Colour saliency on video

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

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

Abstract. Much research has been concerned with the notion of bottom-up *s*aliency in visual scenes, i.e. the contribution of low-level image features such as brightness, colour, contrast, and motion to the deployment of attention. Because the human visual system is obviously highly optimized for the real world, it is reasonable to draw inspiration from human behaviour in the design of machine vision algorithms that determine regions of relevance. In previous work, we were able to show that a very simple and generic grayscale video representation, namely the geometric invariants of the structure tensor, predicts eye movements when viewing dynamic natural scenes better than complex, state-of-the-art models. Here, we moderately increase the complexity of our model and compute the invariants for colour videos, i.e. on the multispectral structure tensor and for different colour spaces. Results show that colour slightly improves predictive power.

Key words: video saliency, eye movements, intrinsic dimension, multispectral structure tensor

1 Introduction

The human visual system uses a sophisticated approach to efficiently cope with the vast amounts of data that enter the eye and which need to be processed in real time. Only information from a small central fraction of the visual field, the fovea, is processed at high spatial resolution; more peripheral information is processed only at a very coarse scale and is used mainly for action guidance. One particular problem that the human vision system seems to solve surprisingly well is then when and where to direct the fovea via eye movements to sample all relevant aspects of a visual scene.

Early work found that fixated image regions differed from non-fixated regions in their low-level features such as contrast [10] or higher-order statistics [7]. Nevertheless, it is still a matter of debate whether these altered image statistics at fixation are actually causal of eye movements [4], or whether it is high-level objects that draw attention [3].

For machine vision applications and systems, however, the distinction between a causal and a mere correlative contribution of saliency to eye movement guidance is rather philosophical. It is safe to assume that the human visual system is highly optimized for the real world, and thus mimicking its performance will find the most informative regions in a scene. Consequently, many models have been developed for saliency on both static images and videos [6, 5, 8, 13]. Typically, these models first extract a range of biologically-inspired low-level features, such as brightness, colour, contrast, orientation, and motion on multiple spatio-temporal scales, and then fuse this information into a single saliency map that assigns a single value of "interestingness" to each image location.

Contrary to these often complex models with a high number of parameters, in previous work we have successfully modelled eye movements using a simple and very generic video representation: the geometric invariants of the structure tensor that capture the amount of spatio-temporal intensity variation [11]. Based on these invariants, we can derive the intrinsic dimensionality of the video, that is the number of degrees of freedom that are used locally. For example, at a stationary edge, the signal changes in only one spatio-temporal direction (orthogonal to the edge), and thus edges constitute $i1D$ regions; transient corners, on the other hand, change in all directions and are therefore $i3D$. One important finding is that the predictive power increases with intrinsic dimensionality: in other words, corners are more informative than edges, and transient features are more informative than their stationary counterparts. A further, surprising finding is that prediction based on this generic video representation outperforms complex state-of-the-art models [12].

So far, the geometric invariants were only computed in grayscale on the luma channel. In the following, we shall compute the invariants on a multispectral structure tensor in order to investigate whether the incorporation of colour information can improve eye movement predictability in dynamic natural scenes.

2 Methods

2.1 The Multispectral Structure Tensor

To estimate the intrinsic dimension of a given video region Ω, we choose a linear subspace $E \subset \mathbb{R}^3$, of highest dimension, such that

$$\frac{\partial f}{\partial \mathbf{v}} = 0 \text{ for all } \mathbf{v} \in E,$$

with the intrinsic dimension of $f = 3 - \dim(E)$ [9]. E can be estimated as the subspace spanned by the set of unity vectors that minimize the energy functional

$$\varepsilon(\mathbf{v}) = \int_\Omega \left| \frac{\partial f}{\partial \mathbf{v}} \right|^2 \mathrm{d}\Omega = \mathbf{v}^T J \mathbf{v},$$

where the *structure tensor* J [1] is given by

$$J = \int_\Omega \nabla f \otimes \nabla f \mathrm{d}\Omega$$

with the tensor product $\otimes$. Alternatively, we can then write

$$J = \omega * \begin{pmatrix} f_x f_x & f_x f_y & f_x f_t \\ f_x f_y & f_y f_y & f_y f_t \\ f_x f_t & f_y f_t & f_t f_t \end{pmatrix}$$

with a spatio-temporal lowpass filter kernel ω and partial derivatives f_x, i.e. $f_x = \partial f / \partial x$. Therefore, E is the eigenspace associated with the smallest eigenvalue of J, and the intrinsic dimension of f corresponds to the rank of J. Instead of performing an eigenvalue analysis, the intrinsic dimension can also be obtained from the symmetric invariants of J:

$$\begin{array}{rlll} H & = 1/3 \operatorname{trace}(J) & = \lambda_1 + \lambda_2 + \lambda_3 & (iD \geq 1) \\ S & = |M_{11}| + |M_{22}| + |M_{33}| & = \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_1\lambda_3 & (iD \geq 2) \\ K & = |J| & = \lambda_1\lambda_2\lambda_3 & (iD = 3). \end{array}$$

For a multispectral image sequence, we look for the subspace E of highest dimension such that, in Ω,

$$\frac{\partial \mathbf{f}}{\partial \mathbf{v}} = 0 \text{ for all } \mathbf{v} \in E.$$

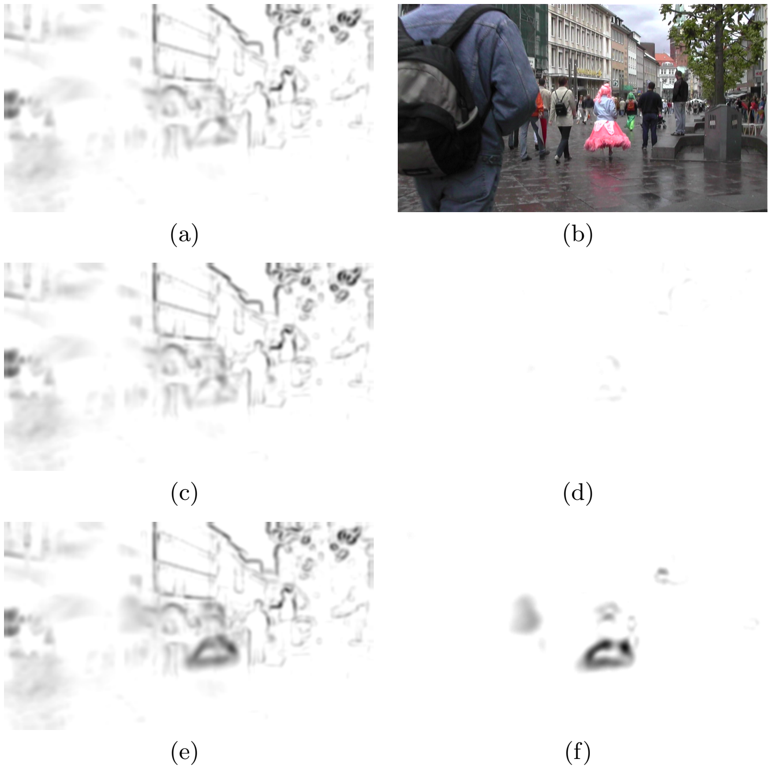
Note that $\mathbf{f}$ is now a vector from $\mathbb{R}^q$ (for an image sequence with q colour channels), so we choose an appropriate scalar product for $\mathbf{y} = (y_1, \ldots, y_q)$ and $\mathbf{z} = (z_1, \ldots, z_q)$ such that $\mathbf{y} \cdot \mathbf{z} = \sum_{k=1}^{q} a_k y_k z_k$, with positive weights a_k that can be used to assign higher importance to certain colour channels, and we arrive at the multispectral structure tensor

$$J = \int_{\Omega} \begin{bmatrix} \|\mathbf{f_x}\|^2 & \mathbf{f_x} \cdot \mathbf{f_y} & \mathbf{f_x} \cdot \mathbf{f_t} \\ \mathbf{f_x} \cdot \mathbf{f_y} & \|\mathbf{f_y}\|^2 & \mathbf{f_y} \cdot \mathbf{f_t} \\ \mathbf{f_x} \cdot \mathbf{f_t} & \mathbf{f_y} \cdot \mathbf{f_t} & \|\mathbf{f_t}\|^2 \end{bmatrix} \mathrm{d}\Omega.$$

In our implementation, we chose 5-tap spatio-temporal binomials for ω and for smoothing the video sequence before taking the derivatives, and J was computed for a spatio-temporally downsampled version of the original video (factor four in space and time). Saliency was then determined as the average energy of the geometric invariants in an 8x8 pixel window around a location.

2.2 Colour spaces

The colour space RGB is commonly used in computer graphics and stores images with red, green, and blue components. Video formats, however, often exploit the reduced colour resolution of the human visual system and thus our original videos had been recorded in the $Y'C_bC_r$ format with one luma and two chroma channels (of halved resolution). When using $Y'C_bC_r$ directly, the dynamic range of the luma channel is much larger than that of the chroma channels, and the contribution of colour to $J_{\mathrm{Y'C_bC_r}}$ is small. We therefore computed the standard deviation of each channel in our set of videos (Y': 124.8; C_b: 39.7; C_r: 44.9) and used their inverse for the weights a_y, a_{cb}, a_{cr} to obtain the colour space $Y'C_bC_r\_weighted$. As a baseline, we computed the invariants in grayscale on the Y' channel only.



(a) (b)

(c) (d)

(e) (f)

Fig. 1. Example frames for colour saliency. a) Invariant H computed on the luma channel only (inverted for legibility). b) Original video frame. c) H on J_{RGB}. d) Absolute difference between a) and c). Even though colour information is represented in RGB, the difference is small. e-f) H on $J_{\mathrm{Y'C_bC_r\text{-}weighted}}$ and absolute difference to a).

2.3 Experimental Data

We used a large eye movement database of about 40000 saccades obtained from 54 subjects watching 18 high-resolution movies (1280 by 720 pixels, 29.97 fps, about 20 s duration each) described in detail elsewhere [2]. For a set of negative examples, we did not generate random data, but shuffled scanpaths on different movies in order to keep the spatio-temporal distribution of positive and negative samples over all movies constant. These samples were then classified with a Maximum Likelihood classifier based on one of the invariants.

3 Results

The ROC scores for the geometrical invariants H, S, and K on the multispectral structure tensor for different colour spaces are shown in Fig. 2. We can replicate the previous result that regions with higher intrinsic dimensionality are also

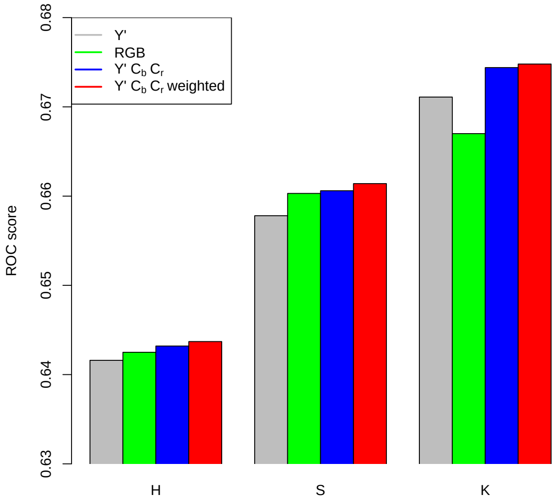


Fig. 2. ROC scores for eye movement predictability of the geometrical invariants of the multispectral structure tensor in different colour spaces. The higher the intrinsic dimension, the higher the predictability ($K > S > H$); saliency on colour video predicts eye movements better than on grayscale video (Y').

more predictive of eye movements ($K > S > H$). Furthermore, the inclusion of colour information improves predictive power, but only slightly. The differences between the different colour spaces are very small, except for the invariant K on RGB, which performs worse even than the grayscale K.

4 Conclusion

We have previously found that a simple, generic model of spatio-temporal intensity variation can predict eye movements on natural videos at least as well as complex state-of-the-art models. In the present manuscript, we have incorporated colour information into our model while maintaining its conceptual simplicity (but at increased computational cost). Results show that indeed colour improves predictive power, but only moderately so. Whether this is due to a ceiling effect or to a relatively small contribution of colour to eye guidance in dynamic natural scenes remains to be determined. Future work will also incorporate further colour spaces, such as HSV, which is particularly sensitive for skin colour, or the perceptually equidistant LAB space.

Acknowledgements Our research has received funding from the European Commission within the GazeCom project (IST-C-033816) of the FP6, and was further supported by NIH grants EY018664 and EY019281. All views herein are those of the authors alone; the European Commission is not liable for any use made of the information.

References

1. J Bigün, G H Granlund, and J Wiklund. Multidimensional orientation estimation with application to texture analysis and optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):775–90, 1991.
2. Michael Dorr, Thomas Martinetz, Karl Gegenfurtner, and Erhardt Barth. Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, 10(10):1–17, 2010.
3. W Einhäuser, M Spain, and P Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):11–26, 2008.
4. Lior Elazary and Laurent Itti. Interesting objects are visually salient. *Journal of Vision*, 8(3):1–15, 2008.
5. Laurent Itti and Pierre Baldi. A principled approach to detecting surprising events in video. In *Proc IEEE Conf on Computer Vision and Pattern Recognition*, pages 631–637, 2005.
6. Laurent Itti, Christof Koch, and Ernst 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.
7. Gerhard Krieger, Ingo Rentschler, Gert Hauske, Kerstin Schill, and Christoph Zetzsche. Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. *Spatial Vision*, 13(2,3):201–214, 2000.
8. Olivier Le Meur, Patrick Le Callet, Dominique Barba, and Dominique Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5):802–817, 2006.
9. Cicero Mota, Ingo Stuke, and Erhardt Barth. The Intrinsic Dimension of Multispectral Images. In *MICCAI Workshop on Biophotonics Imaging for Diagnostics and Treatment*, pages 93–100, 2006.
10. Pamela Reinagel and Anthony M Zador. Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems*, 10:341–350, 1999.
11. Eleonora Vig, Michael Dorr, and Erhardt Barth. Efficient visual coding and the predictability of eye movements on natural movies. *Spatial Vision*, 22(5):397–408, 2009.
12. Eleonora Vig, Michael Dorr, Thomas Martinetz, and Erhardt Barth. A learned saliency predictor for dynamic natural scenes. In K. Diamantaras, W. Duch, and L. S. Iliadis, editors, *ICANN 2010, Part III*, volume 6354 of *Lecture Notes in Computer Science*, pages 52–61, Thessaloniki, Greece, 2010. Springer.
13. Lingyun Zhang, Matthew H. Tong, and Garrison W. Cottrell. SUNDAy: Saliency Using Natural Statistics for Dynamic Analysis of Scenes. In *Proceedings of the 31st Annual Cognitive Science Conference, Amsterdam, Netherlands*, 2009.