salient and non-salient video location a low-dimensional vector of *feature energies* computed on the different pyramid levels (procedure detailed in [8]).

Finally, the predictive power of the different representations is assessed by evaluating (through ROC analysis) the performance of one-dimensional maximum-likelihood classifiers when the feature energies from the single pyramid levels are used as inputs to the decision algorithm. In Table 2, we report average ROC scores (over the 18 movies) obtained for the "most predictive" scale (i.e. the pyramid level with the highest average ROC score).

For comparison, the saliency maps computed by two state-of-the-art algorithms for dynamic scenes (Itti & Koch and SUNDAy [4, 9]) are treated as maximum-likelihood classifiers for discriminating between fixated and not fixated video regions. By thresholding these maps, movie regions above the threshold are classified as salient. A systematic variation of the threshold parameter gives us a single ROC curve per movie and model. The averaged ROC scores over all videos are reported in Table 2.

Table 2. Average ROC scores of the different models and representations.

Model		ROC score	Model		ROC score	Model	ROC score
	x	0.621		xy	0.639	$\mathbf{J_3}(xyt)$	0.673
$\mathbf{J_1}$	y	0.617	$\mathbf{J_2}$	xt	0.637	Itti & Koch	0.644
	t	0.623		yt	0.656	SUNDAy	0.635

4 Discussion and Conclusion

With an average ROC score of 0.673 the three-dimensional structure tensor $\mathbf{J_3}$ is optimal, suggesting that the most predictive regions of a movie are indeed those where intensity varies along all spatial and temporal dimensions. Surprisingly, the second best predictor operates on the axis pair yt; this predictor is sensitive for horizontal translations, which are most common in typical natural scenes. $\mathbf{J_2}$ evaluated on the axes yt outperforms xy and xt by a large margin (with an ROC score of 0.656 compared to 0.639 and 0.637, respectively), and is even superior

to the two baseline models with ROC scores 0.644 (Itti & Koch) and 0.635 (SUNDAy), which incorporate a number of different features such as colour, contrast, and orientation. Although one-dimensional variations perform worst (with $\mathbf{J_1}$ along the vertical axis giving the lowest score – 0.617), their average prediction rate is significantly higher than chance (ROC score of 0.5).

Our results can be used to choose efficient active vision strategies. Under the assumption that the human visual system is near-perfectly optimized for natural environments, the spatio-temporal structure tensor J_3 thus picks the most informative regions. However, with our data, it is now also possible to choose which dimension should be sacrificed for faster computation in resource-limited systems, e.g. in an embedded real-time module of a robot with active vision sensors: for natural environments, the axis pair yt is more informative than xy or xt.

Future work will investigate the predictive power of other tensor representations, such as the Hessian matrix or the energy tensor, and implement the proposed simple saliency models for a real-time system attached to a camera.

Acknowledgments. We would like to thank Karl Gegenfurtner: data were collected in his lab at the Dept. of Psychology of Giessen University. Our research has received funding from the European Commission within the project GazeCom (contract no. IST-C-033816, www.gazecom.eu) of the 6th Framework Programme. All views expressed herein are those of the authors alone; the European Community is not liable for any use made of the information.

References

1. Barth, E., Dorr, M., Vig, E., Pomarjanschi, L., Mota, C.: Efficient Coding and Multiple Motions. Vision Research, DOI: 10.1016/j.visres.2010.08.011 (2010)
2. Bruce, N., Tsotsos, J.K.: Saliency, Attention, and Visual Search: An Information Theoretic Approach. Journal of Vision 9(3), 1–24 (2009)
3. Dorr, M., Martinetz, T., Gegenfurtner, K., Barth, E.: Variability of Eye Movements when Viewing Dynamic Natural Scenes. Journal of Vision 10(10), 1–17 (2010)
4. Itti, L., Koch, C., Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)
5. Krieger, G., Rentschler, I., Hauske, G., Schill K., Zetzsche C.: Object and Scene Analysis by Saccadic Eye-Movements: An Investigation with Higher-Order Statistics. Spatial Vision, 3, 201–214 (2000)
6. Reinagel, P., Zador, A.M.: Natural Scene Statistics at the Centre of Gaze. Network 10, 341–350 (1999)
7. Tatler, B.W., Baddeley, R.J., Gilchrist, I.D.: Visual Correlates of Fixation Selection: Effects of Scale and Time. Vision Research 45, 643–659 (2005)
8. Vig, E., Dorr, M., Barth, E.: Efficient Visual Coding and the Predictability of Eye Movements on Natural Movies. Spatial Vision 22(5), 397–408 (2009)
9. Zhang, L., Tong, M.H., Cottrell, G.W.: SUNDAy: Saliency Using Natural Statistics for Dynamic Analysis of Scenes. In: Proceedings of the 31st Annual Cognitive Science Conference, Amsterdam, Netherlands (2009)