# Learned saliency transformations for gaze guidance

Eleonora Vig[a], Michael Dorr[a,b], and Erhardt Barth[a]

[a]Institute for Neuro- and Bioinformatics, University of Lübeck,
Ratzeburger Allee 160, 23538 Lübeck, Germany;
[b]Schepens Eye Research Institute, Dept. of Ophthalmology, Harvard Medical School,
20 Staniford Street, Boston, MA 02114, USA

## ABSTRACT

The saliency of an image or video region indicates how likely it is that the viewer of the image or video fixates that region due to its conspicuity. An intriguing question is how we can change the video region to make it more or less salient. Here, we address this problem by using a machine learning framework to learn from a large set of eye movements collected on real-world dynamic scenes how to alter the saliency level of the video locally. We derive saliency transformation rules by performing spatio-temporal contrast manipulations (on a spatio-temporal Laplacian pyramid) on the particular video region. Our goal is to improve visual communication by designing gaze-contingent interactive displays that change, in real time, the saliency distribution of the scene.

**Keywords:** spatio-temporal saliency, saliency manipulations, gaze guidance, gaze-contingent displays, Laplacian pyramid

## 1. INTRODUCTION

In our daily life, we are constantly faced with a vast amount of visual information that the human visual system cannot process simultaneously. Despite the illusion that we perceive the entire visual field in full detail, only a small fraction of this information — which falls on the higher resolution area of the retina (the fovea) — can be handled at any one time. The selected locations in the visual field are brought to the fovea and processed by a succession of saccades and fixations that together form a scanpath. The set of mechanisms through which relevant information is selected is called *visual attention* and is known to be determined by two types of factors. First, low-level stimulus properties, such as motion, contrast, and colour, can trigger a fast bottom-up attentional capture. Second, higher cognitive processes, i.e. the viewer's goals and interests, also modulate the attentional selection (top-down attention).

Taking advantage of the gained insights on the cognitive processes involved in attentional selection, various computational models of visual attention have been proposed.[1–4] The ability to predict the salient, and therefore, potentially relevant scene locations has proven to be invaluable for computer vision applications, too, where attention-based algorithms have been proposed e.g. for image and video compression,[5] cropping, quality assessment, and active vision.[6] Such models centre on the concept of a "saliency map", which assigns, to each pixel of an image or video, a saliency value indicating how likely it is that the viewer of the image or video fixates that location due to its (relative) conspicuity. Although the various models differ in their underlying assumptions concerning the model architecture and the formal definition of saliency, they share some properties that make them biologically plausible. In general, by utilizing one or more basic visual features that are known to play a role in attentional control, local contrast of image regions with their surrounding is computed. Visual features, such as orientation, contrast, and colour are extracted separately on multiple scales and then combined together to form a master saliency map. On this map, biological mechanisms, such as winner-take-all competition and inhibition-of-return, are used to shift attention among the salient regions, thus generating a scanpath for an input scene.

Further author information: (Send correspondence to EV)
EV, EB: {vig,barth}@inb.uni-luebeck.de
MD: michael.dorr@schepens.harvard.edu

Apart from the prediction of scanpaths, only very few studies have addressed the intriguing question of how one can change an image or video locally to *influence* the emerging scanpath, i.e. how human gaze can be *guided* by low-level changes to the visual stimulus. In situations where a large visual display (or visual field) needs to be searched for specific information (e.g. driving, analysing medical and geological images), it is often crucial in which order the salient (and relevant) objects and events are attended to, i.e. how we look at a certain visual stimulus. Eye movement studies have shown that in several domains the gaze patterns of experts differ considerably from that of novices. For example, search strategies of expert and novice radiologists are substantially different,[7] and experienced drivers' and pilots' gaze patterns exhibit shorter dwell times and are better defined;[8] in other words, experts have learned to direct their eyes more efficiently. Moreover, in safety-critical situations, such as driving, assistance in where to look next, for example in order not to overlook a pedestrian, can prove more than beneficial.

Recently, we proposed gaze-guidance systems that lead the observer's gaze through a visual scene in order to enforce a predetermined, optimal scanpath and, through this, to aid the information uptake of the human viewer.[9] The goal is to augment human vision with computer vision technology in a least-obtrusive way. Gaze guidance is realized by gaze-contingent interactive displays that use an eye tracker to monitor the viewer's gaze. In order to achieve an alteration of the gaze patterns, the saliency distribution of the visual scene is modified in real time by local changes to the visual input. Based on the original visual input and the eye position of the viewer, first, a limited set of salient, candidate locations is predicted that would attract the user's gaze. Then, using real-time video processing, we increase the probability of being attended (i.e. its saliency) for one candidate location, and simultaneously decrease saliency for all other candidates. That such modifications are not perceived consciously is assured by the fact that they are embedded gaze-contingently in the periphery.

Previously, several attempts had been made to influence gaze patterns, either by filtering potentially salient targets[10, 11] or by adding synthetic gaze attractors such as flashing Gabors.[12, 13] However, these attempts were limited to static natural images and computer-generated content, where eye movements are more idiosyncratic and less driven by bottom-up saliency than on natural movies,[14] and they were also not rendered gaze-contingently, so that subjects presumably quickly became aware of the changes and could consciously decide to ignore their effect.

In the above formulation, a critical issue is to identify optimal image transformations that can make a video region more (or less) eye-catching (i.e. salient) to the viewer. Here, we use a data driven approach to the problem, which aims at *learning*, from eye movements collected on real-world dynamic scenes, how to alter the saliency level of the video locally. We consider a two-class classification scenario in which the video regions fixated by humans form the salient class and non-fixated locations represent the non-salient class. To the best of our knowledge, the general problem of "moving" a sample of a class into the other, in an optimal way and under certain constraints, is novel in the machine learning and computer vision literature.

In the current scenario, transformations are limited to subtle changes in the video patch that go "unnoticed" — as they are embedded gaze-contingently in the periphery — yet still have a gaze guiding effect. In theory, such image modifications could be derived directly in the pixel (or intensity) space of image regions (or patches). However, as natural image patches are known to be samples of an unknown low-dimensional manifold in the space of all possible image patches (i.e. generating an image patch randomly pixel-by-pixel does not give a natural image), transforming them in the original, high-dimensional pixel space will almost always result in unnatural, white noise images. In other words, there are only a limited number of modifications that could be performed on a given image while still preserving its natural look. Moreover, these modification rules would be specific to the image patch at hand, and would not apply to all patches.

Alternatively, one could map the high dimensional pixel patch onto some lower-dimensional (parameter) space by peforming local *feature extraction* on the patch. Such an approach clearly limits the range of possible image modifications to changes in the chosen feature space. This could mean, for example, an increase/decrease in either luminance contrast, colour contrast or intensity, or motion velocity. Nevertheless, it has the advantage that any meaningful feature modification still yields a natural looking image. However, a strong constraint is imposed on the chosen feature space by the need to be able to apply (or map back) the changes in the feature space to the pixel image. Additionally, as we intend to derive transformation rules from information on the salient and non-salient image regions that was obtained with machine learning algorithms, the proposed feature space must be characterized by a good separability of the salient and non-salient classes.

In this paper, we propose to use the *local spectral energy* as a feature space that satisfies the above constraints. It is a low-dimensional representation of a movie patch computed on each level of a spatio-temporal Laplacian pyramid by averaging the squared pixel intensities within the patch. Learned transformations within this space can be implemented as local spatio-temporal contrast manipulations on a spatio-temporal Laplacian pyramid. We show that such transformations lead to a modification of the saliency distribution, which in turn should result in a change in eye movement statistics. In Sec. 2, we present the machine learning framework used for deriving transformations in the spectral energy space. Then, in Sec. 3, we evaluate the effect of the spatio-temporal contrast modifications on saliency distribution in a preliminary experiment, where such energy modifications are embedded offline in a number of real-world videos. The desired effect (an increase or decrease in absolute saliency) is observed in different saliency maps of the modified movies — maps computed by state-of-the-art saliency models for dynamic scenes.

## 2. TRANSFORMATIONS IN THE SPECTRAL ENERGY SPACE

To derive saliency alteration rules, we explore a data-driven approach that takes advantage of learning the discriminative characteristics of salient video regions directly from human-labelled data (i.e. fixated video areas). Note that this approach does not make any assumptions per se on what constitutes saliency in natural movies. Our strategy is to first learn the structural differences between fixated and non-fixated movie regions by building a classifier that operates on the spectral energy representation of the patches, and then use information on the classification boundary to move elements of one class into the other.

### 2.1 Spectral energy as a simple saliency measure

The flow diagram of our joint saliency classification/modification scheme is depicted in Fig. 1. Given a collection of real-world videos, we use eye movements collected on them to label movie areas as either attended or non-attended. The videos are first decomposed into their Laplacian pyramid representation,[15, 16] i.e. a dissection of the original movie into a hierarchy (or pyramid) of videos such that each pyramid level corresponds to a different spatio-temporal frequency band. For each movie location $p = (x, y, z)$ in the two classes (with spatial coordinates $x$ and $y$, and frame number $z$), the local spectral energy is extracted on each level of the spatio-temporal Laplacian pyramid. The spectral energy $e_{s,t}$ on the $s$-th spatial and $t$-th temporal pyramid level ($L_{s,t}$) is computed in a spatial neighbourhood centred around $p$ as

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

where $W_s$ and $H_s$ stand for the width and height of the neighborhood on the $s$-th spatial scale (fewer pixels on lower-resolution spatial scales, but independent of the temporal scale). The spatial coordinates of the location $p$ are also subsampled on the spatial scale $s$: $(x_s, y_s) = (x/2^s, y/2^s)$. We consider a spatial window around a video location because context is known to strongly influence the saliency of the location, and with fixational data one also needs to compensate for spatio-temporal imprecision in both the eye tracking and the human visual system. The size of the window is a free parameter whose value needs to be determined either from data fitting or chosen in accordance with the results of perceptual experiments.

Thus, each video patch, be it attended or not, is described by a feature vector consisting of the spectral energies extracted on the different pyramid levels. With this low-dimensional representation (or energy profile) of a video patch a non-linear kernel support vector machine (SVM) is trained that can discriminate between salient and non-salient movie regions. We here note only briefly that despite its simplicity this algorithm yields similar results to state-of-the-art saliency models. On the collection of videos considered below for evaluation, the leave-one-out ROC score for predicting eye movements — averaged over all 18 movies and after removing biases inherent in eye tracking data — is 0.63 for the above simple algorithm, 0.63 for the classical Itti and Koch model,[1] and 0.64 for SUNDAy.[17] In previous work, however, we could also obtain better results (ROC score of

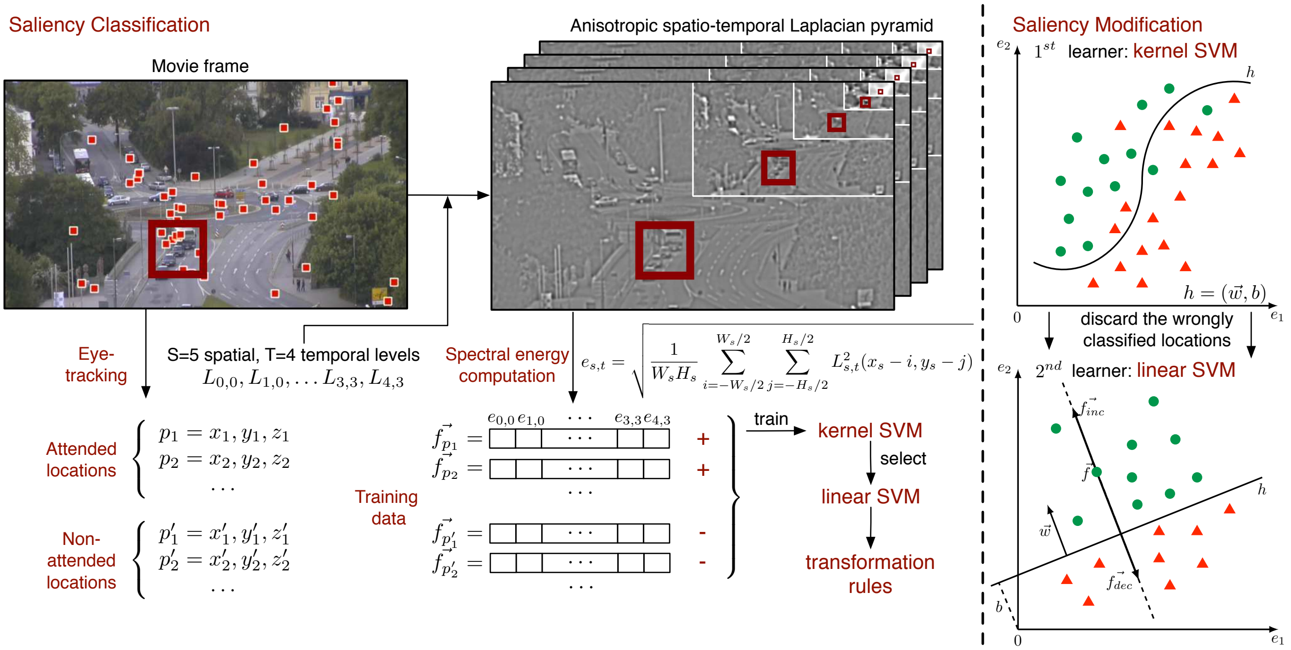

Saliency Classification
Movie frame
Anisotropic spatio-temporal Laplacian pyramid
Eye-tracking
S=5 spatial, T=4 temporal levels
$L_{0,0}, L_{1,0}, \dots L_{3,3}, L_{4,3}$
Spectral energy computation
$e_{s,t} = \sqrt{\frac{1}{W_s H_s} \sum_{i=-W_s/2}^{W_s/2} \sum_{j=-H_s/2}^{H_s/2} L^2_{s,t}(x_s - i, y_s - j)}$
Attended locations
$p_1 = x_1, y_1, z_1$
$p_2 = x_2, y_2, z_2$
$\dots$
Non-attended locations
$p'_1 = x'_1, y'_1, z'_1$
$p'_2 = x'_2, y'_2, z'_2$
$\dots$
Training data
$e_{0,0}\ e_{1,0}\ \cdots\ e_{3,3}\ e_{4,3}$
$\vec{f}_{p_1} =$ +
$\vec{f}_{p_2} =$ +
$\vec{f}_{p'_1} =$ -
$\vec{f}_{p'_2} =$ -
train
kernel SVM
select
linear SVM
transformation rules
Saliency Modification
$1^{st}$ learner: kernel SVM
$e_2$
$e_1$
$h$
$h = (\vec{w}, b)$
0
discard the wrongly classified locations
$2^{nd}$ learner: linear SVM
$\vec{f}_{inc}$
$\vec{f}$
$\vec{w}$
$\vec{f}_{dec}$
$b$

Figure 1. Flow diagram summarizing the proposed approach. In the saliency classification phase (left), a classifier is trained with the spectral energy profiles of attended and non-attended video patches (fixations are denoted by small red (filled) squares in the movie frame). This feature is extracted as the mean-square-root of pixel intensities in a neighbourhood around the locations (large unfilled square) on each level of a spatio-temporal Laplacian pyramid. Right: Schematic view of transformation rules (for illustration purposes, only a two-dimensional feature space is shown: $\vec{f} = (e_1, e_2)$). An iterative SVM approach (kernel + linear SVM) is utilized to learn an optimal separation (a hyperplane $h = (\vec{w}, b)$) of salient (green dots) and non-salient (red triangles) video regions. To avoid saccades to a particular salient region whose energy profile is $\vec{f}$, the patch's energy profile is moved perpendicular to $h$ in the direction of the class of non-salient regions (along $\vec{f}_{\text{dec}}$). To increase the patch's saliency, its energy profile is moved away from $h$ (along $\vec{f}_{\text{inc}}$).

0.67) based on invariants of the structure tensor,[18] which are generic representations that are not invertible and thus cannot be used for saliency modifications.

## 2.2 Spectral energy modification

Support vector machines search for an optimal "hyperplane", a decision boundary that separates the two classes with maximum margin. The hyperplane $h$ is described by a vector $\vec{w}$ perpendicular to the plane and the bias $b$, which specifies its shift from the origin. The closer an instance to the plane, the more difficult it is to classify it into either group, because the more it resembles instances of the other class. The classification confidence of those points located far from the plane is high since, in our case, they are "truly" salient/non-salient video areas. Therefore, in order to change the saliency level of a movie region (in terms of its spectral energy) it suffices to move its energy profile relative to the plane, either towards the plane or away from it. Thus, a separating plane imposes a meaningful direction for transformations of spectral energy profiles in the feature space.

Still, an important question remains: how can we map back a modified feature vector (energy profile) to an image patch? How to apply the learned transformations to the original video patch? Obviously, this mapping can only be approximate, but there are various ways of increasing or decreasing the spectral energy of a video patch. A straightforward approach, applied here, is to multiply every pixel in the patch with the ratio of the desired and actual energy, thus increasing or decreasing contrast in the specific pyramid scale.

One complication in our scenario relates to the fact that the classifier that best discriminates salient video regions from non-salient ones is kernel-based, i.e. it nonlinearly maps its input data into a higher-dimensional space, where the problem becomes linearly separable. The non-linear mapping between the input space and the high-dimensional feature space is performed implicitly using the kernel trick, hence the $\phi$ non-linear embedding function is unknown. As a result, the reverse mapping (with an unknown $\phi^{-1}$) from the feature space back to the input (energy) space of the modified data points is difficult. This is known as the *pre-image problem* in the kernel methods literature. It has been shown that exact pre-images typically do not exist but need to be aproximated, in the process of which they can (easily) get distorted. To remediate the issue of a further non-linear mapping, we reformulate the task of learning a saliency classifier by considering only a subset of the attended and non-attended locations, thereby making the problem "easier". Assuming that the video patches correctly classified by the kernel support vector machine approximate well the manifolds of their respective classes, we train a second, *linear support vector machine* with only these patches, in case of which the separating plane is defined in the input (energy) space — see Fig. 1 for a visual illustration.

Recall that with our problem formulation (gaze guidance through saliency manipulations), the alteration (in terms of the probability of being attended) of only potential gaze-capturing locations is intended. To modify the saliency of a candidate, i.e. salient, video patch, we move its energy profile perpendicular to the separating hyperplane of the linear SVM, either towards the non-salient class (i.e. towards the hyperplane, to make the patch less salient), or away from the hyperplane (to increase its saliency) — as shown schematically in Fig. 1. Thus, for a candidate location with spectral energy vector $\vec{f}$, the transformation rules are defined as

$$\begin{array}{ll} \vec{f_{\text{inc}}} & = \vec{f} + \alpha_1 \vec{w} \frac{b}{||\vec{w}||} \\ \vec{f_{\text{dec}}} & = \vec{f} - \alpha_2 \vec{w} \frac{b}{||\vec{w}||} \end{array} , \tag{2}$$

where $\alpha_i$ denote the degree of change.

One might argue whether the learning of such contrast modification rules (or weights) from eye movement data really is necessary. An analysis of the average spectral energy at attended and non-attended locations reveals that, on every scale, the attended movie regions have higher spectral energy than non-attended ones. Thus, it may suffice to increase/decrease energy by a constant factor — relative to the average spectral energy of the specific class — in each frequency band. However, we chose to learn these weights, since this way the local structure of the manifold of natural video patches is also considered, and the relative weighting of individual frequency bands becomes possible. Different spatio-temporal frequency bands may play different roles in guiding bottom-up attention, and individually weighting them can account for these differences.

To avoid artefacts, such as pixel saturation, due to strong contrast enhancements (occuring in the "saliency-increase" case), elaborate normalization schemes that map back the output videos to pixel intensity values in $[0, 255]$ are required. Because natural videos usually already use up the limited dynamic range of the display, we reduce the to-be-modified videos to $x\%$ overall contrast and adjust the energy weights (through the strength factors $\alpha_i$ in Equation (2) such that the intensity range at the modified location is stretched maximally without overflows. Also, in order to avoid strong and unnatural changes in the candidate video patch, the DC component (i.e. the lowest pyramid level of the Laplacian) is left unaltered.

## 3. EXPERIMENTAL EVALUATION

To evaluate the effect of the spatio-temporal contrast modifications on saliency and eye movements, in a preliminary experiment, we embedded such local energy transformations in high-resolution videos of natural outdoor scenes. Three baseline saliency models for dynamic scenes, the model of Itti and Koch,[1] SUNDAy,[17] and a simple yet powerful approach based on the invariants of the structure tensor,[4] were used to compute saliency maps both for the unmodified and transformed movies. Using statistical tests, we then verify whether the embedded spectral energy modifications really bring the desired change, i.e. an increase or decrease in absolute saliency.

### 3.1 Learning the contrast modification rules

For the experiment, we use a collection of 18 natural videos (1280 by 720 pixels, 29.97 fps, about 20 s duration each, recorded in the $Y'C_bC_r$ format) for which eye movements of 54 human subjects freely viewing these movies are available.[14] From the recorded eye traces overall about 40,000 saccades are extracted using a dual-threshold velocity-based procedure.[19] These fixations are used to find an optimal hyperplane for separating salient and less salient video regions. For the non-salient locations, however, biases inherent in gaze data need to be addressed.

Because eye movements tend to cluster in the centre of the screen (a phenomenon known as the central fixation bias in the human vision literature), one needs to assure a similar, centrally biased distribution of non-attended locations, too. To achieve this, a common approach, which we also follow here, is to use randomly chosen scanpaths from other movies as non-attended locations on a given video. Additionally, assuming that saccades are "responses" to gaze-capturing events, one would also need to consider the oculomotor latency between the event and the saccade associated with it. However, we have previously shown that, possibly due to prediction, the average time lag of saccades in natural videos is near-zero, i.e. no offset needs to be taken into account.[20]

The energy profiles of the attended and non-attended locations are computed on an anisotropic Laplacian pyramid decomposition of the videos (the pyramid having $S = 5$ spatial and $T = 4$ temporal levels), in a 5 by 5 degree spatial neighbourhood on all scales (which corresponds to $128 \times 128$ pixels on the highest spatial levels). In the periphery, the highest spatial and temporal frequency information is known to contribute little to attentional selection, since high spatio-temporal frequency is discernible only near the fovea. Therefore, we leave the energies in these scales (8 out of the 20 pyramid levels) unaltered, i.e. we fix their weights to 1.0. Thus, the soft-margin kernel SVM[21] operates in a low-dimensional space: on the only 12-dimensional vectors containing energies from all but the highest spatial and temporal scales. The optimal SVM parameters, the width of the Gaussian $\gamma$ and the penalty term $C$, are found with 5-fold cross-validation. Different from classical machine learning tasks, here, we do not wish to improve the performance of the above simple classifier on independent test data, but rather optimize it to better fit the given training data. Even though not relevant here, performance on test data is also good (see Sec. 2.1). The quality of prediction on the training data is measured through ROC analysis, which reports an ROC score of 0.82. After discarding the wrongly classified video patches, about 28,000 locations are left per class, with the energy profiles of which a linear SVM is trained. Its $C$ parameter is again determined with 5-fold cross-validation. Now, with this linear SVM, on the selection of "truly" (i.e. easily discriminable) salient and non-salient video patches, an ROC score of 0.819 is achieved. The optimal separating hyperplane $h = (\vec{w}, b)$ found by this linear SVM shall be used to derive the rules in Equation (2).

### 3.2 Embedding the modifications in natural movies

For our evaluations, in the above 18 movies, about every second, 10 candidate locations are determined. In principle, we could have used the above simple saliency predictor based on the spectral energies (or any other state-of-the-art saliency model) to generate these locations. However, for our testing purposes, the most precise determination of gaze-capturing areas is important, and human observers' eye movements are still best predicted by other observers' eye movements. Hence, we created a spatio-temporal fixation density map for each movie by placing a two-dimensional Gaussian with standard deviation 0.75 deg at each gaze sample of the 54 subjects. After normalizing the superposition of these Gaussians, the candidate locations are iteratively extracted from these maps by picking the location with the highest "empirical" saliency, and subsequently laterally inhibiting this location with an inverted Gaussian of standard deviation 2.35 deg. In this way, it is also assured that within a neighbourhood of about $5 \times 5$ deg no overlaps of candidates occur. With Equation (2), for each of these candidates a pair of new spectral energy vectors is computed based on the candidates' actual profiles, which were extracted with the parameters used for the SVM learning. The scalar $\alpha_i$, which controls the degree of change, is first set to a fixed initial value independent of the candidate's energy vector. The rationale is that, at this point, we only define the directions of change in the feature space of spectral energies, and adjust the strength of the modification later, separately for each test condition, in which the effectiveness related to different modification strenghts is examined. Thus, for contrast modifications, initial weights $\vec{w_{\text{inc}}}$ and $\vec{w_{\text{dec}}}$ are derived as the ratio between the desired and actual energies $\frac{\vec{f_{\text{inc}}}}{\vec{f}}$ and $\frac{\vec{f_{\text{dec}}}}{\vec{f}}$, respectively.

As mentioned above, if contrast is increased beyond what the dynamic range of the display allows, artefacts occur. Therefore, to leave room for contrast enhancements, we reduce the overall contrast of our movies by different amounts, and embed modifications in each of these contrast-decreased videos.

The final, video patch-specific saliency-increase weights $\vec{w_{\text{inc}}}'$ are defined for each contrast level so as to stretch the dynamic range in the neighbourhood of the candidate between the extrema (i.e. 0 and 255, black and white – as we are operating on the brightness channel only). Thus, with different overall contrasts, it becomes possible to quantify the strength of the modification and evalutate its effect on saliency. We introduce a simplified notation for the *synthesis* of the Laplacian pyramid: $\sum_{s=0}^{S-1} \sum_{t=0}^{T-1} L_{s,t}$, which in fact involves the iterative upsampling and addition of the Laplacian levels. To avoid overflows, for each pixel $p = (x, y, z)$ in the modified spatio-temporal video patch the following must hold:

$$0 \leq \sum_{s=0}^{S-1} \sum_{t=0}^{T-1} w'_{s,t} L_{s,t}(x, y, z) \leq 255, \tag{3}$$

where $w'_{s,t}$ is the patch-specific weighting coefficient for the spatio-temporal frequency band $(s, t)$. These weights are obtained from the initially derived ones $(w_{s,t})$:

$$w'_{s,t} = (w_{s,t} - 1)\beta + 1, \tag{4}$$

where $\beta$ takes now the role of $\alpha_i$ from Equation (2) in controlling the strength of the manipulation. To stretch the intensity range to the extrema but not beyond, $\beta$ is derived from Equation (3) for each spatio-temporal video patch individually as

$$\beta = \min_{(x,y,z)\in\text{patch}} \begin{cases} \dfrac{255 - \sum_s \sum_t L_{s,t}(x,y,x)}{\sum_s \sum_t w_{s,t} L_{s,t}(x,y,z) - \sum_s \sum_t L_{s,t}(x,y,z)}, & \text{if the denominator} > 0 \\ -\dfrac{\sum_s \sum_t L_{s,t}(x,y,z)}{\sum_s \sum_t w_{s,t} L_{s,t}(x,y,z) - \sum_s \sum_t L_{s,t}(x,y,z)}, & \text{if the denominator} < 0 \end{cases}. \tag{5}$$

For each pixel in the video patch, exactly one of the following conditions holds: (1) the denominator is larger than zero, i.e. the manipulation brings an increase in pixel intensity, and so (in the limit) $\beta$ should stretch the new intensity to 255; (2) the denominator is negative, i.e. the modified pixel intensity is smaller than the original, and should, therefore, be decreased further to 0; (3) the denominator is zero, i.e. the pixel intensity remains the same, hence, $\beta$ is not affected. By picking the smallest ratio over all pixels in the patch, we assure that the modified intensities remain in the allowed range.

The quantification of the strength of the saliency-decrease rules cannot be tied to the overall contrast level of the video. Instead, the strength is varied by scaling the initial $w_{\vec{\text{dec}}}$ weights so that $n$ weights closest to zero are actually brought to zero $(0 \leq n < S * T)$. Setting the energies in certain scales to zero means removing those frequencies.

For the experiment, three saliency-increase and one saliency-decrease strengths were tested; for the increase rules, the original videos were decreased to 70, 80, and 90 percent overall contrast. For simplicity, we only report results for one condition: the 80% overall contrast case. The same qualitative results were obtained in the other two conditions, with the obvious difference that saliency-increase modifications at 70% contrast were stronger than at 80% or 90%. For decrease rules, $n$ was set to 4, i.e. frequencies in four spatio-temporal levels – with weights closest to zero – were removed. Every second, the saliency of 5 randomly chosen candidate points was increased further, and the remaining 5 candidates were decreased in their saliency. For the results reported below, spatio-temporal contrast manipulations were embedded in a 5 by 5 deg spatial and 700 ms temporal neighbourhood centred around the candidates. An example stillshot from a movie and its altered version is shown in the first row of Fig. 2. Lack of temporal change in the printed figure renders the modifications less visible than in the actual movie; however, in the difference map of the two the 10 modified patches are clearly discernible. In this specific frame of the "roundabout" scene, the 5 locations in the upper part of the scene are decreased in saliency, while those in the lower part are increased.

### 3.3 Results

The effect of spectral energy modifications on overall saliency is evaluated by pairwise comparison of the saliency maps of unmodified and transformed videos – maps which were generated by three independent models of bottom-up attention. The first of these is an implementation of the classical Itti and Koch saliency map, the architecture of which we reviewed in the introduction. Here, we use the Maxnorm normalization scheme (based on normalized summation) to fuse the separate feature maps into one master map. SUNDAy, the second model, uses natural image statistics in combination with a Bayesian approach to detect gaze-capturing events. To create the saliency maps for both of these models, their publicly available implementations were used with their default

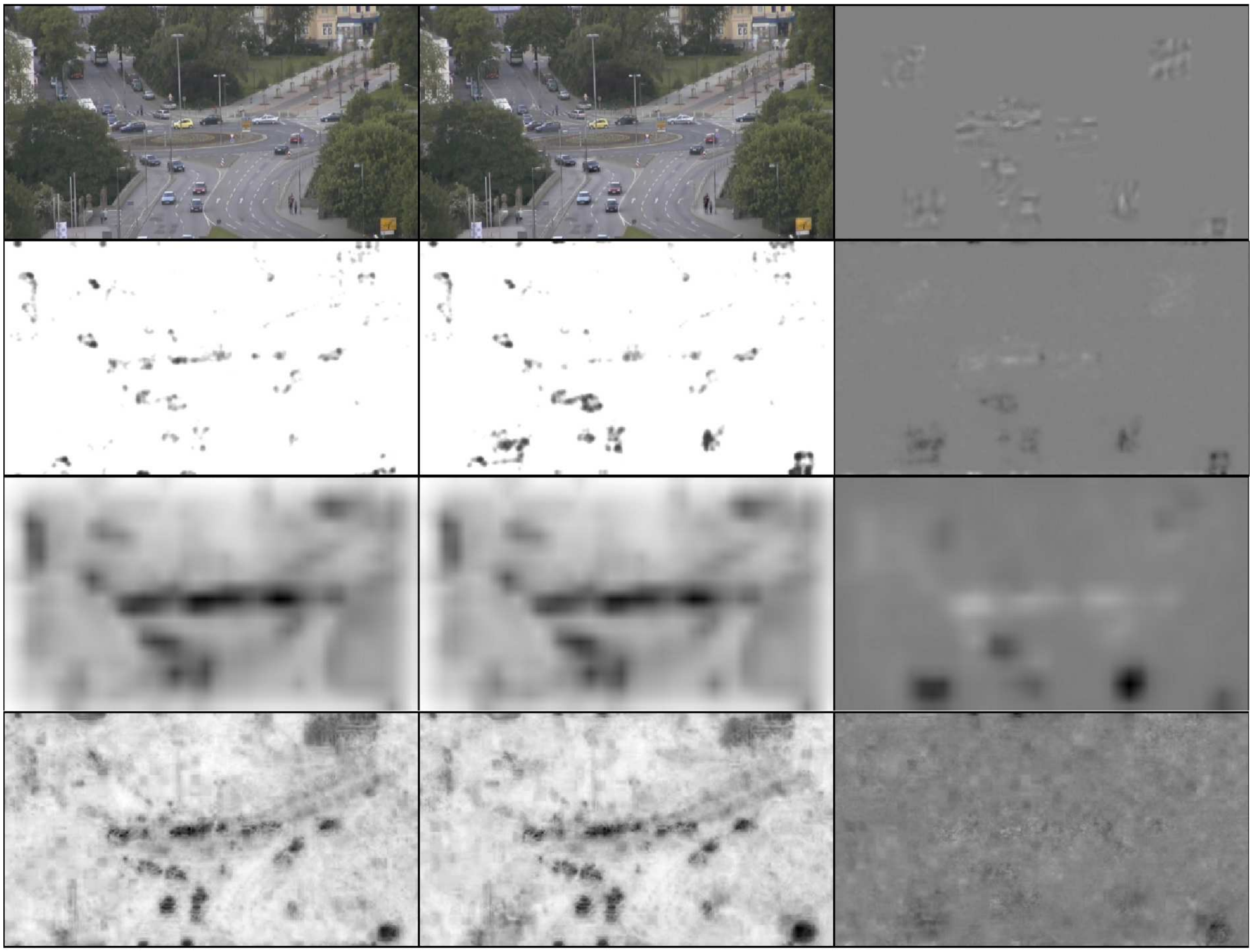
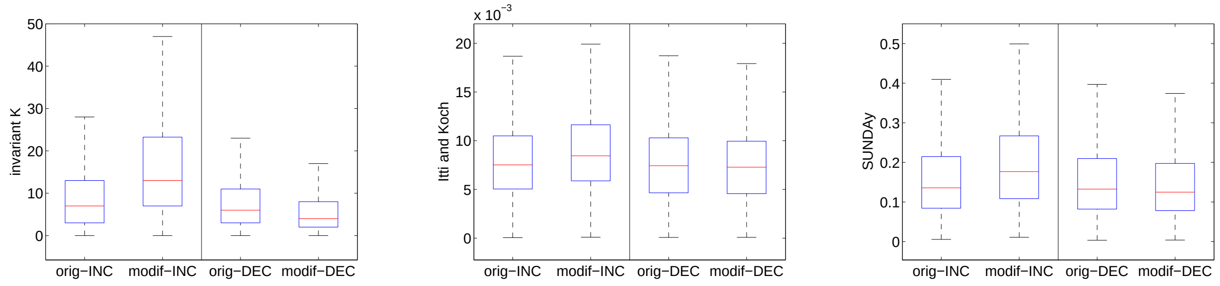


Figure 2. Saliency maps for one frame of an original (first column) and altered (second column) video. Ten non-overlapping candidate (i.e. salient) regions undergo saliency manipulations: the five candidates in the upper part of the scene are reduced in saliency, while the remaining five in the lower part are rendered more salient. Three baseline models are used to obtain the saliency maps: the geometrical invariant $K$ (second row), the model of Itti and Koch (third row), and SUNDAy (last row). In the differences of the saliency maps before and after the modification (third column), the desired alteration in saliency can be clearly detected for the saliency-increase case (dark areas in the difference maps), while the decrease rules (bright areas in the differences) have a weaker effect on the saliency (in particular for SUNDAy).



(a) Invariant K

(b) Itti and Koch

(c) SUNDAy

Figure 3. Box plots comparing the saliency distributions of candidate locations extracted from the saliency maps of original and modified videos. The distributions at saliency-increase (INC) and decrease (DEC) locations are treated separately. In all cases, the differences between the original and modified saliency distributions are statistically significant (paired Wilcoxon signed rank test) (middle line: median, box: upper and lower quartile, whiskers: data extent, outliers not shown).

parameters. Finally, the third algorithm, introduced in Ref. 4, is a simple and fast alternative to state-of-the-art saliency algorithms, and relies on the estimation of the intrinsic dimension, i.e. the degree of variation of the signal, by means of computing the geometrical invariants of the structure tensor. We have previously shown that the more the video signal varies locally, i.e. the higher its intrinsic dimension is, the more it attracts human gaze. The geometrical invariant $K$, which encodes spatial and temporal changes and is computed as the product of the eigenvalues of the structure tensor, even outperforms baseline models in predicting eye movements.[18] The parameters for computation of $K$ were as in Ref. 18. All of these models compute saliency on spatially downsampled versions of the original movie in order to reduce computational cost and to increase resilience against noise. The lowpass-filtered videos (6.6 cycles/degree) were created by filtering the video with a 5-tap spatial binomial filter and downsampling it (in space) by a factor of two. Note, though, that the highest spatial levels remained unchanged in our transformations anyway.

Saliency maps for the “roundabout” scene from Fig. 2 are shown in subsequent rows of the same figure (in the order: invariant $K$, Itti and Koch, and SUNDAy – second to fourth rows). Alterations in the saliency distribution are visually more striking in the image differences (third column) between the saliency maps of unchanged and modified videos. Here, a deviation from the gray value indicates an alteration in the saliency level: at darker areas a saliency-increase occurs, while brighter regions experience a decrease in saliency. Visually, saliency-increase seems to have a more pronounced effect than decrease, especially in the case of the maps computed by SUNDAy.

The saliency of candidate locations before and after the energy modification was compared with a paired Wilcoxon signed rank test, and proved to be significantly different for all three saliency models and both increase and decrease (see Fig. 3). Results confirm our observation on the effectiveness of the modifications: the differences in saliency level are significantly greater where a saliency-increase manipulation was performed than at decrease locations. However, comparing the effectiveness of the two types of changes is not entirely fair, as the quantifications of the strength of manipulation for increase rules is independent of that of the decrease rules. Also, the modifications are the most effective (in changing the saliency distribution) for invariant $K$ ($p = 1.9 \cdot 10^{-240}$ for increase rules, $p = 1.7 \cdot 10^{-165}$ for decrease rules). Nevertheless, the desired effect is reached also in the saliency maps of Itti and Koch (increase, $p = 4.9 \cdot 10^{-213}$; decrease, $p = 5.0 \cdot 10^{-67}$) and SUNDAy (increase, $p = 1.0 \cdot 10^{-145}$; decrease, $p = 8.0 \cdot 10^{-25}$). Unlike invariant $K$, which detects spatio-temporal intensity variations (space-time corners, non-constant translations), the two state-of-the-art models base their prediction of saliency on additional low-level features, such as colour and orientation. This explains why modifications to contrast only have a more modest (yet significant) impact on overall saliency in the case of the Itti and Koch and SUNDAy models.

## 4. CONCLUSION AND OUTLOOK

Redirecting visual attention to certain goal-relevant areas in the visual field is a promising new strategy to integrate into future visual and communication systems. Our goal in this paper was to explore techniques that allow to alter the saliency distribution of the scene, by embedding subtle low-level changes in the visual stimulus. With effective changes that do not introduce objectionable image artefacts, an unconscious gaze guiding process may be achieved.

In this paper, we proposed a generic saliency modification scheme in which, first, the structural differences between attended and non-attended video locations are learnt. The information on the class boundary that separates the two classes was then used to derive the desired image transformations that lead to an alteration in saliency. Our scheme is generic because it does not assume any specific low- or high-level image feature space in which the manipulation rules are derived. However, two constraints have to be met by the selected feature(s). First, for effective saliency transformations, in this space, a high separability of the salient and non-salient video areas is highly desirable. Second, modifications in the chosen feature space need to be mapped to manipulation rules in the original input or pixel space of videos.

The spectral energy, computed on a spatio-temporal Laplacian pyramid, has proven to be a simple feature that fulfils the above constraints. Transformations performed in this low-dimensional space were implemented as local spatio-temporal contrast manipulation rules (on the spatio-temporal Laplacian). Normalization schemes to avoid visual artefacts and ways to quantify the modification strengths were also discussed. Finally, in a preliminary experiment, which aimed at evaluating the potential of such local video manipulations, we used three independent saliency models to compare the saliency maps of the unmodified and altered videos. The desired effect was reached in the saliency maps of modified movies, where a saliency-increase (or -decrease) rule applied to a video patch led to an increase (or decrease) in absolute saliency relative to the original movie patch.

It should be also noted that, since the saliency transformation rules are learned from eye movement data and validated on existing saliency models, our results are also indicative of the biological relevance of these models.

In future work, we shall investigate and empirically validate the effect of gaze-contingent energy modifications on eye movements in a psychophysical experiment.

## 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) 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] Itti, L., Koch, C., and Niebur, E., "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**, 1254–1259 (Nov 1998).

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

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

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

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

[6] Aloimonos, Y., Weiss, I., and Bandyopadhyay, A., "Active vision," *International Journal of Computer Vision* **1**(4), 333–356 (1988).

[7] Nodine, C. F. and Mello-Thoms, C., "The nature of expertise in radiology," in [*Handbook of Medical Imaging, Volume 1. Physics and Psychophysics*], Metter, R. L. V., Beutel, J., and Kundel, H. L., eds., SPIE Press, Bellingham, WA (2000).

[8] Kasarskis, P., Stehwien, J., Hickox, J., and Aretz, A., "Comparison of expert and novice scan behaviors during vfr flight," in [*The 11th International Symposium on Aviation Psychology*], (2001).

[9] Barth, E., Dorr, M., Böhme, M., Gegenfurtner, K. R., and Martinetz, T., "Guiding the mind's eye: improving communication and vision by external control of the scanpath," in [*Human Vision and Electronic Imaging*], Rogowitz, B. E., Pappas, T. N., and Daly, S. J., eds., *Proc. SPIE* **6057** (2006). Invited contribution for a special session on Eye Movements, Visual Search, and Attention: a Tribute to Larry Stark.

[10] Su, S. L., Durand, F., and Agrawala, M., "An inverted saliency model for display enhancement," in [*Proceedings of 2004 MIT Student Oxygen Workshop*], 119–124 (2004).

[11] Nyström, M. and Holmqvist, K., "Effect of compressed offline foveated video on viewing behavior and subjective quality," *ACM Trans. Multimedia Comput. Commun. Appl.* **6**, 4:1–4:14 (February 2010).

[12] McNamara, A., Bailey, R., and Grimm, C., "Search task performance using subtle gaze direction with the presence of distractions," *ACM Transactions on Applied Perception* **6**(3), 1–19 (2009).

[13] Einhäuser, W., Rutishauser, U., Frady, E. P., Nadler, S., König, P., and Koch, C., "The relation of phase noise and luminance contrast to overt attention in complex visual stimuli," *Journal of Vision* **6**(11), 1148–1158 (2006).

[14] Dorr, M., Martinetz, T., Gegenfurtner, K., and Barth, E., "Variability of eye movements when viewing dynamic natural scenes," *Journal of Vision* **10**(10), 1–17 (2010).

[15] Adelson, E. H. and Burt, P. J., "Image data compression with the Laplacian pyramid," in [*Proceeding of the Conference on Pattern Recognition and Image Processing*], 218–223, Los Angeles, CA: IEEE Computer Society Press (1981).

[16] Uz, K. M., Vetterli, M., and LeGall, D. J., "Interpolative multiresolution coding of advanced television with compatible subchannels," *IEEE Transactions on Circuits and Systems for Video Technology* **1**(1), 86–99 (1991).

[17] Zhang, L., Tong, M. H., and 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).

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

[19] Böhme, M., Dorr, M., Krause, C., Martinetz, T., and Barth, E., "Eye movement predictions on natural videos," *Neurocomputing* **69**(16–18), 1996–2004 (2006).

[20] Vig, E., Dorr, M., Martinetz, T., and Barth, E., "Eye movements show optimal average anticipation with natural dynamic scenes," *Cognitive Computation* (2010). (in press).

[21] Chang, C.-C. and Lin, C.-J., *LIBSVM: a library for support vector machines* (2001). Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.