How many bits does it take for a stimulus to be salient?

Sayed Hossein Khatoonabadi[1] Nuno Vasconcelos[2] Ivan V. Bajić[3] Yufeng Shan[4]
[1,3] Simon Fraser University, Burnaby, BC, Canada, skhatoon@sfu.ca, ibajic@ensc.sfu.ca
[2] University of California, San Diego, CA, USA, nuno@uscd.edu
[4] Cisco Systems, Boxborough, MA, USA, yshan@cisco.com

Abstract

Visual saliency has been shown to depend on the unpredictability of the visual stimulus given its surround. Various previous works have advocated the equivalence between stimulus saliency and uncompressibility. We propose a direct measure of this quantity, namely the number of bits required by an optimal video compressor to encode a given video patch, and show that features derived from this measure are highly predictive of eye fixations. To account for global saliency effects, these are embedded in a Markov random field model. The resulting saliency measure is shown to achieve state-of-the-art accuracy for the prediction of fixations, at a very low computational cost. Since most modern cameras incorporate video encoders, this paves the way for in-camera saliency estimation, which could be useful in a variety of computer vision applications.

1. Introduction

Visual attention mechanisms play an important role in the ability of biological vision to quickly parse complex scenes, as well as their robustness to scene clutter. In humans, attention is driven by both the visual stimuli that compose the scene and observer biases that derive from high-level perception. Recently, there has been substantial interest in the modeling of attention mechanisms in computer vision. Most of these efforts have addressed the stimulus driven component, typically through the development of models of visual saliency. While some work has attempted to model the influence of perceptual cues in the saliency process, most works have addressed what is usually defined as *bottom-up* or purely stimulus driven saliency. This has long been believed to be implemented in the early stages of vision, via the projection of the visual stimulus along the features computed by the early stages of visual cortex, and to consist of a center-surround operation. In general, regions of the field of view that are distinctive compared to their surroundings attract attention [9].

A considerable research effort has recently been devoted to the development of computational models of saliency. Early approaches pursued a circuit driven view of the center-surround operation, modeling saliency as the result of center-surround filters and normalization [26]. Under these models, saliency is computed by a network of neurons, where a stimulus similar to its surround suppresses neural responses, resulting in low saliency, while a stimulus that differs from its surround is excitatory, leading to high saliency values. More recently, several works have tried to identify general computational principles for saliency, also applicable to the development of other classes of saliency mechanisms, such as those responsible for top-down saliency effects, or even broader perception [11, 25, 15, 47, 20].

A particularly fruitful line of research has been to connect saliency to probabilistic inference. This draws on a long established view, in cognitive science, of the brain as a probabilistic inference engine [28], tuned to the visual statistics of the natural world [4, 5, 6, 49]. In the cognitive science literature, it has long been proposed that the brain operates as a universal compression device [5], where each layer eliminates as much signal redundancy as possible from its input, while preserving all the information necessary for scene perception. This principle is at the root of many posterior developments in signal processing and computer vision, such as wavelet theory [36], the now widely popular use of sparse representations [42], and, more recently, compression based models of saliency.

These models can be divided into two main classes. A first class of approaches models saliency as a measure of stimulus information. For example, [11, 57, 41] advocate an information maximization view of visual attention, where the saliency of the stimulus at an image location is measured by the self-information [12] of that stimulus, under the distribution of feature responses throughout the visual field. If feature responses at the location have low probability under this distribution, self-information is high and the location considered salient. Otherwise, the stimulus is not salient. [25] proposes a similar idea, denoted Bayesian surprise, which equates saliency to the divergence between a prior feature distribution, collected from surround, and a posterior distribution, computed after observation of feature responses in the center. A second class of approaches equates saliency to a measure of signal compressibility. This consists of producing, at each location, a compressed representation of the stimulus, through a principal component analysis [22, 37, 16], wavelet [46], or sparse decomposition [23, 31] , and measuring the error of stimulus reconstruction from this compressed representation. Incompressible image locations, which produce large reconstruction error, are then considered salient.

In parallel to these conceptual developments, there has also been an emphasis on performance evaluation of different approaches to saliency [10]. These efforts have shown that saliency models based on the compression principle tend to make accurate predictions of eye fixation data. In fact, several of these models predict saliency with accuracy close to the probability of agreement of human eye fixations. It could thus be claimed that "the bottom-up saliency problem is solved." There are, nevertheless, three main problems with the current state-of-the-art. First, while it is true that high accuracy has been extensively documented for image saliency, the same is not true for video, which has been the subject of much less attention. Second, while many implementations of the "saliency as compression" principle have been proposed, much smaller attention has been devoted to implementation complexity. This is of critical importance for many applications of saliency, such as anomaly detection [35] or background subtraction [50] in large camera networks. For such applications, the saliency operation should ideally be performed in the cameras themselves, which would only consume the power and bandwidth necessary to transmit video when faced with salient or anomalous events. This, however, requires highly efficient saliency algorithms. Finally, while many implementations of the compression principle have been proposed for saliency, none has really used a direct measure of compression efficiency. From a scientific point of view, this weakens the arguments in support of the principle.

These observations have motivated us to investigate an alternative measure of saliency, directly tied to compression efficiency. The central idea is that there is no need to define new indirect measures of compressibility, since a direct measure is available at the output of any modern video compressor. In fact, due to the extraordinary amount of research in video compression over the last decades, modern video compression systems operate close to the rate-distortion bounds. It follows that the number of bits produced by a modern video codec is a fairly accurate measure of the compressibility of the video being processed. In fact, because modern codecs work very hard to assign bits efficiently to different locations of the visual field, *the spatial distribution of bits can be seen as a saliency measure,* which directly implements the compressibility principle. Under this view, regions that require more bits to compress are more salient, while regions that require fewer bits are less.

We formalize this idea by proposing the *operational block description length (OBDL)* as a measure of saliency. The OBDL is the minimum number of bits required to compress a given block of video data under a certain distortion criterion. This saliency measure addresses the three main limitations of the state of the art. First, it is a direct measure of stimulus compressibility, namely "how many bits it takes to compress." By leveraging extensive research in video compression, this is a far more accurate measure of compressibility than previous proposals, such as surprise, mutual information, or reconstruction error. Second, it is equally easy to apply to images and video. For example, it does not require *weighting* the contributions of spatial and temporal errors, as the video encoder already uses motion estimation and compensation, and performs rate-distortion optimized bit assignments. Finally, because most modern cameras already contain an on-chip video compressor, it has trivial complexity for most computer vision applications. In fact, it only requires partial decoding of the compressed bit stream, namely the amount of decoding required to determine the number of bits assigned to each image region.

We propose an implementation of the OBDL measure, and show that saliency can be encoded with a simple feature derived from it. However, while video compression systems produce very effective measures of compressibility, this measure is strictly local, since all processing is restricted to image blocks. Saliency, on the other hand, has both a local and global component, e.g. saliency maps are usually smooth. To account for this property we embed the OBDL features in a Markov random field (MRF). Extensive experiments show that the resulting OBDL-MRF saliency measure has excellent accuracy for the prediction of eye fixations in dynamics scenes.

2. Related work

The overwhelming majority of existing saliency models operate on raw pixels, rather than compressed images or video. An excellent review of the state of the art is given in [9, 10]. Nevertheless, some previous works have attempted to make use of compressed video data, such as motion vectors (MVs), block coding modes, motion-compensated prediction residuals, or their transform coefficients, in saliency modeling [33, 2, 32, 40, 14]. This is typically done for efficiency reasons, i.e., to avoid recomputing information already present in the compressed bitstream. The extracted data is a proxy for many of the features frequently used in saliency modeling. For example, the MV field is an approximation to optical flow, while block coding modes and prediction residuals are indicative of motion complexity. Furthermore, the extraction of these features only requires partial decoding of the compressed video file, the recovery of actual pixel values is not necessary.

Our approach is quite different from the majority of these methods, most of which do not even explicitly equate stimulus saliency to compressibility. On the contrary, we pursue the compressibility principle to the limit, proposing to measure saliency with a compressibility score that has not been previously used in the literature. This score, denoted the operational block description length (OBDL), is the total number of bits spent on the encoding of a block of video data. This leverages the fact that modern video compressors encode blocks differentially. A block of image pixels is first predicted from either its temporal (neighboring frames) or spatial (neighboring blocks) surround. The prediction residual is then compressed, using a combination of quantization and entropy coding, and transmitted. When both prediction operations are ineffective, the process results in large prediction residuals and the block requires more bits to compress. By measuring this number of bits, the OBDL is an indicator of the predictability of the block.

The OBDL also generalizes many of the previously proposed compression-based measures of saliency. For example, the representation of a block into a series frequency discrete cosine transform (DCT) coefficients resembles the subspace [22, 16, 37], sparse [23, 31] or independent component [11] decompositions at the core of various saliency measures, the differential encoding of DCT coefficients, by subtracting the values of neighboring blocks, resembles the center-surround operations of [26], and the encoding of motion compensated residuals resembles the surprise mechanism of [25]. In fact, given the well known convergence of modern entropy coders to the entropy rate of the source being compressed

$$H = \frac{1}{n} \sum_i \log \frac{1}{p(x_i)}, \tag{1}$$

where $p(x)$ is the probability of symbol x, the number of bits produced by the entropy coder is a measure of the self information of each block. Hence, a video compressor is a very sophisticated implementation of the saliency principle of [11], which evaluates saliency as

$$S(x) = \log \frac{1}{p(x)}. \tag{2}$$

While [11] proposes a simple independent component analysis to extract features x from the image pixels, the video compressor performs a sequence of operations involving motion compensated prediction, DCT transform of the residuals, predictive coding of DCT coefficients, quantization, and entropy coding, all within a rate-distortion optimization framework.

This results in a much more accurate measure of information and, moreover, is much simpler to obtain in practice, given the widespread availability of video codecs. The proposed OBDL is even simpler to extract from compressed bitstreams than the other forms of compressed-domain information mentioned above, because the recovery of MVs or residuals is not required. Overall, the OBDL combines the accuracy of the non-compressed domain saliency measures with the computational efficiency of their compressed-domain counterparts.

3. Features derived from OBDL

In this section we introduce the OBDL and provide some evidence for its ability to predict eye fixations.

3.1. The OBDL

Typical video compression consists of motion estimation and motion-compensated prediction, followed by intra-prediction, transformation, quantization and entropy coding of prediction residuals and motion vectors. Most of these steps have been in place since the earliest video coding standards, albeit becoming more sophisticated over time. While, for concreteness, we focus on the H.264/AVC coding standard [56], the feature computations proposed here can be adjusted to other video coding standards, including the latest high efficiency video coding (HEVC) [51]. Due to the focus on H.264/AVC, our "block" is a 16×16-pixel macroblock, abbreviated MB.

The OBDL is computed directly from the output of the entropy decoder, which is the first processing block in a video decoder. No further decoding of the compressed bitstream is needed. The number of bits spent on encoding each MB is extracted and mapped to the unit interval $[0, 1]$, where the value of 0 is assigned to the MB(s) requiring the least bits to code and the value of 1 is assigned to the MB(s) requiring the most bits to code, among all MBs in the frame. The normalized OBDL map is smoothed by convolution with a 2D Gaussian of standard deviation equal to 2° of visual angle. Although the spatially smoothed OBDL map is already a solid saliency measure, we observed that an additional improvement in the accuracy of saliency predictions is possible by performing further temporal smoothing. This conforms with what is known about biological vision [3, 1, 39], where temporal filtering is known to occur in the earliest layers of visual cortex. Specifically, we apply a simple causal temporal averaging over 100 ms to obtain a feature derived from the OBDL.

3.2. Prediction of eye fixations

We have performed some preliminary experiments to compare the statistics of the OBDL feature at human fixation points and non-attended locations, in video. These experiments were based on the protocol of Reinagel and

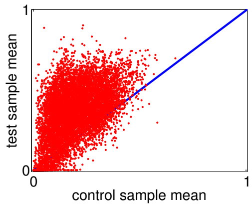

Figure 1. Scatter plots of the pairs (control sample mean, test sample mean) in each frame for OBDL-derived feature. Dots above the diagonal show that feature values at fixation points are higher than at randomly selected points.

Zador [45], who performed an analogous analysis for spatial contrast and local pixel correlation, in natural images. Their analysis compared fixation to random points of still images, showing that, on average, spatial contrast is higher and local pixel correlation lower around fixation points.

We follow the same protocol, using two eye-tracking datasets, DIEM [38] and SFU [18], which contain fixation points of human observers on video clips. Each video was encoded in the H.264/AVC format using the FFMPEG library (www.ffmpeg.org) with a quantization parameter (QP) of 30 1/4-pixel motion vector accuracy with no range restriction, and up to four motion vectors per MB. In each frame, feature values at fixation points were selected as the test sample, while feature values at non-fixation points were used as the control sample. The latter was obtained by applying a nonparametric bootstrapping technique [13] to all non-fixation points of the video frame. Control points were sampled with replacement, multiple times, with sample size equal to the number of fixation points. The average of the feature values over all bootstrap samples was taken as the control sample mean.

Fig. 1 presents pairs of (control sample mean, test sample mean) values for the spatio-temporal filtered OBDL feature. Each red dot represents a video frame. It is clear that, on average, feature values are higher at fixation points than at randomly-selected non-fixation points. To validate this hypothesis, we perform a two-sample t-test [29] using the control and test sample of each sequence. The null hypothesis was that the two samples originate in populations of the same mean. This hypothesis was rejected by the two-sample t-test, at the 0.1% significance level, for all sequences. Note that we have used a very strict 0.1% significance level, as compared to the more conventional (and looser) 1% and 5% levels. The p-values obtained for each video sequence are listed in Table 1, along with the percentage of frames where the test sample mean is greater than the control sample mean. Overall, these results confirm that the OBDL-derived feature is a strong predictor of fixations.

4. OBDL-MRF saliency estimation model

In this section, we describe a measure of visual saliency based on a Markov random field (MRF) model of OBDL feature responses.

4.1. MRF model

While video compression algorithms are very sophisticated estimators of local information content, they only produce *local* information estimates, since all the processing is spatially and temporally localized to the MB unit. On the other hand, saliency has both a local and a global component. For example, many saliency models implement inhibition of return mechanisms [26], which suppress the saliency of image locations in the neighborhood of a saliency peak. To account for these effects, we rely on a MRF model [55, 30].

More specifically, the saliency detection problem is formulated as one of inferring the *maximum a posteriori* (MAP) solution of a spatio-temporal Markov random field (ST-MRF) model. This is defined with respect to a binary classification problem, where salient blocks of 16×16 pixels belong to class 1 and non-salient blocks to class 0. The goal is to determine the class labels $\omega^t \in \{0, 1\}$ of the blocks of frame t, given the labels $\omega^{1\cdots t-1}$ of the previous frames, and all previously observed compressed information $o^{1\cdots t}$. The optimal label assignment ω_*^t is that which maximizes the posterior probability $p(\omega^t|\omega^{1\cdots t-1}, o^{1\cdots t})$. By application of Bayes rule this can be written as

$$\begin{aligned} & p(\omega^t|\omega^{1\cdots t-1}, o^{1\cdots t}) \propto \\ & \quad \propto \quad p(\omega^{1\cdots t-1}|\omega^t, o^{1\cdots t}) \cdot p(\omega^t|o^{1\cdots t}) \\ & \quad \propto \quad p(\omega^{1\cdots t-1}|\omega^t, o^{1\cdots t}) \cdot p(o^{1\cdots t}|\omega^t) \cdot p(\omega^t), \end{aligned} \tag{3}$$

where $\propto$ denotes equality up to a normalization constant. Using the Hammersley-Clifford theorem [7], the MAP solution is

$$\begin{aligned} \omega_*^t = \arg\min_{\psi \in \Omega^t} \Bigg\{ & \frac{1}{T_t} E(\psi; \omega^{1\cdots t-1}, o^{1\cdots t}) \\ & + \frac{1}{T_o} E(\psi; o^{1\cdots t}) + \frac{1}{T_c} E(\psi) \Bigg\}, \end{aligned} \tag{4}$$

where Ω^t is the set of all possible labeling configurations for frame t, $E(\cdot)$ are energy functions, and T_i a constant.

The energy functions $E(\psi; \omega^{1\cdots t-1}, o^{1\cdots t})$, $E(\psi; o^{1\cdots t})$, and $E(\psi)$ measure the degree of *temporal* consistency of the saliency labels, the *coherence* between labels and feature observations, and the *spatial compactness* of the label field, respectively. A more precise definition of these three components is given in the following sections. Finally, the minimization problem (4) is solved by the method of iterated conditional modes (ICM) [8].

Table 1. Results of statistical comparison of test and control samples. For each sequence, the p-value of a two-sample t-test and the percentage (%) of frames where the test sample mean is larger than the control sample mean are shown.

Seq.	*Bus*	*City*	*Crew*	*Foreman*	*Garden*	*Hall*	*Harbour*	*Mobile*	*Mother*	*Soccer*	*Stefan*	*Tempete*	*blicb*	*bws*	*ds*
p-value	10^{-112}	10^{-16}	10^{-29}	10^{-10}	10^{-52}	10^{-211}	10^{-83}	10^{-58}	10^{-120}	10^{-68}	10^{-53}	10^{-31}	10^{-79}	10^{-82}	10^{-50}
%	99	73	83	56	90	96	98	88	100	94	98	89	96	98	92

Seq.	*abb*	*abl*	*ai*	*aic*	*ail*	*hp6t*	*mg*	*mtnin*	*ntbr*	*nim*	*os*	*pas*	*pnb*	*ss*	*swff*	*tucf*	*ufci*
p-value	10^{-36}	10^{-86}	10^{-37}	10^{-192}	10^{-162}	10^{-42}	10^{-49}	10^{-87}	10^{-57}	10^{-16}	10^{-123}	10^{-112}	10^{-12}	10^{-135}	10^{-16}	10^{-22}	10^{-7}
%	76	92	73	100	98	84	91	83	98	68	97	99	56	96	60	86	65

4.2. Temporal consistency

Given a block at image location $\mathbf{n} = (x, y)$ of frame t, the spatio-temporal neighborhood $N_\mathbf{n}$ is defined as the set of blocks $\mathbf{m} = (x', y', t')$ such that $|x - x'| \leq 1$, $|y - y'| \leq 1$ and $t - L < t' < t$ for some L. The temporal consistency of the label field is measured locally, using

$$E(\psi; \omega^{1\cdots t-1}, o^{1\cdots t}) = \sum_{\mathbf{n}} E_t(\mathbf{n}), \tag{5}$$

where $E_t(\mathbf{n})$ is a measure of inconsistency within $N_\mathbf{n}$, which penalizes temporally inconsistent label assignments, i.e., $\omega^t(x, y) \neq \omega^{t'}(x', y')$.

The saliency label $\omega(\mathbf{m})$ of block $\mathbf{m}$ is assumed to be Bernoulli distributed with parameter proportional to the strength of features $o(\mathbf{m})$, i.e. $P(\omega(\mathbf{m})) = o(\mathbf{m})^{\omega(\mathbf{m})}(1 - o(\mathbf{m}))^{1-\omega(\mathbf{m})}$. It follows that the probability $b(\mathbf{n}, \mathbf{m})$ that block $\mathbf{m}$ will bind with block $\mathbf{n}$ (i.e. have label $\psi(\mathbf{n})$) is

$$b(\mathbf{n}, \mathbf{m}) = o(\mathbf{m})^{\psi(\mathbf{n})}(1 - o(\mathbf{m}))^{1-\psi(\mathbf{n})}. \tag{6}$$

The consistency measure weights this probability by a similarity function, based on a Gaussian function of the distance between $\mathbf{n}$ and $\mathbf{m}$,

$$d(\mathbf{n}, \mathbf{m}) \propto \exp\left(\frac{-d_s(\mathbf{m}, \mathbf{n})}{2\sigma_s^2}\right) \exp\left(\frac{-d_t(\mathbf{m}, \mathbf{n})}{2\sigma_t^2}\right), \tag{7}$$

where $d_s(.,.)$ and $d_t(.,.)$ are the Euclidean distances along the spatial and temporal dimension, respectively, and σ_s^2, σ_t^2 two normalization parameters. The expected consistency between the two locations is then

$$c(\mathbf{n}, \mathbf{m}) = \frac{b(\mathbf{n}, \mathbf{m}) d(\mathbf{n}, \mathbf{m})}{\sum_{\mathbf{m} \in N_\mathbf{n}} b(\mathbf{n}, \mathbf{m}) d(\mathbf{n}, \mathbf{m})}. \tag{8}$$

This determines a prior expectation for the consistency of the labels, based on the observed features $o(\mathbf{m})$. The energy function then penalizes inconsistent labelings, proportionally to this prior expectation of consistency

$$E_t(\mathbf{n}) = \sum_{\mathbf{m} \in N_\mathbf{n}} c(\mathbf{n}, \mathbf{m}) \left(1 - \omega(\mathbf{m})\right)^{\psi(\mathbf{n})} \omega(\mathbf{m})^{1-\psi(\mathbf{n})}. \tag{9}$$

Note that $E_t(\mathbf{n})$ ranges from 0 to 1, taking the value 0 when all neighboring blocks $\mathbf{m} \in N_\mathbf{n}$ have the same label as block $\mathbf{n}$, and the value 1 when neighboring blocks all have label different than $\psi(\mathbf{n})$.

4.3. Observation coherence

The incoherence between the observation and label fields at time t is measured with an energy function $E(\psi; o^{1\cdots t})$. While this supports the dependence of w^t on all prior observations ($o^{1\cdots t-1}$), we assume that the current labels are dependent only on the current observations (o^t). Incoherence is then measured by the energy function

$$E(\psi; o^{1\cdots t}) = \sum_{\mathbf{n}} \left(\inf_{\mathbf{p}} o(\mathbf{p})\right)^{1-\psi(\mathbf{n})} \left(1 - \sup_{\mathbf{p}} o(\mathbf{p})\right)^{\psi(\mathbf{n})}, \tag{10}$$

where infimum $\inf(.)$ and supremum $\sup(.)$ are defined over $\mathbf{p} = (x', y')$ such that $|x - x'| \leq 1$, $|y - y'| \leq 1$. This is again in $[0, 1]$ and penalizes the labeling of block $\mathbf{n}$ as non-salient, i.e., $\psi(\mathbf{n}) = 0$ when the infimum of feature value $\inf_\mathbf{p} o(\mathbf{p})$ is large, or as salient, i.e., $\psi(\mathbf{n}) = 1$, when the supremum of feature value $\sup_\mathbf{p} o(\mathbf{p})$ is small.

4.4. Compactness

In general, the probability of a block being labeled salient should increase if many of its neighbors are salient. The last energy component in (4) encourages this type of behavior. It is defined as

$$E(\psi) = \sum_{\mathbf{n}} \Phi(\mathbf{n})^{1-\psi(\mathbf{n})} \left(1 - \Phi(\mathbf{n})\right)^{\psi(\mathbf{n})}, \tag{11}$$

where $\Phi(\mathbf{n})$ is a measure of saliency in the neighborhood of $\mathbf{n}$. This is defined as

$$\Phi(\mathbf{n}) = \alpha \sum_{\mathbf{m} \in \mathbf{n}^+} \psi(\mathbf{m}) + \beta \sum_{\mathbf{m} \in \mathbf{n}^\times} \psi(\mathbf{m}), \tag{12}$$

where $\mathbf{n}^+$ and $\mathbf{n}^\times$ are, respectively, the first-order (North, South, East, and West) and the second-order (North-East, North-West, South-East, and South-West) neighborhoods of block $\mathbf{n}$. In our experiments, we set $\alpha = \frac{1}{6}$ and $\beta = \frac{1}{12}$, to give higher weight to first-order neighbors.

4.5. Optimization

The solution of (4) can be found with many numerical procedures. Two popular methods are stochastic relaxation (SR) [17] and ICM [8]. SR has been reported to have some advantage in accuracy over ICM, but at a higher computational cost [53]. In this work, we adopt ICM, mainly due

Table 2. Saliency prediction algorithms used in our evaluation. D: input data (cmp: compressed; pxl: pixel); I: Implementation (M: Matlab; P: Matlab p-code; C: C/C++; E: Executable).

#	Algorithm	First Author	Year	D	I
1	**PMES**	Ma [33]	2001	cmp	M
2	**MAM**	Ma [34]	2002	cmp	M
3	**PIM-ZEN**	Agarwal [2]	2003	cmp	M
4	**PIM-MCS**	Sinha [48]	2004	cmp	M
5	**MCSDM**	Liu [32]	2009	cmp	M
6	**MSM-SM**	Muthuswamy [40]	2013	cmp	M
7	**PNSP-CS**	Fang [14]	2014	cmp	M
8	**MaxNorm**	Itti [26]†	1998	pxl	C
9	**Fancy1**	Itti [24]†	2004	pxl	C
10	**SURP**	Itti [25]†	2006	pxl	C
11	**GBVS**‡	Harel [19]	2007	pxl	M
12	**STSD**	Seo [47]	2009	pxl	M
13	**SORM**	Kim [27]	2011	pxl	E
14	**AWS**	Diaz [16]	2012	pxl	P

† ilab.usc.edu/toolkit
‡ DIOFM channels (DKL-color, Intensity, Orientation, Flicker, and Motion)

to its simplicity. The label of each block is initialized according to the corresponding feature value, $o(\mathbf{n})$, i.e. the block is labeled salient if $o(\mathbf{n}) > 0.5$ and non-salient otherwise. Each block is then relabeled with the label (0 or 1) that produces the largest reduction in the energy function. This relabeling is iterated until no energy reduction is possible. We limit the iterations to eight in our experiment. It is worth mentioning that ICM is prone to local minimum and the results are dependent on the initial labeling.

4.6. Final saliency map

The procedure above produces the most probable, a posteriori, map of salient block labels. To emphasize the locations with higher probability of attracting attention, the OBDL of a block declared salient (non-salient) by the MRF is increased (decreased) according to the OBDLs in its neighborhood. The process is formulated as

$$S(\mathbf{n}) = \begin{cases} \inf_{\mathbf{q}} \{o(\mathbf{q}) d(\mathbf{q}, \mathbf{n})\} & \psi(\mathbf{n}) = 1 \\ 1 - \inf_{\mathbf{q}} \{(1 - o(\mathbf{q}))\, d(\mathbf{q}, \mathbf{n})\} & \psi(\mathbf{n}) = 0 \end{cases} \quad (13)$$

where $\mathbf{q} = (x', y', t')$ is defined as the set of blocks such that $|x - x'| \leq 1$, $|y - y'| \leq 1$ and $t - L < t' \leq t$, and $d(\mathbf{q}, \mathbf{n})$ as in (7). In this way, a block $\mathbf{n}$ labeled as salient by the MRF inference is assigned a saliency equal to the largest feature value within its neighborhood, weighted by its distance from $\mathbf{n}$. On the other hand, for a block $\mathbf{n}$ declared as non-salient, this operation is applied to the complement of the saliency values within $N_{\mathbf{n}}$. The complement of this value is then assigned as the saliency value of $\mathbf{n}$.

5. Experimental Results

In this section, we report on various experiments designed to evaluate the performance of the OBDL-MRF. The MATLAB code and data used in this study is available at www.sfu.ca/∼ibajic/software.html and www.svcl.ucsd.edu/publications.

Table 3. Average processing time (ms) per frame.

Algorithm	AWS	GBVS	PNSP-CS	MAM	PMES	SURP	STSD	Fancy1	SORM	MaxNorm	PIM-ZEN	**OBDL-MRF**	MCSDM	PIM-MCS	MSM-SM
T	1559	923	895	778	579	323	227	98	92	89	43	**39**	15	10	8

5.1. Experimental set-up

The proposed algorithm was compared with a number of state-of-the-art algorithms for saliency estimation in video, which are listed in Table 2. Among these, only AWS [16] is a purely spatial model. AWS is frequently cited as an accurate predictor of eye fixations in still images and videos. For each algorithm, the table indicates whether its features are computed from raw pixel values (pxl) or compressed video data (cmp), as well as some implementation details.

Evaluation was carried out on the DIEM [38] and SFU [18] datasets. On DIEM, 20 sequences, similar to those used by [10], were chosen. To match the length of the SFU sequences, only the first 300 frames of the DIEM sequences were used in the experiments. Since DIEM videos have various resolutions, they were first resized to 288 pixels in height, while preserving the original aspect ratio. This resulted in five resolutions: 352×288, 384×288, 512×288, 640×288 and 672×288. All sequences were encoded in the H.254/AVC format with FFMPEG (www.ffmpeg.org) using QP $\in \{3, 6, ..., 51\}$ in the baseline profile, with default Group-of-Pictures (GOP) structure.

Two popular metrics were used to evaluate the accuracy of the eye fixation predictions of the various algorithms: area under the receiver operating characteristic Curve (AUC) [52] and normalized scanpath saliency (NSS) [44]. Both were corrected for center bias and border effects (what is usually referred to as "shuffled") as suggested by [43, 54], i.e. by sampling control points more often from the middle of the frame than from its boundaries.

5.2. MRF configurations

We started with a number of experiments that tested the role of the different components of the saliency detector in its performance. The first set of experiments tested the impact of the MRF inference in the saliency judgments. We compared the performance of the saliency measure of (13) for various MRF settings. These included 1) no MRF, where saliency was measured by the OBDL feature responses $o(\mathbf{m})$, including raw OBDL, spatial-filtered OBDL (OBDL-S) and spatio-temporal filtered OBDL (OBDL-T) and 2) MRFs that implemented only subsets of the energy function of (4), indicated by T, O and C for temporal consistency, observation coherence and compactness, respectively. For example, OBDL-MRF-TC means that the MRF model imple-

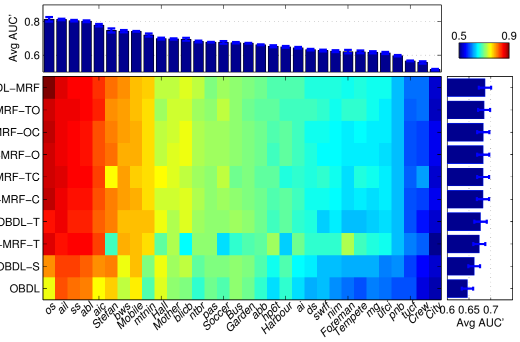

Figure 2. Accuracy of various MRF settings. Each 2D color map shows the average AUC score of each setting on each sequence. *Topbar*: Average AUC score for each sequence, across all settings. *Sidebar*: Average AUC scores each setting across all sequences. Error bars represent standard error of the mean, $\sigma/\sqrt{n}$, where σ is the sample standard deviation of n samples. Sequences from the SFU dataset are indicated with capital first letter.

mented only temporal consistency and compactness. This was implemented by setting subsets of the temperature values to infinity. In our example, $T_o = \infty$. The temperature constants were otherwise set to 1. The temporal support parameter L of Section 4.2 was set to 500ms. Fig. 2 shows the average AUC score of the various MRF settings, across test sequences. The average AUC performance across sequences/settings is shown in the sidebar/topbar. Note that the simple temporally filtered OBDL (OBDL-T) achieves good performance. The global fusion of saliency information, by the MRF provides some additional gains. In general, the addition of more components to the energy function results in improved predictions, with best results produced by the full-blown OBDL-MRF.

5.3. Comparison to the state-of-the-art

A set of experiments was performed to compare the OBDL-MRF to state-of-the-art saliency algorithms. These experiments used quantization parameter QP $= 30$, i.e. reasonably good video quality - average peak signal-to-noise (PSNR) across encoded sequences of 35.8 dB.

We start by comparing the processing times of the various saliency measures in Table 3. All times report to implementation on an Intel (R) Core (TM) i7 CPU at 3.40 GHz and 16 GB RAM running 64-bit Windows 8.1. As expected, compressed-domain measures tend to require far less processing time than their pixel-domain counterparts. The proposed OBDL-MRF, implemented in MATLAB, required an average of 39 ms per video frame. While this is slower than some of the compressed-domain algorithms, it enables the computation of saliency at close to 30 fps. This is enough for most applications of computer vision.

Fig. 3 shows the average AUC (top figure) and NSS scores (bottom figure) of various algorithms, across test sequences. On average, pixel-based methods perform better than those based on compressed video. This trend is however disrupted by the OBDL-MRF, which achieves the best performance. Figure 5 illustrates the differences between the saliency predictions of various algorithms.

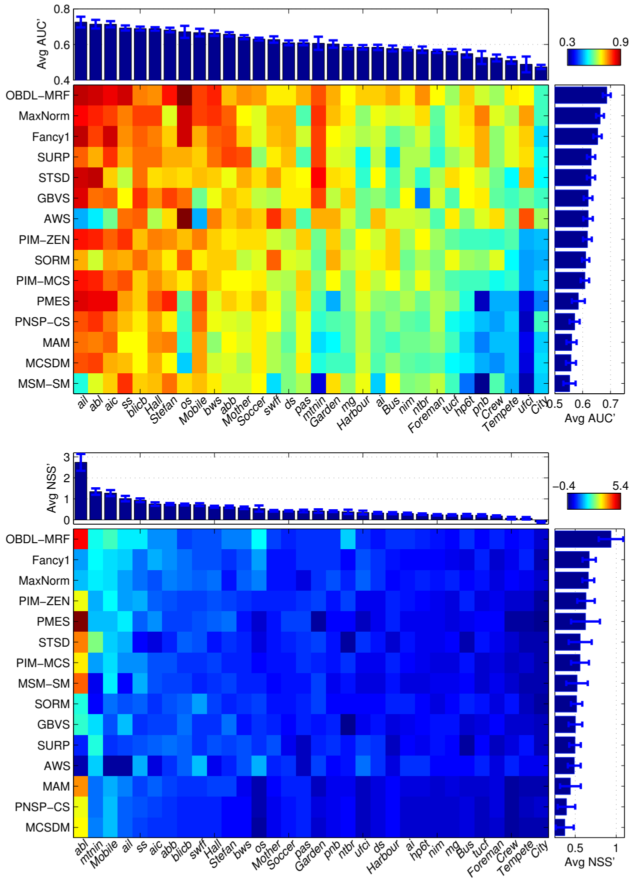

Figure 3. Accuracy of various saliency algorithms over the two datasets according to (top) AUC and (bottom) NSS scores.

The performances of the different saliency measures were also evaluated with a multiple comparison test [21]. This involves computing, for each sequence and measure, both the average score (across all frames) of the saliency measure and the 95% confidence interval for the average score. A set of top performers is then selected for the sequence. This includes the measure of highest average score and all other measures whose 95% confidence interval overlaps with that of the highest-scoring measure. The number of appearances of the different saliency measures among the top performer class is shown in Fig. 4(a). Again, pixel-based methods tend to do better than compressed-based methods and all methods underperform the OBDL-MRF.

5.4. Sensitivity to the amount of compression

Since a compressed video representation always involves some amount of information loss, it is important to deter-

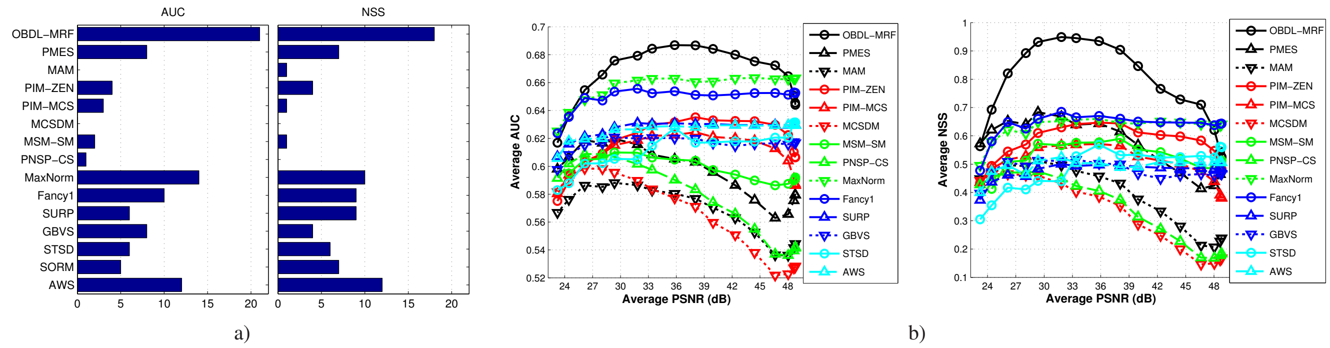

Figure 4. a) Number of appearances among top performers, under AUC and NSS. b) Impact of average PSNR on saliency predictions.

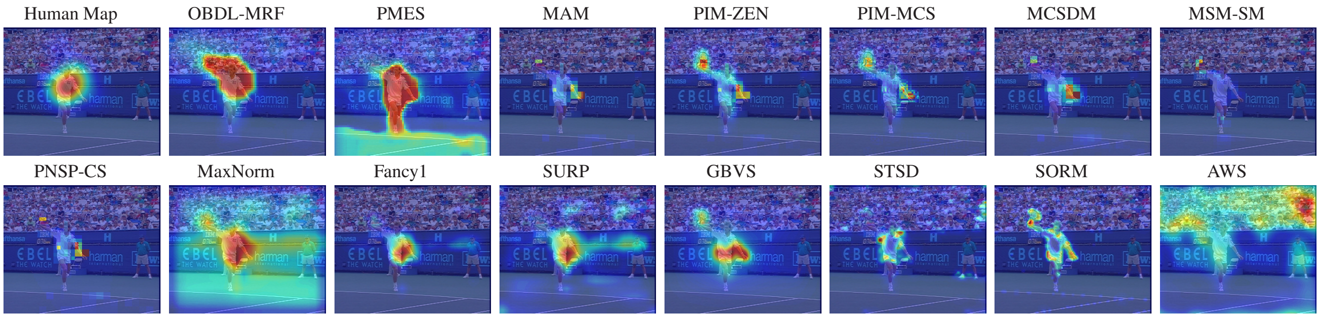

Figure 5. Saliency maps obtained by various algorithms on a video frame.

mine the sensitivity of the saliency measure to the amount of this loss. This question is particularly pertinent for the OBDL, since the predictive power of bit counting could change dramatically across compression regimes. Obviously, in the limit of "zero-bit encoding," the proposed OBDL-MRF will not be a very good saliency predictor. To study this question, we repeated the experiments above for different amounts of compression loss, by varying the QP parameter. The quality of the encoded video, measured in terms of PSNR, drops as the QP increases. Fig. 4(b) shows how the average AUC and NSS scores change as a function of the average PSNR (across sequences), by choosing $QP \in \{3, 6, ..., 51\}$. Interestingly, some of the methods that exhibit largest sensitivity to compression artifacts (such as AWS or GBVS) are not compressed-domain approaches.

Somewhat surprisingly, saliency predictions degrade for *both* very low and high quality video. For most methods, it appears that an intermediate PSNR leads to the best performance. This could be because, at intermediate PSNRs, compression algorithms act as mild low-pass filtering operators, eliminating some of the video sequence noise. It appears that many of the algorithms are sensitive to such noise. With respect to the OBDL-MRF, the accuracy of saliency predictions degrades substantially at the extremes of the compression range. While at low rates there are too few bits to enable a precise measurement of saliency, at high rates there are too many bits available, and all image regions become salient. In any case, the OBDL-MRF achieves the top scores for the overwhelming majority of the compression range. It is also encouraging that saliency estimation is most accurate in the middle of this range, since this is the preferred operating point for most vision applications.

6. Conclusion

We proposed a model of visual saliency based on the compressibility principle. While, at a high level, this is similar to well-known saliency models, such as those based on self-information and surprise, it has the distinct advantage of being readily available at the output of any video encoder, which already exists in most modern cameras. Furthermore, the compressibility measure now proposed naturally takes into account the trade-off between spatial and temporal information, because the video encoder already performs rate-distortion optimization to produce the best predictions for different video regions. In this sense, the proposed solution is a much more sophisticated measure of compressibility than previous measures, based on reconstruction error or cruder measurements of self information. The resulting saliency measure was shown highly accurate for the prediction of eye fixations, achieving state of the art results on standard benchmarks. This is complemented by very low complexity, which makes it appropriate for in-camera saliency estimation.

Acknowledgments: This work was supported in part by NSERC grant RGPIN 327249, NSF award IIS-1208522, and Cisco Research Award CG# 573690.

References

[1] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am*, 2(2):284–299, 1985. 3

[2] G. Agarwal, A. Anbu, and A. Sinha. A fast algorithm to find the region-of-interest in the compressed MPEG domain. In *Proc. IEEE ICME'03*, volume 2, pages 133–136, 2003. 2, 6

[3] S. M. Anstis and D. M. Mackay. The perception of apparent movement [and discussion]. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 290(1038):153–168, 1980. 3

[4] F. Attneave. Informational aspects of visual perception. *Psychological Review*, 61:183–193, 1954. 1

[5] H. Barlow. Cerebral cortex as a model builder. In *Models of the Visual Cortex*, pages 37–46, 1985. 1

[6] H. Barlow. Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12:241–253, 2001. 1

[7] J. Besag. Spatial interaction and the spatial analysis of lattice systems. *Journal of the Royal Statistical Society. Series B*, 36:192–236, 1974. 4

[8] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B*, 48:259–302, 1986. 4, 5

[9] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):185–207, 2013. 1, 2

[10] A. Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Trans. Image Process.*, 22(1):55–69, 2013. 2, 6

[11] N. Bruce and J. Tsotsos. Saliency based on information maximization. *Advances in Neural Information Processing Systems*, 18:155, 2006. 1, 3

[12] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006. 1

[13] B. Efron and R. Tibshirani. *An introduction to the bootstrap*, volume 57. CRC press, 1993. 4

[14] Y. Fang, W. Lin, Z. Chen, C. M. Tsai, and C. W. Lin. A video saliency detection model in compressed domain. *IEEE Trans. Circuits Syst. Video Technol.*, 24(1):27–38, 2014. 2, 6

[15] D. Gao and N. Vasconcelos. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, 8, 7:1–18, 2008. 1

[16] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil. Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, 30(1):51–64, 2012. 2, 3, 6

[17] S. Geman and D. Geman. Stochastic relaxation, gibbs distribution and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, 1984. 5

[18] H. Hadizadeh, M. J. Enriquez, and I. V. Bajić. Eye-tracking database for a set of standard video sequences. *IEEE Trans. Image Process.*, 21(2):898–903, Feb. 2012. 4, 6

[19] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. *Advances in neural information processing systems*, 19:545–552, 2007. 6

[20] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282–4286, 1995. 1

[21] Y. Hochberg and A. C. Tamhane. *Multiple comparison procedures*. John Wiley & Sons, Inc., 1987. 7

[22] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *Proc. IEEE CVPR'07*, pages 1–8, 2007. 2, 3

[23] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. *Advances in Neural Information Processing Systems*, 21:681–688, 2008. 2, 3

[24] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Trans. Image Process.*, 13(10):1304–1318, 2004. 6

[25] L. Itti and P. F. Baldi. Bayesian surprise attracts human attention. *Advances in Neural Information Processing Systems*, 19:547–554, 2006. 1, 3, 6

[26] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998. 1, 3, 4, 6

[27] W. Kim, C. Jung, and C. Kim. Spatiotemporal saliency detection and its applications in static and dynamic scenes. *IEEE Trans. Circuits Syst. Video Technol.*, 21(4):446–456, 2011. 6

[28] D. Knill and W. Richards. *Perception as Bayesian Inference*. Cambridge Univ. Press, 1996. 1

[29] E. Kreyszig. *Introductory mathematical statistics: principles and methods*. Wiley New York, 1970. 4

[30] S. Z. Li. *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009. 4

[31] X. Li, H. Lu, L. Zhang, X. Ruan, and M. H. Yang. Saliency detection via dense and sparse reconstruction. In *Proc. IEEE ICCV'13*, pages 2976–2983, 2013. 2, 3

[32] Z. Liu, H. Yan, L. Shen, Y. Wang, and Z. Zhang. A motion attention model based rate control algorithm for H. 264/AVC. In *The 8th IEEE/ACIS International Conference on Computer and Information Science (ICIS'09)*, pages 568–573, 2009. 2, 6

[33] Y. F. Ma and H. J. Zhang. A new perceived motion based shot content representation. In *Proc. IEEE ICIP'01*, volume 3, pages 426–429, 2001. 2, 6

[34] Y. F. Ma and H. J. Zhang. A model of motion attention for video skimming. In *Proc. IEEE ICIP'02*, volume 1, pages 129–132, 2002. 6

[35] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *Proc. IEEE CVPR'10*, pages 1975–1981, 2010. 2

[36] S. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 11:674–693, July 1989. 1

[37] R. Margolin, A. Tal, and L. Zelnik-Manor. What makes a patch distinct? In *Proc. IEEE CVPR'13*, pages 1139–1146, 2013. 2, 3

[38] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, 3(1):5–24, 2011. 4, 6

[39] B. Moulden, J. Renshaw, and G. Mather. Two channels for flicker in the human visual system. *Perception*, 13(4):387–400, 1984. 3

[40] K. Muthuswamy and D. Rajan. Salient motion detection in compressed domain. *IEEE Signal Process. Lett.*, 20(10):996–999, Oct. 2013. 2, 6

[41] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006. 1

[42] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996. 1

[43] D. J. Parkhurst and E. Niebur. Scene content selected by active vision. *Spatial Vision*, 16(2):125–154, 2003. 6

[44] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397–2416, 2005. 6

[45] P. Reinagel and A. M. Zador. Natural scene statistics at the center of gaze. *Network: Computation in Neural Systems*, 10:1–10, 1999. 4

[46] N. Sebe and M. S. Lew. Comparing salient point detectors. *Pattern Recognition Letters*, 24(1):89–96, 2003. 2

[47] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):15, 2009. 1, 6

[48] A. Sinha, G. Agarwal, and A. Anbu. Region-of-interest based compressed domain video transcoding scheme. In *Proc. IEEE ICASSP'04*, volume 3, pages 161–164, 2004. 6

[49] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S. C. Zhu. On advances in statistical modeling of natural images. *Journal of mathematical imaging and vision*, 18(1):17–33, 2003. 1

[50] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. IEEE CVPR'99*, pages 246–252, 1999. 2

[51] G. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand. Overview of the high efficiency video coding (HEVC) standard. *IEEE Trans. Circuits Syst. Video Technol.*, 22(12):1649–1668, 2012. 3

[52] J. A. Swets. *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers.* Lawrence Erlbaum Associates, Inc., 1996. 6

[53] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for Markov random fields. In *Proc. ECCV'06*, volume 2, pages 16–29, 2006. 5

[54] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist. Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, 45(5):643–659, 2005. 6

[55] C. Wang, N. Komodakis, and N. Paragios. Markov random field modeling, inference & learning in computer vision & image understanding: A survey. *Computer Vision and Image Understanding*, 117(11):1610–1627, 2013. 4

[56] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra. Overview of the H. 264/AVC video coding standard. *IEEE Trans. Circuits Syst. Video Technol.*, 13(7):560–576, 2003. 3

[57] 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*, 8(7), 2008. 1