Intrinsic Dimensionality Predicts the Saliency of Natural Dynamic Scenes

Eleonora Vig, Michael Dorr, Thomas Martinetz, *Senior Member, IEEE*, and Erhardt Barth, *Member, IEEE*

***Abstract*— Since visual attention-based computer vision applications have gained popularity, ever more complex, biologically-inspired models seem to be needed to predict salient locations (or interest points) in naturalistic scenes. In this paper, we explore how far one can go in predicting eye movements by using only basic signal processing, such as image representations derived from efficient coding principles, and machine learning. To this end, we gradually increase the complexity of a model from simple single-scale saliency maps computed on grayscale videos to spatio-temporal multiscale and multispectral representations. Using a large collection of eye movements on high-resolution videos, supervised learning techniques fine-tune the free parameters whose addition is inevitable with increasing complexity. The proposed model, although very simple, demonstrates significant improvement in predicting salient locations in naturalistic videos over four selected baseline models and two distinct data labelling scenarios.**

***Index Terms*— Computational models of vision, video analysis, computer vision, spatio-temporal saliency, eye movement prediction, intrinsic dimension, visual attention, interest point detection.**

I. Introduction

THE vast amount of visual information available in the world requires selective mechanisms that direct and limit the processing of incoming information to the relevant scene locations. In biological vision, effective attentional processes exist that guide our gaze to informative, or "salient", parts of the visual field. The cognitive processes that underlie visual attention have been extensively investigated both through psychophysical as well as neurophysiological studies. More recently, computational models of attention have been proposed, which are inspired by the findings of these studies, and which attempt to predict where people look when watching complex scenes.

In computer vision, where meaningful descriptions of scenes need to be generated in real time and under computational constraints, the usefulness of such selective processing has been recognized early. Interest point detection (e.g. [1]) is in common use in several areas, such as object recognition, tracking, image and video retrieval, and stereo matching. The connection between visual attention and interest operators, which use local cues to limit the processing to informative content, has been stressed recently, e.g. [2], [3]. As a result, visual attention-based approaches to various computer vision tasks, such as image coding and compression [4], [5], quality assessment [6], image cropping [7], and object recognition [3], [8], [9], gained an increasing popularity. Such models are often preferred over simpler methods from image processing for their enhanced performance and their ability to biologically motivate the major computational steps. However, to grant biological plausibility, sophisticated models are required that rest on several assumptions about perceptual processes, demand high computational costs, and whose results depend on the optimal choice of many free parameters. With such (overly) complex models, however, the possibility of overfitting arises, that is, the model may be "over-tuned" for specific assumptions and data, and therefore fail to generalize. In this paper, we propose a rather simplistic model of bottom-up saliency for dynamic scenes with the aim to keep the number of assumptions (and, implicitly, the number of free parameters) to a minimum. This model is also related to the neurobiological principle of efficient coding [10]. To test our model, we evaluate how well it predicts human eye movements on naturalistic videos both in absolute terms and in comparison with more complex, state-of-the-art saliency models.

The authors are with the Institute for Neuro- and Bioinformatics, University of Lübeck, Ratzeburger Allee 160, D-23538 Lübeck, Germany. E-mail: {vig, dorr, martinetz, barth}@inb.uni-luebeck.de

M. Dorr is now affiliated with the Schepens Eye Research Institute, Dept. of Ophthalmology, Harvard Medical School, 20 Staniford Street, Boston, MA 02114, USA. E-mail: michael.dorr@schepens.harvard.edu

A. Related Work

Visual attention is a function of the continuous interaction between two different mechanisms: on the one hand, top-down or goal-driven, and bottom-up or stimulus-driven on the other [11]. The former is a voluntary, conscious form of attention control, where the task at hand and the observer's intentions determine the locations to be fixated. The latter refers to a set of processes by which eye movements are driven involuntarily by the "salience" of a stimulus, defined by its low-level visual features such as contrast, colour, and motion.

Due to the complexity of high-level cognitive functions, research has focused on bottom-up factors, investigating the relationship between eye movements and low-level image features at fixations[1]. For instance, it has been found that spatial contrast tends to be higher at the centre of fixation than at random control locations [12], [13], and there are regularities in the higher-order image statistics at fixation as well. Using bispectrum analysis, Krieger et al. [14] examined higher-order spatial correlations of the image intensities and found that intrinsically two-dimensional features such as curves and T- or L-junctions draw eye movements more often. Also, the eyes are often directed at regions with temporal change (motion). Therefore, bottom-up models of attention have been proposed that predict gaze based on visual attributes that are relevant in capturing stimulus-driven attention. These models centre on the concept of a "saliency map", which topographically encodes the salience of a location over the entire scene. Bottom-up saliency modelling is the focus of this paper.

Most bottom-up attention models that are biologically inspired (e.g. [15]–[19]) follow the Feature Integration Theory of Treisman [20] by first decomposing the visual input into separate low-

[1]In the process of seeing, our eyes alternate between fixations, when they are aimed at a fixed point, and rapid reorienting movements called saccades.

level feature maps, such as orientation, contrast and colour, on multiple scales. Normalized centre-surround difference maps are then computed for individual features and later combined by a weighting scheme to form a master saliency map. Attention is guided to peaks in this map in a winner-take-all fashion. An inhibition-of-return mechanism prevents attention from returning to an already attended location. This initial saliency model has undergone several modifications and extensions since Koch and Ullmann's [21] original description. It has been, for instance, extended to the temporal domain, and top-down priors have been incorporated to model phenomena beyond attention. For example, a low-dimensional signature vector, called the gist of the scene and acquired at multiple scales from basic visual features, has been used to perform scene classification [22].

Existing bottom-up saliency models, be they purely computational or biologically inspired, differ in their underlying *computational principles* they use to formally define the concept of saliency and motivate the model architecture (i.e. the choice of optimal features and major computational steps). A number of recent approaches turn to information theory to define "distinctiveness", i.e. conspicuity. The model of Bruce and Tsotsos [17] aims at maximizing Shannon's self-information to find the most informative locations in the image. Gao et al. [18], [19] introduced the concept of "discriminant saliency", which based on the definition of the target and null hypotheses (e.g. centre vs. surround, object class of interest vs. all other object classes) can act both as a bottom-up saliency predictor or top-down object detector. In this context, salient locations are those where the discrimination between target and non-target (in terms of some selected optimal features) can be made with minimum probability of error. Discrimination and classification confidence are defined with respect to a number of existing computational principles for perceptual organization (e.g. infomax or Barlow's inference by detection of suspicious coincidences).

The authors in [23] present a region-based bottom-up model for images, which uses roughly segmented regions as candidates for salient objects. The most salient segment is found through graphical model approximation. This stochastic model quantifies a number of intuitive observations, such as the likelihood of correspondence between visually similar image regions, and the assumption that the number of interesting objects in the scene is small.

Often, the problem of predicting eye movements on complex scenes is formulated in a Bayesian framework. This kind of approach provides an elegant way to, again, incorporate prior knowledge, e.g. about the statistics of visual attributes in specific scene types, or descriptions and layout of the scene. Itti and Baldi [24], for instance, proposed a Bayesian notion of surprise measured in "wows", by calculating the mismatch (or Kullback-Leibler divergence) between expectations of the observer, i.e. priors, and the perceived reality, i.e. posteriors. The model SUN [25] also uses a Bayesian framework to analyze fixations. Similarly to [17], novelty is defined as self-information of the visual features, but the feature statistics used to detect outliers are learned from previous examples, and are not based only on the current image or video. An alternative interpretation of Bayesian surprise, in the spatial rather than temporal domain, is proposed in [26].

While most approaches described above strive to address biological plausibility, the resulting models tend to be complex and have a large number of free parameters that need to be tuned by hand. Learning techniques are increasingly being employed as a practical solution to the parameter tuning problem (e.g. as above in [25]). Such models even allow to infer the model structure from the data, without the need to quantify several assumptions about perceptual processes. Still, the usefulness of learning in visual saliency modelling has been recognized only recently. Kienzle et al. were the first to derive saliency-based interest operators from human eye movement data using machine learning techniques that operated directly on the pixel intensities of static scenes [27] and Hollywood movies [28]. They showed that the learned discriminative features have a centre-surround pattern. Due to constraints imposed by the reduced ability of learning algorithms to operate in high-dimensional (pixel) spaces given a limited number of training samples, the algorithms in [27], [28] were limited to a single spatial scale. A data-driven approach is used in [29], too, where optimal parameters are learned (from fixation data on static scenes) for an attention model that is based on low-, mid- and high-level features calculated by several existing saliency methods. In [9], another supervised approach aims at learning to detect salient objects from manually labelled examples. Here, a set of novel features, such as multiscale contrast, centre-surround histogram, and colour spatial distribution, is combined through conditional random field learning.

While several models exist for saliency prediction on still images, only recently the number of studies that deal with scene sequences increased. Although some of the static approaches have been generalized to videos (e.g. [24], [28], [30]), these models often lack a unified framework for the static (spatial) and space-time saliency domains. Traditional ways to incorporate temporal information have often simply complemented the feature set with dynamic features, e.g. the optic flow information. In [9], for instance, the same set of novel features proposed for still images are defined on the motion field to capture spatio-temporal cues. The authors in [31] extend the bottom-up discriminant centre surround saliency model of [18] to background subtraction in highly dynamic scenes. Incorporating temporal information is also not straightforward in a learning context, where the task of eye movement prediction is further complicated by the increased number of (pixel-) dimensions.

Since most saliency models for videos are sensitive to dynamic content, camera motion and film-editing (e.g. jump cuts and gradual transitions) pose difficulties — even for the most advanced predictors — by causing false alarms in the salient features. This shortcoming is typically corrected with compensation of camera motion and shot boundary elimination. Shot boundary detection, too, can be tackled with an attentional paradigm. In [32], for example, saliency maps of nearby frames are compared for consistency and shot boundaries are detected when the similarity is below a given threshold.

B. Motivation

As seen above, computational saliency models range in complexity from empirical models with few parameters to more complex, multi-parameter ones. While ever more complex models seem to be needed to better predict gaze behaviour on realistic scenes, there are also a few counterexamples to the trend [27], [33].

This paper contributes to this latter line of research by exploring the potential of models that make as few assumptions as possible.

Once we have established a baseline, we can then investigate (and quantify) the potential gain from gradually increasing complexity. We propose to go back to the basics of signal processing to obtain efficient image representations, and, if required, utilize powerful learning algorithms on these representations to predict visual saliency in videos. We begin with the simple observation that many video regions, such as homogeneous areas, are highly redundant, and that local *changes*, i.e. intensity variations (along edges, corners, etc.) are informative. The degree of this signal redundancy can be mathematically described by the *intrinsic dimension* of an image or video region, and we here use this concept as a simple measure of saliency. In order to further tune the model parameters so as to predict bottom-up attention on complex scenes, we adopt data-driven machine learning techniques. However, given the high dimensionality of a pixel-based video representation, current learning algorithms would require very large amounts of data and thus have only limited practical applicability. Even with only a moderate amount of training data, i.e. human fixations on videos, we here overcome the curse of dimensionality through dimensionality reduction (specifically by spatial pooling of features). This allows us to incorporate more information by computing features on multiple spatio-temporal scales. Furthermore, the concept of intrinsic dimensionality naturally leads to a unified representation of spatial and temporal saliency, such that no fusion of separate static and dynamic maps is required. Similarly, the definition can be extended to multispectral sequences, so that it becomes no longer necessary to combine separate saliency maps from each colour channel. In order to test the performance of our model, we use a large data set of human fixations on a large collection of high-resolution videos. Since top-down processes strongly modulate gaze behaviour, we cannot expect any bottom-up model to fully account for the complex nature of attentional orienting. Nevertheless, we shall show that our simple assumptions already account reasonably well for eye movements during free-viewing of dynamic real-world scenes. Indeed, the proposed simple approach shows significant improvement over several state-of-the-art models of bottom-up saliency, which base their prediction on numerous assumptions on perceptual processes and incorporate several basic features. Through a systematic analysis, we set out to quantitatively evaluate the gain from more complex features by gradually extending a simple single-scale saliency map computed on the intensity videos to a multiscale and multispectral model. Our results support the (intuitive) assumption that a higher degree of variation in the visual signal leads to higher saliency.

The remainder of this paper is organized as follows. We start, in Section II, by describing the computational steps of a simple and efficient algorithm for bottom-up saliency. Then, in Section III, we demonstrate its performance in predicting human fixations on high-resolution natural videos. There, we shall test the validity of the approach for two distinct data labelling scenarios, discuss implementation issues, and present a systematic analysis of how the choice of free parameter values affects prediction performance. In Section III-E, we compare our results to those of four baseline models for bottom-up saliency. Then, in Section IV, we interpret the results and summarize the major findings. Finally, we provide some concluding remarks in Section V. A preliminary version of our algorithm with only a brief empirical analysis was published in [34].

II. Saliency Computation

An outline of our approach is schematically illustrated in Fig. 1. In this work, we learn the structural differences between salient and non-salient video locations on simple video representations that characterize different types of spatio-temporal intensity changes. Given a collection of image sequences and a large set of recorded eye movements on them, we label areas in the videos as either salient or non-salient. For each video, we compute low-level feature maps that encode the intrinsic dimensionality of video regions. Such maps are computed on several spatio-temporal levels of multiresolution image pyramids. In a neighbourhood around each location (be it salient or not), we extract the *feature energy* from these maps: the root-mean-square of the pixels in the spatio-temporal neighbourhood. Feature energy (a single scalar) is computed on each pyramid level; thus, each location is described by a low-dimensional vector whose components are the energy values on different scales. Such feature energy vectors are finally fed into a classifier, which learns a mapping between feature energy vectors and the saliency level of a certain location.

Before we describe the above steps in greater detail, we first recall the definition of the intrinsic dimension and review one method (based on the geometrical invariants of the structure tensor) used here for the estimation of the intrinsic dimension of both grayscale and multispectral image sequences.

A. Intrinsic Dimension

The *intrinsic dimension* (iD) [35] quantifies the information content of a signal. It describes the number of degrees of freedom needed to locally represent the observed signal. Thus, for a video, static and homogeneous locations are intrinsically zero dimensional ($i0D$), stationary edges and uniform regions that change in time have an intrinsic dimension of one ($i1D$), stationary corners and edges that change in time are $i2D$, while transient corners and non-uniform motion are intrinsically three-dimensional ($i3D$). The concept of intrinsic dimension is particularly relevant for image and video coding, because in natural scenes regions with high intrinsic dimension are less frequent than regions with low intrinsic dimension [36]. Moreover, an image or video can be fully reconstructed from only those regions where the iD is greater than one, i.e. $i0D$ and $i1D$ regions are redundant [37], [38].

Let a grayscale video be represented by the function $f : \mathbb{R}^3 \to \mathbb{R}$. 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

$$\frac{\partial f}{\partial \mathbf{v}} = 0 \quad \text{for all } \mathbf{v} \in E \,, \tag{1}$$

where the intrinsic dimension of Ω is $3 - \dim(E)$.

B. Invariants of the Structure Tensor

The subspace E can be estimated as the subspace spanned by the set of unity vectors that minimize the energy functional

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

where the *structure tensor* $\mathbf{J}$ [39] is given by

$$\mathbf{J} = \int_\Omega \nabla f \otimes \nabla f \; \mathrm{d}\Omega = \int_\Omega \begin{bmatrix} f_x^2 & f_x f_y & f_x f_t \\ f_x f_y & f_y^2 & f_y f_t \\ f_x f_t & f_y f_t & f_t^2 \end{bmatrix} \mathrm{d}\Omega \,. \tag{3}$$

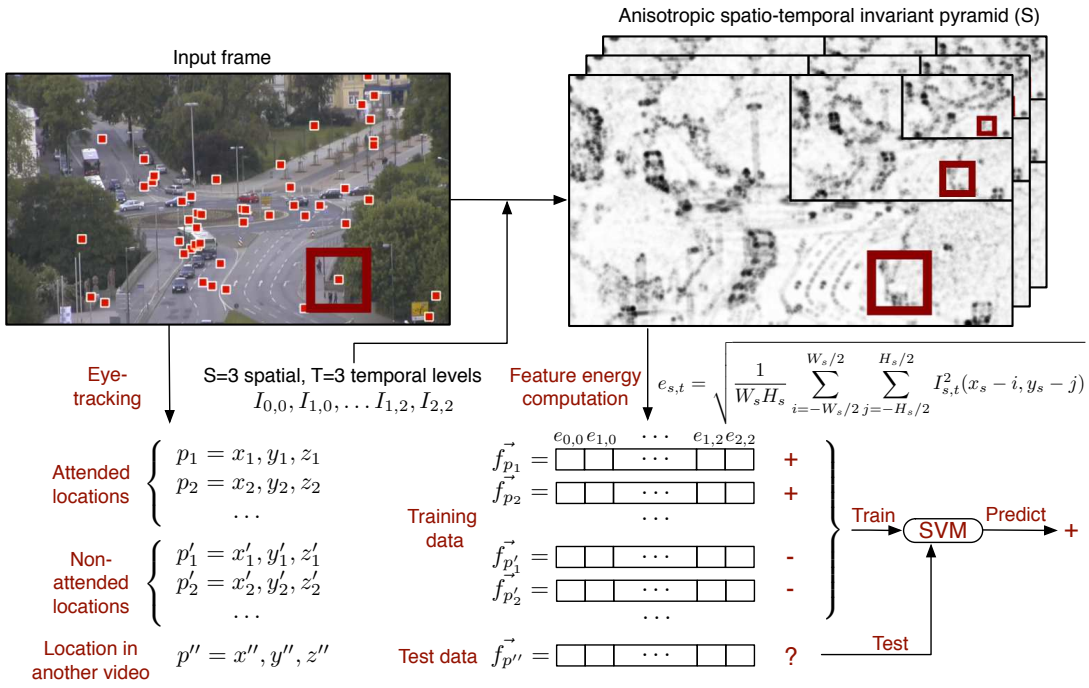


Fig. 1. Flow diagram summarizing our approach. Using eye-tracking data (fixations are denoted by small filled squares in the movie frame on the left), we label video regions as attended or non-attended. Image features — the geometrical invariants — are extracted on multiple scales of an anisotropic spatio-temporal pyramid. For a neighbourhood (large unfilled square shown schematically) around each location, the average feature energy is computed on each scale of the spatio-temporal pyramid. An SVM is trained on the obtained energy vectors and is then used to predict whether test locations of a new video will be attended or not.

In the above formula, $\otimes$ denotes the tensor product, the integral over Ω can be implemented a spatio-temporal Gaussian smoothing function, and f_x, f_y, and f_t stand for the first-order partial derivatives. E is the eigenspace associated with the smallest eigenvalue of $\mathbf{J}$, and the intrinsic dimension of f corresponds to the rank of $\mathbf{J}$. To avoid the computationally costly eigenvalue analysis, the intrinsic dimension can, alternatively, be obtained from $\mathbf{J}$'s symmetric invariants H, S, and K [40]:

$$\begin{array}{lll} H & = 1/3 \text{ trace}(\mathbf{J}) & = \lambda_1 + \lambda_2 + \lambda_3 \\ S & = M_{11} + M_{22} + M_{33} & = \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_1\lambda_3 \\ K & = |\mathbf{J}| & = \lambda_1\lambda_2\lambda_3 \end{array} , \quad (4)$$

where λ_i are eigenvalues and M_{ij} are minors of $\mathbf{J}$. If $K \neq 0$, the intrinsic dimension is 3 ($i3D$); if $S \neq 0$ it is at least $i2D$; and if $H \neq 0$ it is at least $i1D$. Example stillshots of the invariants of a natural scene are rendered in Fig. 2. The video itself and additional demos of the spatio-temporal invariants are available online at `http://www.inb.uni-luebeck.de/tools-demos/saliency`.

C. Multispectral Invariants

The concept of intrinsic dimension has been also extended to multispectral signals [41]. Given a multispectral image sequence $\mathbf{f}$ with q colour channels ($\mathbf{f} : \mathbb{R}^3 \to \mathbb{R}^q$), we choose an appropriate scalar product for $\mathbf{y} = (y_1, \ldots, y_q)$ and $\mathbf{z} = (z_1, \ldots, z_q)$ such that $\mathbf{y} \cdot \mathbf{z} = \sum_{k=1}^{q} a_k y_k z_k$. The positive weights a_k are meant to emphasize certain colour channels. The multispectral structure tensor can now be written as

$$\mathbf{J} = \int_\Omega \begin{bmatrix} \|\mathbf{f_x}\|^2 & \mathbf{f_x} \cdot \mathbf{f_y} & \mathbf{f_x} \cdot \mathbf{f_t} \\ \mathbf{f_x} \cdot \mathbf{f_y} & \|\mathbf{f_y}\|^2 & \mathbf{f_y} \cdot \mathbf{f_t} \\ \mathbf{f_x} \cdot \mathbf{f_t} & \mathbf{f_y} \cdot \mathbf{f_t} & \|\mathbf{f_t}\|^2 \end{bmatrix} \mathrm{d}\Omega \, . \quad (5)$$

Note that the above formulation does not assume any particular colour space. Videos are often represented in the $Y'C_bC_r$ colour space (instead of RGB, for instance) because the luma (Y') and the two chroma (C_b, C_r) channels are less correlated and the chroma channels are subsampled to take advantage of the lower colour sensitivity of the human visual system. However, when using $Y'C_bC_r$, the dynamic range of the luma channel is much greater than that of the chroma channels, so that the contribution of colour to $\mathbf{J}_{\mathbf{Y'C_bC_r}}$ is small. To compensate for this, we compute the standard deviation of each channel and use their inverse for the weights $a_{Y'}, a_{C_b}$, and a_{C_r}.

D. Multiscale Feature Extraction

The scale on which the intrinsic dimension is estimated depends on the bandwidths of the Gaussian smoothing function Ω and of the derivative operators. Therefore, the above geometrical invariants are computed on each scale of an anisotropic spatio-temporal multiresolution pyramid. As opposed to an isotropic pyramid, where spatial and temporal frequencies vary together, here each level of a spatial pyramid is decomposed further into its temporal bands. The resulting finer partition of the spectrum allows for the consideration of a higher number of subbands that encompass e.g. high spatial and low temporal frequencies. In principle, the anisotropic decomposition could also be applied to the spatial smoothing (i.e. separately on the horizontal and vertical spatial frequencies); however, this comes at considerable computational cost and is therefore avoided here.

E. Dimensionality Reduction

The saliency of a video location is strongly influenced by its spatio-temporal context. Centre-surround models exploit this

Fig. 2. Stillshot from a video (top left quadrant) and the corresponding geometrical invariants. For invariant K (bottom right quadrant), non-white locations change in all three spatio-temporal directions, whereas for S (bottom left), the video signal changes in at least two directions. Additionally, invariant H (top right) also responds to stationary edges and uniform regions that change in time (i.e. one dimensional changes). The (small) response even of K at the corners of the windows is due to small camera vibrations and noise. For the invariants, the brightness has been thresholded and inverted for better legibility.

property when they define saliency as the ability of some features to best discriminate between image structure in a centre and a surround window. In a data-driven approach, where fixational data is utilized to tune the model parameters, one also has to compensate for possible inaccuracies in both the eye tracking and the biological system. The size of the spatio-temporal neighbourhood that needs to be considered is still a matter of debate in the human vision community. While some studies use windows of the size of the high-resolution centre of the retina, the fovea (2–3 degrees), one can also optimize it with respect to the available eye movement data. Learning in the pixel space determined by the number of pixels of the neighbourhood is often problematic as the feature space dimensionality of a reasonable sized image patch, e.g. 64 by 64 pixels (2.5×2.5 deg) grows rapidly (more than 4000 dimensions). In such a scenario, given a limited number of training data, the effects of the "curse of dimensionality" seriously degrade classification performance. Because of these constraints, the learning algorithm in [27], for instance, was restricted to a single spatial scale.

In order to tackle the above problem and use information from multiple scales, we perform a *spatial pooling*. We reduce pixel information in a window around the location to a single scalar, by taking the root-mean-square of the feature values (i.e. geometrical invariants) in the window. Through pooling, an invariant representation of the local neighbourhood emerges. This allows us to compute the *feature energy* on every scale of the multiresolution pyramids, as the dimensionality remains low. Here, we use a spatial neighbourhood only, as the uncertainty induced by measurement errors and saccade imprecision is higher in the spatial domain than in the temporal one.

More formally, for a movie location $p = (x, y, z)$ (with spatial coordinates x and y, and frame number z), we compute a vector $\mathbf{f_p} = (e_{0,0}, e_{0,1}, \cdots, e_{S-1,T-1})$ consisting of the feature energies extracted from each scale of an anisotropic pyramid with S spatial and T temporal levels. The feature energy of a window (centred around the location p) computed on the s-th spatial and t-th temporal pyramid level is defined as

$$e_{s,t} = \sqrt{\frac{1}{W_s H_s} \sum_{i=-W_s/2}^{W_s/2} \sum_{j=-H_s/2}^{H_s/2} I_{s,t}^2(x_s - i, y_s - j)}\ , \quad (6)$$

where $I_{s,t}$ represents the s-th spatial and t-th temporal level of one of the invariant pyramids, H, S, and K, computed beforehand for every pixel. W_s and H_s stand for the (subsampled) spatial width and height of the neighbourhood on the s-th spatial scale (independent of the temporal scale). W_s and H_s are decreased by a factor of two per level, so that the effective window size is the same on all scales. The spatial coordinates of the location are also subsampled on the spatial scale s: $(x_s, y_s) = (x/2^s, y/2^s)$. In time, one frame of a lower pyramid level corresponds to several frames on the original level, so that we implicitly integrate over time as well. Given a learning scenario, the optimal window size can be inferred from the eye movement data by systematically evaluating, in terms of performance in predicting fixations, a range of different neighbourhood sizes.

F. Learning

Given a collection of videos together with a set of salient and non-salient locations on these videos, the task of predicting interesting locations can be naturally viewed as a binary decision problem, to which efficient methods from machine learning can be applied.

Thus, the task of learning to distinguish salient locations consists in finding a confidence value quantifying the patch's level of interestingness. Formally, we look for a function $g : \mathbb{R}^{S \times T} \to \mathbb{R}$ that returns such a confidence value for a new movie location p, based on its energy vector $\mathbf{f_p}$. The training data comprises the feature energy vectors of previously seen locations and associated class labels (salient or not), $(\mathbf{f_{p_i}}, l_i) \in \mathbb{R}^{S \times T} \times \{-1, 1\}$.

The data is partitioned "movie-wise" into a training and a test set: gaze data of all viewers on one movie are retained for testing, while the fixations on the remaining movies are used for the training. For the classification we use a standard soft margin Support Vector Machine with Gaussian kernels. Prior to training, we linearly scale each attribute (i.e. the feature energy on a particular spatio-temporal scale) to $[-1, 1]$. Optimal model parameters are found with cross-validation on the training sequence. To measure the quality of prediction, we perform an ROC analysis using the collected human gaze data as ground truth. Based on the resulting ROC curve, a single scalar, called the ROC score (and also known as the Area Under the Curve AUC), will provide an estimate of the prediction quality.

To quantify the benefits of incorporating information from multiple scales, we compare the model with simpler variants of the above classifier that operate on *single scales* only. For this, we evaluate the performance of one-dimensional maximum-likelihood classifiers when the feature energies from individual pyramid levels are treated as inputs to the decision algorithm. Results for the "most predictive" scale are then compared to the performance of the (learned) multiscale model.

III. Experimental Evaluation

Here, we test the quality of the structure tensor-based predictors on a large set of eye movement data and compare their predictive power with that of four state-of-the-art models of bottom-up saliency.

Fig. 3. Stillshots from four movies: beach, breite_strasse, ducks_boat, holsten_gate.

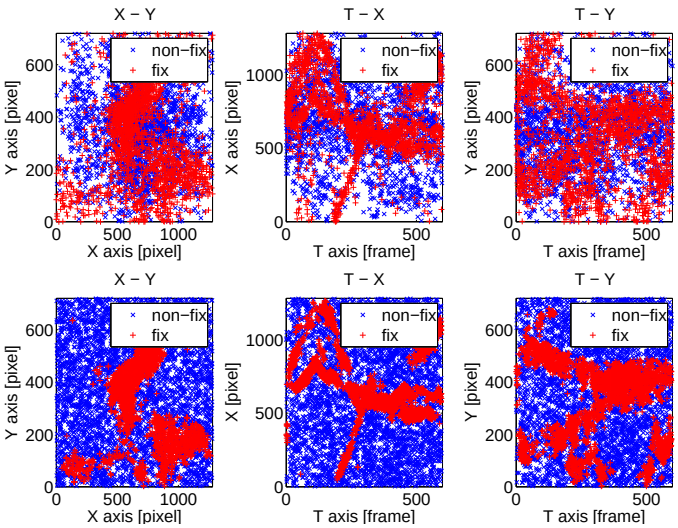


Fig. 4. Salient (red plus) and non-salient (blue cross) locations on a movie. These locations are shown on the 2D projections (xy, tx, and ty) of the 3D spatio-temporal volume of the video. Upper row: "bias-free" labelling with saccade landing points in the salient class, and fixations on other movies in the non-salient class. Lower row: "default" labelling — salient and non-salient locations are chosen from the maxima and minima of the empirical saliency measure. Note the difference in overlap between the two classes under the two labelling schemes.

A. Videos and Eye Movement Data

Our experiments examined the performance of the proposed approach for eye movement prediction on the public data set of [42]. This set consists of 18 high-resolution movie clips (1280 by 720 pixels, 29.97 fps, about 20 s duration each, recorded in the $Y'C_bC_r$ format) of natural outdoor scenes, and the gaze data of 54 human subjects freely viewing these videos. Stillshots from four videos are shown in Fig. 3. For more details about the recording setup we refer to [42]. From the recorded gaze data, about 40,000 saccades were extracted using a dual-threshold velocity-based procedure [43].

B. Data Set Labelling

The learning algorithm takes as input a set of positive, salient examples and a set of negative, non-salient ones. Whereas the set of fixations, more precisely saccade landing points, appears as a straightforward choice for the positive class, obtaining negative examples is non-trivial. An intuitive and commonly used approach is to arbitrarily pick locations from a uniform distribution either from the entire scene or (better) from areas that were not fixated, i.e. where spatio-temporal distance to the nearest fixation is large enough. However, several recent studies have pointed out that such approaches do not account for a common problem inherent in most eye movement data sets: the tendency of viewers to fixate preferably in the centre of the display [13], [44]. To remove possible artefacts due to the *centrally biased* distribution of gaze positions, it has been suggested that the non-salient locations of a video should be taken from real scanpaths on *different* movies. That way, an identical spatio-temporal distribution of the positive and negative examples over the set of all movies is obtained, but such artefact minimization also comes at a price. The above procedure of picking the negative examples may lead to overlap between the two classes and, hence, to an underestimation of the real model performance.

Existing approaches typically report results for only one of the aforementioned methods, so that it is not clear how sensitive the models are to labelling conditions, and whether or not the different conditions lead to significant deviation in performance. To investigate this and provide a fair comparison of the different models that might otherwise benefit from (labelling) biases, we consider both of the above labelling procedures: the "bias-free", where we account for the central fixation bias and allow for overlap, and the "default" one, which minimizes the overlap. Loosely speaking, the "bias-free" scheme samples negative training data from different movies, whereas the "default" scheme samples from different spatial locations.

In the first case, the full set of saccade landing points is used to label the salient locations (about 40,000 over all movies and subjects). For the negative class, the non-salient locations of a movie are chosen using randomly selected scanpaths from different movies (see upper row of Fig. 4). Because of latencies of the oculomotor system, the time of the gaze response to a specific salient event does not necessarily coincide with the time of the event. Hence, existing approaches usually introduce a temporal offset (between 150-250 ms) based on well-established results on reaction time to synthetic stimuli. However, we have previously shown that the typical reaction time is stimulus dependent, and in natural scenes this average lag is near zero (i.e. no offset needs to be considered) due to the highly predictive nature of salient real-world events [45].

As argued before, such a "bias-free" labelling procedure introduces overlap between the salient and non-salient classes, i.e. the data set is contaminated with wrongly labelled samples (outliers) that deteriorate the model performance. In an attempt to avoid such overlap, in the "default" labelling scheme, we rank video regions according to an *"empirical" saliency measure*, which is derived from the recorded eye movement data. Such maps are defined as the density of the gaze points averaged over all viewers and therefore constitute an upper limit of prediction, i.e. an inter-subject agreement. We compute a probability map for each video, by superposing spatio-temporal Gaussians placed at each gaze location of all subjects. Samples of the salient and non-salient classes are picked from regions with the highest (for the positive class) and lowest (for the negative class) density of fixations. In our analysis, the Gaussian filter had a spatial support of 2.4 degrees of visual angle, a temporal one of 0.17 s, with standard deviations of 0.6 degrees (spatial) and 600 ms (temporal).

An equal number (40,000) of salient (non-salient) locations are then chosen randomly from locations where the empirical saliency exceeds (is below) a given global threshold (see lower row of Fig. 4). Threshold values were set at the upper ten percent (for salient) and lower one percent (for non-salient locations) of the maximum empirical saliency estimated over all movies. These values were chosen so as to obtain an equal number of data points in the two (salient and non-salient) classes.

C. Implementation

Here, we provide a more detailed discussion of how implementation considerations were integrated in our analysis.

To extract the proposed salient features (the geometrical invariants) on different spatio-temporal scales, we constructed an anisotropic pyramid with $S = 5$ spatial and $T = 5$ temporal levels, as described in Section II-D. This rather high number of pyramid levels (a free parameter) was chosen so as to ensure that frequency components that are potentially relevant for visual saliency are represented. For the structure tensor $\mathbf{J}$, partial derivatives in Eq. 3 were calculated by first smoothing the input with spatio-temporal 5-tap binomial kernels $(1, 4, 6, 4, 1)/16$ and then applying $[-1, 0, 1]$ kernels to compute the differences of neighbouring pixel values. For the smoothing of the products of derivatives (with Ω), we chose the same spatio-temporal 5-tap Gaussian.

Besides being symmetric, the above filter kernels are non-causal, so that the temporal filtering requires video frames with future time stamps. As a consequence, depending on the number of temporal scales, a certain number of the initial and final output frames of the invariants are distorted. To avoid such temporal border effects, we only considered fixations from (and restricted the analysis to) valid frames. For a temporal pyramid with $T = 5$ levels, this meant discarding quite a notable number of frames: the first and last 3.2 s (96 frames) were not considered for further analysis. Since the invariants H, S, and K comprise of products of one, two, and three eigenvalues, respectively, their dynamic range is not identical. For a fair comparison of the three, we therefore mapped them to the same dynamic range: they were raised to the power of six, three, and two, respectively.

To increase computational efficiency in the subsequent steps, the invariants were stored to disk using lossless compression. We normalized output invariant videos to pixel intensity values between $[0, 255]$ by taking the eighth root and linearly scaling the maximum over all levels to 255.

Once these features were extracted on multiple scales, we computed the feature energy in windows of varying size at each salient and non-salient location (about 25,000 per class over all movies, after discarding invalid invariant frames). We cropped the window at the boundaries if it was too large.

Finally, a classifier was trained with feature energy vectors on all but one video from the movie set and testing was performed on the withheld movie. The optimal parameters of the kernel Support Vector Machine (i.e. the width γ of the Gaussian and the penalty term C) were found by 8-fold cross-validation on the training sequence. Given a low number of videos (18 in total), and since eye movement predictability varies quite considerably between different video clips, the whole procedure (including the training and search for optimal parameters) was repeated 18 times so that each movie served as test data once.

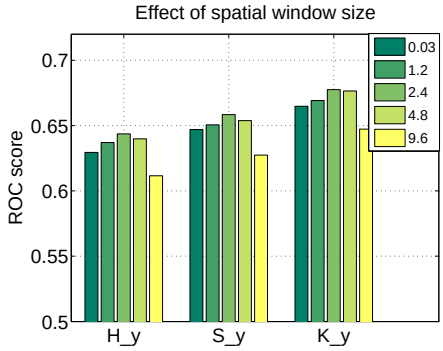
To estimate the performance gain from incorporating information from multiple spatio-temporal scales, the predictability of the single scales was also tested. For this, an ROC analysis was performed (without further SVM prediction) on the energies from single pyramid levels. Here, multiscale results are compared with the outcome of the single "best" scale over all movies (in terms of ROC analysis), i.e. the frequency component that is most relevant for attentional selection. In case of multiscale analysis, the delivered decision values on the test movie are determined with respect to the training data, that is, the energy vectors from the remaining 17 videos. For single scales, however, a separate ROC analysis on each single movie would not take into account the overall distribution of feature energies in the two classes, and thus overestimate performance. Therefore, for single scales, instead of 18 ROC tests for the individual movies, we perform a single ROC analysis on the *entire* set of salient and non-salient locations from *all* 18 videos. This assures that during decision making the approximated true distribution of the fixated and non-fixated energies is used.

D. Quantitative Analysis

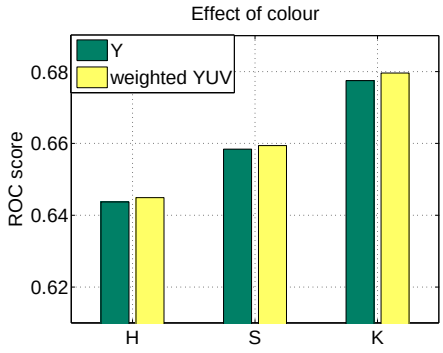
In this section, we systematically investigate how different feature types contribute to model performance. We vary three main variables: the window size considered in extracting the feature energy, the colour channels (luminance alone or multispectral representations) on which the geometrical invariants are extracted, and, finally, the number of pyramid scales considered (single-scale vs. multiscale approach). The following analysis was performed for all three geometrical invariants. Since the qualitative results for the two types of data set labelling were identical, in this section, we only consider one: the "bias-free" labelling.

We started with the simplest scenario, considering salient features that are extracted on single spatio-temporal scales of the grayscale videos (i.e. no multiscale and multispectral analysis yet). Here, we report results for the pyramid level that gave best predictability, in terms of a single ROC analysis over the entire set of fixated and non-fixated locations from all 18 movies. To quantify the gain of the final spatial pooling (i.e. feature energy computation) on predictability, we varied the spatial window in size between a single pixel (i.e. no spatial pooling) to about 10 degrees of visual angle, with the exact window sizes used as follows: 0.03, 1.2, 2.4, 4.8, and 9.6 degrees. As seen in Fig. 5(a), the trend is consistent for all three invariants: predictability increases with the window size, peaking at around 2.5 degrees, after which it slowly decreases. A window of 4.8 degrees still yields prediction rates close to the maximum. This is in agreement with psychophysical studies that claim the size of the influencing spatio-temporal context has roughly the size of the fovea. Since the relative gain in predictive power from no window to one of 2.4 degrees is 11% for invariant H, and 8% for S and K, a rather large pooling is justified. Therefore, for further analysis we fix the window size to the optimal 2.4 degrees.

The qualitatively most relevant result, however, is that the prediction performance increases with the intrinsic dimension: invariants that extract features with higher intrinsic dimension are more predictive. Thus, invariant K with an ROC score of 0.68 is best, followed by S (AUC of 0.66), whereas the worst performing is H with an AUC of 0.64. Similar results that showed this ranking were published in [10] on a substantially different problem: there we were predicting gaze behaviour of new viewers on videos that



(a)



(b)

Fig. 5. (a) Eye movement predictability as a function of window size for the "bias-free" labelling. Range tested: $\{0.03, 1.2, 2.4, 4.8, 9.6\}$ degrees of visual angle. For all three invariants, highest ROC scores were found at 2.4 deg. (b) Predictability using the geometrical invariants of the structure tensor on the luminance channel (Y) and of the multispectral structure tensor (YUV) given an optimal window size of 2.4 deg. Performance does not increase much with the addition of the UV colour channels. In both (a) and (b), invariants that extract features with higher intrinsic dimensions (K) are more predictive than lower intrinsic dimensions (S and H).

have already been "seen" (i.e. learned on) by the classifier, as opposed to predicting eye movements on new videos. Due to the significant differences in problem formulation (e.g. training and test set division, data labelling, type of pyramids and number of scales used, etc.) results of the two scenarios cannot be compared directly.

Results for geometrical invariants computed on the luminance channel alone versus on multispectral representations (the weighted $Y'C_bC_r$ colour space) are shown in Fig. 5(b). Colour information has surprisingly little effect on saliency: it improves prediction performance, but only slightly.

Finally, we evaluate how much improvement can be achieved when including information from multiple scales. Thus, the single-dimensional ROC analysis is replaced by a kernel SVM that operates on 25-dimensional feature energy vectors computed on anisotropic invariant pyramids with $S = 5$ spatial and $T = 5$ temporal levels. As expected, results in Fig. 6 show some benefits of multiscale processing: prediction performance improved by 11% for invariant H, for S by 7%, while a slightly smaller increase of 4.5% is found for K.

E. Comparison to Existing Bottom-up Models

We compared the proposed generic method with four state-of-the-art models of bottom-up saliency for dynamic scenes: the

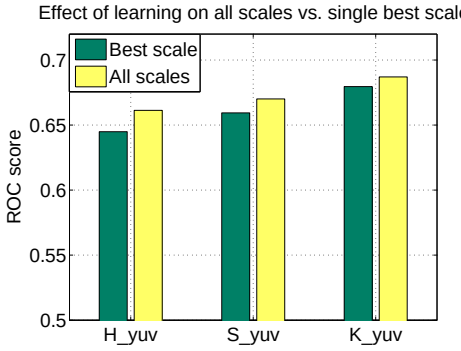


Fig. 6. Predictive power of single-scale (i.e. "best" scale in terms of ROC analysis) and multiscale approaches (window size = 2.4 degrees, multispectral structure tensor, "bias-free" labelling). Using information from multiple scales improves performance, but only slightly.

Bayesian "surprise" [24], SUNDAy [30], and the models of [15] and [46] (denoted by "Maxnorm" and "Fancy"). The last two are in fact implementations of the classical saliency map of Koch and Ullman [21] but which employ different fusing schemes of the individual saliency maps into a master map. Default model parameters were used to obtain saliency maps for the same video set. To discriminate between salient and non-salient movie locations, these maps were treated as maximum likelihood binary classifiers. By thresholding these maps, movie regions above the threshold were classified as salient. A systematic variation of the threshold – "movie-wise" – resulted in 18 ROC scores listed in Table I. As before, the labelling scheme used to obtain the results in Table I was the "bias-free". For comparison, the geometrical invariants were extracted from multiscale and multispectral representations (with feature energies computed in the optimal window of 2.4 degrees). The prediction performance of the various models was compared with a paired Wilcoxon signed rank test. Statistical significance was obtained for $K > H$ ($p = 0.034$) and $K > S$ ($p = 0.013$), but not for $S > H$ ($p = 0.395$). Also, results on the invariants proved to be significantly different from those of the four baseline models (except for $H >$ SUNDAy with $p = 0.07$). However, no statistical differences were found among the four state-of-the-art models.

Possible ROC scores range from 0.5, which indicates chance performance, to 1.0, which means perfect discrimination. Note, however, that different class labelling strategies narrow the effective range of ROC scores. On the one hand, the "bias-free" method that accounts for the central fixation bias may lead to erroneous labelling, which results in lower prediction rates. On the other hand, with no bias-correction ("default" labelling), the model benefits from the differences in the spatio-temporal location distributions, which amounts to a substantial jump in performance. To estimate the effective performance range related to the two different labelling strategies, we additionally considered two simple control measures: (1) the spatial distance of the salient/non-salient location to the video-centre as a (possible) lower bound to this range, and (2) the "empirical saliency" measure – a fixation density map – as a "perfect" predictor of eye movements and, as such, as an upper bound. Note that when existing scanpaths from other movies serve as non-fixated points, the salient and non-salient location distributions are identical, hence, the distance to centre performs roughly at chance level. However, the empirical saliency is obviously an optimal predictor

TABLE I
ROC SCORES OF VARIOUS BOTTOM-UP SALIENCY MODELS ON THE COLLECTION OF 18 OUTDOOR VIDEOS ("BIAS-FREE" LABELLING; NUMBERS IN BOLD INDICATE HIGHEST PREDICTION RATE). REGIONS WITH HIGHER INTRINSIC DIMENSION (ENCODED BY INVARIANT K) ARE SIGNIFICANTLY MORE PREDICTIVE FOR SALIENCY (PAIRED WILCOXON'S TEST).

Movie	H	S	K	Maxn	Fancy	Surp	SUN
beach	0.67	0.68	**0.71**	0.64	0.61	0.61	0.65
breite_strasse	0.71	**0.76**	**0.76**	0.73	0.70	0.70	0.70
bridge_1	**0.63**	0.61	0.59	0.53	0.48	0.52	0.50
bridge_2	0.57	0.53	0.53	0.59	0.61	**0.64**	0.60
bumblebee	0.57	0.54	**0.63**	0.53	0.55	0.54	0.56
doves	0.80	0.82	**0.83**	0.67	0.70	0.71	0.72
ducks_boat	0.58	0.64	**0.70**	**0.70**	0.63	0.65	0.63
ducks_children	0.73	**0.78**	**0.78**	0.48	0.59	0.56	0.70
golf	0.75	0.76	**0.77**	0.70	0.60	0.67	**0.77**
holsten_gate	0.62	0.62	**0.66**	0.61	0.53	0.51	0.61
koenigstrasse	**0.64**	0.62	0.60	0.57	0.53	0.60	0.62
puppies	0.68	0.73	0.75	0.68	**0.76**	0.71	0.65
roundabout	0.68	0.69	**0.70**	0.63	0.63	0.62	0.63
sea	0.84	**0.86**	**0.86**	0.82	0.77	0.83	0.84
st_petri_gate	0.56	0.58	**0.60**	0.52	0.56	0.56	0.51
st_petri_market	0.62	0.60	**0.63**	0.57	0.56	0.52	0.58
st_petri_mcdon.	0.51	0.48	0.50	0.51	**0.59**	0.51	0.57
street	0.74	0.76	**0.77**	0.71	0.68	0.58	0.68
Average	0.66	0.67	**0.69**	0.62	0.61	0.61	0.64

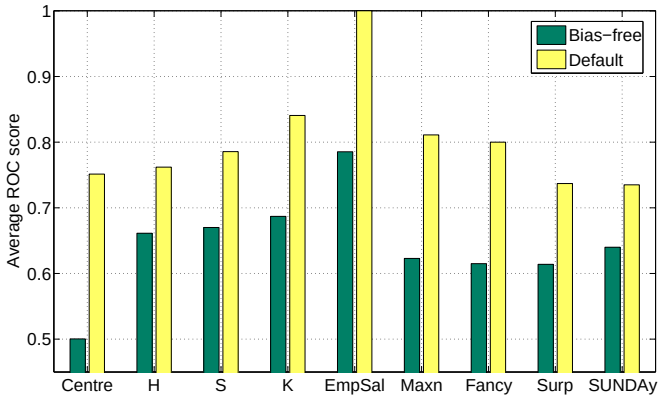


Fig. 7. Average ROC scores of the various models for the prediction of eye movements on naturalistic videos. The two data labelling scenarios (green – "bias-free" and yellow – "default") differ on whether or not viewing biases are accounted for, and whether all fixations or only the most salient areas are modelled. To estimate the effective performance range, two control measures were introduced: (1) Centre – distance to the video-centre as a lower bound and (2) EmpSal - the empirical saliency as the upper bound. The invariants (H, S, and K) were computed with the optimal parameters: a multispectral anisotropic pyramid with five spatial and five temporal levels, and feature energy was averaged in a window of 2.4 degrees. Performance is compared to that of four baseline models: Itti's Maxnorm (Maxn) and Fancy algorithm, Itti and Baldi's Surprise model (Surp) and SUNDAy.

(with an AUC of 1.0) when the locations of the two classes are picked by thresholding this map.

The performance of the various methods for the two labelling strategies is summarized as averages over all 18 test sets/movies in Fig. 7. With no bias-correction ("default" labelling), the distance to the centre alone achieves a mean ROC score of 0.75, which is in agreement with previously reported results [25], [29]. At the same time, in the case of "bias-free" labelling, an empirical saliency measure built on the fixation positions discriminates these same locations from non-salient ones with a mean AUC of only 0.79. The non-optimal performance is here due to noisy labelling and overlap in the two classes.

Despite its simplicity, our generic model based on the invariants of the structure tensor outperforms all four baseline models when accounting for the central fixation bias. Invariant K (average 0.69) comes closest to the upper bound marked by empirical saliency (0.79), but even the "weaker" invariants S and H still perform better than the baseline models; of those, SUNDAy achieves the highest average AUC (0.64).

Invariant K gives best prediction results (0.84) also for the second labelling procedure. Here, the two Itti models ("Maxnorm" and "Fancy", 0.81 and 0.80) perform better than SUNDAy and Surprise; the latter two surprisingly seem to be only as good as the "distance to centre" classifier.

IV. DISCUSSION

In this paper, we have derived a generic yet powerful model for bottom-up saliency from the simple assumption that the degree of local intensity variation is related to the informativeness of an image region. The concept of intrinsic dimensionality measures this degree and yields a basic description (or "alphabet") of how a multidimensional signal may change. We characterize typical video structures based on the geometrical invariants H, S, and K of the structure tensor, which correspond to the minimum intrinsic dimension of a movie region. Our model of bottom-up saliency combines such simple low-level visual features — the geometrical invariants extracted on multiple spatio-temporal scales — with machine learning to predict salient locations in natural dynamic scenes. We found that this simple approach proves successful in explaining human fixation data on a diverse collection of real-world videos. All three geometrical invariants were found to have good prediction capability. More importantly, however, our results provide strong evidence that the human visual system preferentially allocates its processing resources to more informative image regions; invariants that extract features with higher intrinsic dimension yield a sparser representation and are more predictive for eye movements. Conversely, movie regions with lower intrinsic dimension, i.e. redundant locations in case of $i0D$ and $i1D$, are less often fixated. Taken together, this provides indirect evidence for the efficient coding strategy of the brain [47], and indeed $i2D$ operators emerge as non-linear filters when sparse overcomplete bases are learnt [48]. Our structure tensor-based approach is closely related to the space-time interest points of Laptev [49]. In their approach, the spatio-temporal structure tensor is employed to detect local 3D corners in videos, which are highly useful in providing a compact representation of a movie. Such space-time interest points are popular in computer vision, e.g. for learning and recognizing human activities in videos.

Despite being based on simple, low-dimensional representations (1 to max. 25 scalars), the proposed model shows significant improvement over the four selected baseline models of bottom-up saliency. This finding becomes even more striking given the fact that such cognitive models rest on several assumptions, employ a high number of hand-tuned parameters, and involve complex

computations. However, the straightforward hypothesis that during visual processing signals with lower intrinsic dimension are suppressed renders also our model biologically plausible as well. Indeed, previous work has shown that this simple hypothesis can already explain the occurrence of lateral inhibition ($i0D$ signals are suppressed), end-stopping ($i1D$ signals are suppressed) [35], and motion selectivity [50].

Existing approaches are typically tuned towards optimal performance for specific tasks: while the SUNDAy model yields smooth, continuous saliency maps that are more adequate for the prediction of real fixations, the Itti models (especially the normalization scheme "Fancy") produce sparser maps with few peaks that rather account for the most salient scene locations only. To test how well our simple approach can generalize to both tasks, we defined two data-labelling scenarios: one that aims to model all human fixations, but picks non-salient locations so as to account for viewing biases, too; and a second, where salient and not salient locations are chosen from the most and least salient video regions without viewing bias correction. To our surprise, we find that while existing models typically excel in only one scenario, our approach, more specifically invariant K, is generic enough to provide optimal prediction for both problems.

We also have shown that although different labelling schemes allow the comparison of the relative performance of the different models, they also narrow down the effective performance range. Knowledge of the upper and lower bounds of the model performance is essential as it allows the assessment of the true performance gain and the estimation of the closeness to the optimal model behaviour achievable for a given problem formulation.

In order to understand the potential gain from more complex (but biologically motivated) features, that is from additional information (be it for instance multiscale or multispectral), we performed a comprehensive analysis by gradually extending our simplest saliency map, the geometrical invariants computed on a single scale of the intensity videos. With the integration of more features, the introduction of additional free parameters becomes inevitable, but their values are here fine-tuned in a supervised learning scenario.

Our first extension, the spatial pooling through feature energy computation, allowed us to consider movie sub-volumes (i.e. a salient context) of arbitrary size around the fixation. Thus, we could overcome the limitations of learning algorithms operating in high-dimensional (pixel) spaces. This is, however, only one simple way of decreasing dimensionality, and we are aware that by such a notable reduction also an information loss is introduced. Still, this step enabled the computation of visual features on multiple spatio-temporal scales, while only modestly increasing the dimensionality again.

A key issue in the design of bottom-up saliency maps is how to combine separate feature maps coming from different modalities to create a unique master map. A main advantage of the concept of intrinsic dimensionality is that it leads to a unified representation of spatial and temporal saliency and, moreover, that it can be readily extended to multispectral sequences. However, we found no strong difference between the invariants on luminance and those on a multispectral representation. This could be partly due to the fact that colour channels are highly correlated with each other, so that only redundant information is added with colour. Also, other colour spaces, such as the perceptually uniform CIELAB space, as well as the approximately equidistant HSV space, may better capture the true role of colour in attentional guidance.

Overall, we found that including more information and fine-tuning the model parameters through learning algorithms increased the predictability, but the gain was less than intuitively expected. Learning appears to partially compensate for the lower quality of an image or video representation, when quality is measured in terms of how compact a representation is. Note, however, that our eye movement prediction results are better than those of the reference models even without multiscale learning.

Obviously, as with any purely bottom-up model of visual saliency, the present approach cannot fully account for the complex nature of human fixation patterns. Nevertheless, such models may predict top-down behaviour reasonably well when the high-level task is implicit or unknown [51]. Indeed, our proposed model further improves upon previous approaches and successfully predicts human eye movements during free-viewing of dynamic real-world scenes. Note that incorporating other known properties of active vision, such as scanpath statistics, temporal correlations of scanpaths, and preference for the centre, could lead to even better performance.

V. Conclusion

In summary, we have demonstrated how standard supervised learning techniques can fine-tune the free parameters of a simple image processing-based model of bottom-up saliency to account for eye movements in natural dynamic scenes. Grounded in the intuitive assumption that the visual signal must change in order to attract attention, we proposed a generic model and tested its predictive power on a large set of eye movements in two distinct data labelling scenarios. Despite its conceptual simplicity, our model outperforms state-of-the-art baseline models.

Acknowledgment

We would like to thank Karl Gegenfurtner; data was collected in his lab at the Dept. of Psychology of Giessen University and is available at [42] and `www.inb.uni-luebeck.de/tools-demos/gaze`. Our research has received funding from the European Commission within the project GazeCom (contract no. IST-C-033816, see `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] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of interest point detectors," *Int. J. Comput. Vision*, vol. 37, pp. 151–172, June 2000.

[2] S. J. Dickinson, H. I. Christensen, J. K. Tsotsos, and G. Olofsson, "Active object recognition integrating attention and viewpoint control," *Computer Vision and Image Understanding*, vol. 63, no. 67–3, pp. 239–260, 1997.

[3] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004, pp. 37–44.

[4] W. S. Geisler and J. S. Perry, "A real-time foveated multiresolution system for low-bandwidth video communication," in *Human Vision and Electronic Imaging: SPIE Proceedings*, B. E. Rogowitz and T. N. Pappas, Eds., 1998, pp. 294–305.

[5] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. on Image Processing*, vol. 13, no. 10, pp. 1304–1318, Oct 2004.

[6] A. Ninassi, O. L. Meur, P. L. Callet, and D. Barba, "Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric," in *Proceedings of the International Conference on Image Processing (ICIP)*, 2007, pp. 169–172.

[7] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen, "Gaze-based interaction for semi-automatic photo cropping," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, ser. CHI '06. New York, NY, USA: ACM, 2006, pp. 771–780.

[8] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 411–426, 2007.

[9] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2011.

[10] E. Vig, M. Dorr, and E. Barth, "Efficient visual coding and the predictability of eye movements on natural movies," *Spatial Vision*, vol. 22, no. 5, pp. 397–408, 2009.

[11] A. L. Yarbus, *Eye Movements and Vision*. New York: Plenum Press, 1967.

[12] P. Reinagel and A. M. Zador, "Natural scene statistics at the centre of gaze," *Network: Comput Neural Syst*, vol. 10, pp. 341–350, 1999.

[13] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time," *Vision Research*, vol. 45, pp. 643–659, 2005.

[14] G. Krieger, I. Rentschler, G. Hauske, K. Schill, and C. Zetzsche, "Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics," *Spatial Vision*, vol. 13, no. 2,3, pp. 201–214, 2000.

[15] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[16] O. L. Meur, P. L. Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 802–817, 2006.

[17] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006, pp. 155–162.

[18] D. Gao and N. Vasconcelos, "Decision-theoretic saliency: Computational principles, biological plausibility, and implications for neurophysiology and psychophysics," *Neural Computation*, vol. 21, no. 1, pp. 239–271, 2009.

[19] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 989–1005, 2009.

[20] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[21] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry." *Human neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[22] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 300–312, 2007.

[23] T. Avraham and M. Lindenbaum, "Esaliency (extended saliency): Meaningful attention using stochastic image modeling," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 693–708, 2010.

[24] L. Itti and P. Baldi, "Bayesian Surprise Attracts Human Attention," *Vision Research*, vol. 49, no. 10, pp. 1295–1306, May 2009.

[25] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, pp. 1–20, 12 2008.

[26] I. Gkioulekas, G. Evangelopoulos, and P. Maragos, "Spatial Bayesian surprise for image saliency and quality assessment," in *Proc. IEEE Int'l Conf. on Image Processing (ICIP-10)*, Hong Kong, September 2010.

[27] W. Kienzle, F. A. Wichmann, B. Schölkopf, and M. O. Franz, "A Nonparametric Approach to Bottom-Up Visual Saliency," in *Advances in Neural Information Processing Systems*. Cambridge, Mass. USA: MIT Press, 2007, pp. 689–696.

[28] 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)*. Berlin, Germany: Springer Verlag, 2007, pp. 405–414.

[29] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to Predict Where Humans Look," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2009.

[30] L. Zhang, M. H. Tong, and G. W. Cottrell, "SUNDAy: Saliency Using Natural Statistics for Dynamic Analysis of Scenes," in *Proceedings of the 31st Annual Cognitive Science Conference, Amsterdam, Netherlands*, 2009.

[31] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 171–177, 2010.

[32] G. Boccignone, A. Chianese, V. Moscato, and A. Picariello, "Foveated shot detection for video segmentation," *IEEE Trans. Circuits Syst. Video Technol*, vol. 15, pp. 365–377, 2005.

[33] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. on Image Processing*, vol. 19, no. 1, pp. 185–198, 2010.

[34] E. Vig, M. Dorr, T. Martinetz, and E. Barth, "A learned saliency predictor for dynamic natural scenes," in *ICANN 2010, Part III*, ser. Lecture Notes in Computer Science, K. Diamantaras, W. Duch, and L. S. Iliadis, Eds., vol. 6354. Thessaloniki, Greece: Springer, 2010, pp. 52–61.

[35] C. Zetzsche and E. Barth, "Fundamental limits of linear filters in the visual processing of two-dimensional signals," *Vision Research*, vol. 30, pp. 1111–1117, 1990.

[36] C. Zetzsche, E. Barth, and B. Wegmann, "The importance of intrinsically two-dimensional image features in biological vision and picture coding," in *Digital Images and Human Vision*, A. B. Watson, Ed. MIT Press, Oct. 1993, pp. 109–38.

[37] E. Barth, T. Caelli, and C. Zetzsche, "Image encoding, labeling, and reconstruction from differential geometry," *CVGIP: Graphical Models and Image Processing*, vol. 55, no. 6, pp. 428–446, November 1993.

[38] C. Mota and E. Barth, "On the uniqueness of curvature features," in *Dynamische Perzeption*, ser. Proceedings in Artificial Intelligence, G. Baratoff and H. Neumann, Eds., vol. 9. Köln: Infix Verlag, 2000, pp. 175–178.

[39] B. Jähne, H. Haußecker, and P. Geißler, Eds., *Handbook of Computer Vision and Applications*. San Diego, USA: Academic Press, 1999.

[40] C. Mota, I. Stuke, and E. Barth, "Analytic solutions for multiple motions," in *Proc. IEEE Int. Conf. Image Processing*, vol. II. Thessaloniki, Greece: IEEE Signal Processing Soc., October 7-10, 2001, pp. 917–20.

[41] ——, "The intrinsic dimension of multispectral images," in *MICCAI Workshop on Biophotonics Imaging for Diagnostics and Treatment*, 2006, pp. 93–100.

[42] M. Dorr, T. Martinetz, K. Gegenfurtner, and E. Barth, "Variability of eye movements when viewing dynamic natural scenes," *Journal of Vision*, vol. 10, no. 10, pp. 1–17, 2010.

[43] M. Böhme, M. Dorr, C. Krause, T. Martinetz, and E. Barth, "Eye movement predictions on natural videos," *Neurocomputing*, vol. 69, no. 16–18, pp. 1996–2004, 2006.

[44] P.-H. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *Journal of Vision*, vol. 9, no. 7, pp. 1–16, 7 2009.

[45] E. Vig, M. Dorr, T. Martinetz, and E. Barth, "Eye movements show optimal average anticipation with natural dynamic scenes," *Cognitive Computation*, vol. 3, no. 1, pp. 79–88, 2011.

[46] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 161–169, Jan 2001.

[47] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[48] K. Labusch, E. Barth, and T. Martinetz, "Sparse Coding Neural Gas: Learning of Overcomplete Data Representations," *Neurocomputing*, vol. 72, no. 7-9, pp. 1547–1555, 2009.

[49] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2, pp. 107–123, 2005.

[50] E. Barth and A. B. Watson, "A geometric framework for nonlinear visual coding," *Optics Express*, vol. 7, no. 4, pp. 155–165, 2000.

[51] L. Elazary and L. Itti, "Interesting objects are visually salient," *Journal of Vision*, vol. 8, no. 3, pp. 1–15, 3 2008.

Eleonora Vig graduated in 2006 with an MSc in Intelligent Systems from the Babeş-Bolyai University, Cluj-Napoca, Romania. In July 2011, she obtained a PhD in Computer Science at the University of Lübeck, Germany. Come September 2011, she will join the Vision Lab at The Rowland Institute at Harvard as a postdoctoral fellow. Her main research focuses on computational vision problems such as saliency prediction and object recognition.

Michael Dorr obtained his undergraduate degree in Computer Science in 2004 and his PhD in Engineering in 2010, both in Erhardt Barth's group at the Institute for Neuro- and Bioinformatics of the University of Lübeck. He is now a postdoctoral fellow at the Schepens Eye Research Institute, Harvard Medical School, where he uses eye tracking and gaze-contingent video processing algorithms to study human vision under natural conditions.

Thomas Martinetz is full professor of computer science and the director of the Institute for Neuro- and Bioinformatics at the University of Lübeck. He studied Physics at the TU München and obtained his doctoral degree in Biophysics at the Beckman Institute for Advanced Science and Technology of the University of Illinois at Urbana-Champaign. From 1991 to 1996 he was project leader "Neural Networks for automation control" at the Corporate Research Laboratories of the Siemens AG in Munich. From 1996 to 1999 he was Professor for Neural Computation at the Ruhr-University of Bochum and head of the Center for Neuroinformatics. Thomas Martinetz is Chairman of the German Chapter of the European Neural Network Society.

Erhardt Barth received the PhD degree in Electrical and Communications Engineering from the Technical University of Munich, Munich, Germany. He is a Professor at the Institute for Neuro- and Bioinformatics, University of Lübeck, Lübeck, Germany, where he leads the research on human and machine vision. He has conducted research at the Universities of Melbourne and Munich, the Institute for Advanced Study in Berlin, and the NASA Vision Science and Technology Group in California. Dr. Barth received a Schloessmann Award in May 2000.