mathematics

MDPI

Article

Saliency Map Estimation Using a Pixel-Pairwise-Based Unsupervised Markov Random Field Model

Max Mignotte

Vision Laboratory, Département d'Informatique et de Recherche Opérationnelle (DIRO), Faculté des Arts et des Sciences, Université de Montréal, Montréal, QC H3C 3J7, Canada; mignotte@iro.umontreal.ca

Abstract: This work presents a Bayesian statistical approach to the saliency map estimation problem. More specifically, we formalize the saliency map estimation issue in the fully automatic Markovian framework. The major and original contribution of the proposed Bayesian–Markov model resides in the exploitation of a pixel pairwise modeling and a likelihood model based on a parametric mixture of two different class-conditional likelihood distributions whose parameters are adaptively and previously estimated for each image. This allows us to adapt our saliency estimation model to the specific characteristics of each image of the dataset and to provide a nearly parameter-free—hence dataset-independent—unsupervised saliency map estimation procedure. In our case, the parameters of the likelihood model are all estimated under the principles of the iterative conditional estimation framework. Once the estimation step is completed, the MPM (maximum posterior marginal) solution of the saliency map (which we show as particularly suitable for this type of estimation), is then estimated by a stochastic sampling scheme approximating the posterior distribution (whose parameters were previously estimated). This unsupervised data-driven Markovian framework overcomes the limitations of current ad hoc or supervised energy-based or Markovian models that often involve many parameters to adapt and that are finely tuned for each different benchmark database. Experimental results show that the proposed algorithm performs favorably against state-of-the-art methods and turns out to be particularly stable across a wide variety of benchmark datasets.

Keywords: statistical estimation; iterative conditional estimation (ICE); Markov random field (MRF); mode of posterior marginal (MPM); regions of interest; saliency map estimation; salient object detection; stochastic optimization; unsupervised Markovian segmentation; pixel pairwise modeling

MSC: 60J10

Citation: Mignotte, M. Saliency Map Estimation Using a Pixel-Pairwise-Based Unsupervised Markov Random Field Model. *Mathematics* **2023**, *11*, 986. https://doi.org/10.3390/math11040986

Academic Editors: Vitaly Schetinin, Livija Jakaite and Dayou Li

Received: 28 December 2022
Revised: 31 January 2023
Accepted: 6 February 2023
Published: 15 February 2023

1. Introduction

Saliency detection can be defined as the selection of visually interesting and important regions or objects in an image which catch immediate attention or naturally grab and hold the viewer's attention. In fact, this task tries to fundamentally simulate the early processing stage (namely, *eye fixations* and *selective processing*) of the human vision system (HVS), which has the natural and astonishing ability to quickly and accurately identify the most visually noticeable and informative foreground object, in a (possibly complex) scene, in order to then adaptively focus the attention, via the *visual attention mechanism*, on such perceived important regions. This allows humans (and some mammals) to efficiently and quickly analyze the scene with a minimal allocated processing (visual) resources.

Recent years have witnessed rapidly increasing interest in saliency detection since this task plays an important role in a variety of applications, especially as a first step of many graphics/vision applications for which it is necessary, before all, to tackle the problem of information overload. It includes content-based image or video retrieval, categorization, summarization, compression, browsing and/or resizing, automatic image

cropping, adaptive image display on small device, infrared small target detection [1], object cosegmentation or advertising design to name a few.

Since the pioneer work of Itti et al. [2], who was one of the first authors to propose a saliency-based computational model relying on contrast-based image local features combined and computed across different scales, a rich literature on image saliency analysis has then been proposed to date. Within that literature, a significant number of methods have exploited the same principle based on local contrast concepts relying on pixel/region difference (with possibly different features) in the vicinity and expressing that a salient region/object exhibits a significant contrast to its immediate surroundings [3–13]. Another approach relies on the assumption that the salient object is globally distinct, i.e., it possesses discriminative color distributions with respect to the rest of the image or color uniqueness in terms of global statistics [7,11,14–20].

Local-contrast-based methods tend to produce higher saliency values near edges instead of uniformly highlighting salient objects (whose inner region can be discarded) and are also degraded with the presence of high-frequency noise, while global-contrast-based methods cannot clearly distinguish among regions and are degraded with the presence of background with complex (or salient small-scale) patterns [21,22].

That is why some saliency detection (SD) models propose to mix local- and global- (and region) contrast-based features [23], to combine local contrast cues via a multilayer or multiscale approach [20,21], a tree structure [19] or to base their saliency map estimation on a multilevel segmentation [18]. Unlike those who use local- or global-contrast-based saliency features that depend on the statistics of the particular image being viewed, Zhang et al. derived a saliency measure from natural image statistics obtained in advance from a database of natural images [24].

Most of these above-mentioned saliency-based computational methods are ad hoc designed, partly because the overarching goal of these models (i.e., the criterion used to select the optimal solution or simply what it is designed to optimize) is not specified, and also because these techniques have many parameters which must either be finely tuned (hand-selected) or require a high degree of supervision and machine learning, which in turn makes them entirely dataset-dependent. In other words, there is no guarantee that these ad hoc algorithms can achieve similar performance on another dataset for which they have not been trained or tuned for [22,25].

In order to remedy these drawbacks and shortcomings, some saliency detection models have been guided by the Gestalt law [26], developed in the matrix decomposition model framework [27], incorporated into the framework of graph cuts [28] or conceptualized within the framework of quantum mechanics [22]. Some other models have been developed in the energy-based model framework [21,29] or equivalently in the regression framework [18] and thus consider the saliency estimation as a global optimization problem. Other saliency detection models have been designed in the Bayes statistical framework [30–32] or with the conditional random field (CRF)-based approach (which is often and simply built from an ad hoc conditional distribution with an associated graphical structure) [23,33–36] or finally described in the (supervised) Markov Random Field (MRF) theory [5,37–42].

More precisely, regarding the above-mentioned MRF models, most authors [5,38–41] formulate the saliency estimation as a random-walk problem using the equilibrium distribution [5,38–40,43–46] to simulate human attention or exploit this equilibrium distribution to weight the absorbed time, thereby suppressing the saliency of long-range background regions with a homogeneous appearance [41]. More generally, an absorbing Markov chain [47,48] possibly using different hierarchies of deep features extracted from fully convolutional networks [49] or guided by depth information [50] can also be used for this purpose. Finally, Han et al. proposed to formulate the saliency estimation problem as a maximum a posteriori (MAP) estimation in a more classical contextual classification problem [37]. More precisely, the saliency map was estimated in a two-step process. The first step allowed the authors to generate a rough saliency map obtained by Itti's algo-

rithm [2]. In the second step, only a few attention seeds were first selected, according to the previously estimated saliency map and were combined and integrated, with some low-level features, in an MRF model to sequentially grow the attention objects from these selected attention seeds. Nevertheless, since no distribution or mixture of distributions were actually estimated from the image, this approach was not, stricto sensu, a Markovian approach. In addition, the resulting final algorithm remained an adhoc combination of several techniques with two algorithmic postprocesses to refine the extracted results. Moreover, that method relied on many internal parameters which had to be finely hand-selected and thus suffered from severe parameter dependencies, since the model was not cast within an unsupervised Markovian framework, which would have made it possible to estimate, in a criterion sense, the important internal parameters of the model.

Contrary to [37], we herein consider an unsupervised Markovian approach based on a field of observation built from a modeling by a pair of pixels and encoding the nonlocal pairwise pixel interactions (NLPPI) existing in the image. Based on these NLPPIs, our likelihood model is given by a mixture of class-conditional likelihood distributions of pairwise features (for each pairwise pixels existing in the image), whose parameters are (adaptively) estimated for each image. This allows us to adapt our saliency estimation model to the specific characteristics of each image of the dataset and to provide a nearly parameter-free—hence dataset-independent—unsupervised saliency map estimation procedure. In our case, the likelihood model parameters are fully estimated under the principles of the ICE (iterative conditional estimation) framework [51,52] with an ML (maximum likelihood) estimator (obtained with an iterative procedure minimizing the mean square error). Finally, the MPM (maximum posterior marginal) estimation of the saliency map is then computed via a Markov chain Monte Carlo (MCMC) sampling method to approximate the posterior distribution with the previously estimated parameters. This Bayesian criterion [53] is specifically well suited to our problem since it provides a practical way to obtain either the binary saliency map or its *soft* probabilistic version with values varying from zero to one.

Let us underline that, up to now, relatively few research works have been proposed in vision and image processing about energy-based or MRF models encoding or modeling via a likelihood (energy) function or (mixture of) distributions, all (or a subset of) the NLPPIs existing in an image. We can nevertheless mention some research works using energy-based models for texture synthesis [54], image segmentation [55,56], three-band compression model (for the color visualization of hyperspectral images) in remote sensing [57,58], gait analysis [59], edge histogram specification [60], for the compression of high-dynamic-range (HDR) images [61], with MRF models in segmentation fusion [62] or recently, in multimodal (heterogeneous) change detection in remote sensing [63].

The rest of this paper is structured as follows: section 2 details the proposed automatic Markovian model for the saliency map estimation problem by first defining and combining the likelihood and prior density functions and then the proposed two-step procedure, namely, a parameter estimation step followed by a saliency map estimation step, based on the previously estimated parameters. Section 3 presents a set of experimental results and comparisons with existing saliency map estimation methods, as well as the evaluation of the robustness of the proposed Bayesian approach. Finally, Section 4 concludes the paper.

2. Unsupervised Markovian Model For The Saliency Map Estimation Problem

Herein, we formalize the saliency estimation issue in the fully automatic Bayesian framework. To this end, an efficient and reliable method is a two-step approach [64]. First of all, a parameter estimation step is carried out to deduce the parameters of the likelihood model (in the sense of ML). Based on the value of these parameters, a second step is then devoted to the estimation of the saliency measure map (with range values between 0 and 1) or the binary saliency map.

2.1. Observation Field

Let us introduce some notation which is used throughout the paper. We first commonly consider a couple of random fields $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$, with $\mathbf{Y} = \{Y_s, s \in S\}$ the input image (for which a saliency map must be estimated and assumed to be toroidal) with N pixels located on a lattice S of N sites (or pixels) s, and $\mathbf{X} = \{X_s, s \in S\}$, the pixelwise random vector defining the label field or the saliency map. Each Y_s takes its value in a color space and each X_s is defined either in the discrete binary *state space* $\Lambda = \{e_0 = nonsalient,\ e_1 = salient\}$ or (it will be explicit in the following) in the probabilities range from 0 to 1.

In addition to that, we also consider that the set of $y_{<s,t>}$ values are a realization of a random variable (rv) vector $\mathbf{Y}_{<s,t>} = \{Y_{<s,t>}, Y_{<s,u>}, \ldots, Y_{<u,v>}, \ldots\}$ gathering the $N(N-1)$ random variables associated to each site (or pixel) pair, that we herein call the random (pixel-pairwise) observation field and secondly that $\mathbf{X}_{<s,t>}$ is its corresponding random (pairwise) label field taking its value in the discrete *state space* $\Lambda_{<s,t>} = \{id, di\}$. The pixel-pairwise label *id* means that the pixel at location s an t must share the same (**id***entical*) class label in the final saliency map $\hat{x}$ to be estimated (leading either to the configuration $<x_s{=}salient, x_t{=}salient>$ or $<x_s = nonsalient, x_t{=}nonsalient>$). Conversely, $x_{<s,t>} = di$ means that we have a **di***fferent* configuration, i.e., either the configuration $<x_s{=}salient, x_t{=}nonsalient>$ or $<x_s{=}nonsalient, x_t{=}salient>$.

In our pixel-pairwise modeling approach, in order to keep a quasi-linear complexity with respect to the number of pixels in the image (and therefore reduce the computational complexity of our stochastic algorithm), we consider for each pixel s and centered around it, a subsample $\mathcal{G}_s$ of 8 pairs of pixels evenly located around an $N_w \times N_w$ square window (see Figure 1).

Furthermore, we consider two feature vectors at site s, namely, $V_s^{\circlearrowleft}$ and $V_s^{\blacktriangleright}$, encoding first the textural and structural information existing around each local squared region of size $N_T = 16 \times N_T = 16$ centered at s and the color information, via the feature vector $V_s^{\bullet}$, existing within a local squared region of size $N_c = 5 \times N_c = 5$.

More precisely, in our application, we first estimate the discrete cosine transform (DCT) of each (gray-scale version of the) local squared window $N_T \times N_T$, compute its module (i.e., its absolute value since the DCT is real) and then apply a half-circular or radial integration transform (RIT) (i.e., to get the mean of the absolute DCT coefficient values for each discrete radius using a bilinear interpolation) to estimate a spectral descriptor vector $V_s^{\circlearrowleft}$ of size $N_T/2$ (see Figure 1). From this DCT transform, we also compute an axial integration, for each of the five discretized possible orientations, as described in Figure 1 to get $V_s^{\blacktriangleright}$.

As this texture descriptor is obtained from the compressed domain, it has the ability to be both greatly reduced in size and also robust to noise (since several denoisers are built from a filtering in this DCT domain [65,66]). Moreover, this strategy also allows one to efficiently code a texture with the properties of translation and rotation invariance for $V_s^{\circlearrowleft}$ and scale invariance for $V_s^{\blacktriangleright}$.

Moreover, remember that the DCT has a better compressive power than the discrete Fourier transform (DFT) and also the ability to estimate a less biased spectrum than that obtained by a DFT (especially when it is calculated on small images). This particularity comes from the even or mirror symmetry properties of the DCT which thus avoid the creation of false spectral components (or artifacts) generated by edge effects induced by the intrinsic periodic property of the DFT. In addition, the DCT calculates a spectrum which is real, contrary to the DCT, which estimates a complex spectrum (as a result, this also makes the calculation of the DCT extremely fast). (To this end, we have exploited the fast and optimized 16×16 DCT function implemented in C code by T. Euro (i.e., DDCT16X16S) and extracted from the FFT2D package available online at the http address given in [67]. The size $N_T = 16 \times N_T = 16$ was herein chosen because $N_T = 16$ is, before all, a power of 2 allowing to compute the DCT with linear complexity. Let us note that a size $N_T = 8 \times N_T = 8$ would have given almost the same (but very slightly less good) result.)

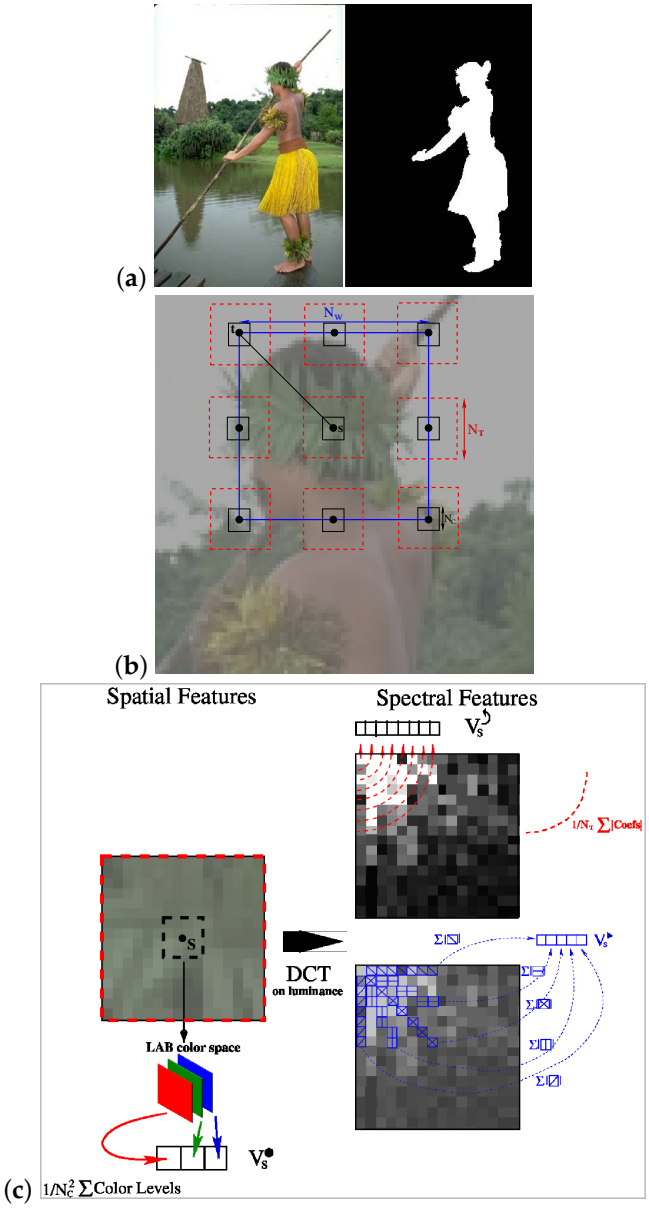

Figure 1. From top to bottom. (**a**) Input (and first) image of the DB-ECSSD database [21,68], assumed to be toroidal and showing a salient object and its ground truth. (**b**) Subsample $\mathcal{G}_s$ of 8 pairs of pixels $<s,t>$ associated to each pixel s (in which the pixel t is evenly located all around an $N_w \times N_w$ square window (with $N_w = 50$ in our case) with the two local square subwindows associated to each site s or t. (**c**) The spatial and spectral feature vectors $V_s^{\bullet}$, $V_s^{\circlearrowright}$ and $V_s^{\blacktriangleright}$ were built from the two local squared subwindows associated to the site s.

Finally, from the $N_c \times N_c$ local window, the color information around the pixel is taken into account, via the feature vector $V_s^{\bullet}$ of length 3 by specifying the mean L (luminance or lightness), A and B values contained in this subwindow (see Figure 1).

In our approach, we based our likelihood model on the assumption that the salient region or object was globally distinct, i.e., it possessed both discriminative colors and (to a lesser extent) discriminative textural properties with respect to the rest of the image. To this end, in our pixel-pairwise modeling using a subset of pixel pairs $<s,t>$ existing in S (see Figure 1), $y_{<s,t>}$ was computed from each considered pairwise feature vectors $(V^{\circlearrowright}, V^{\blacktriangleright}, V^{\bullet})$ extracted from the pixel pair $<s,t>$, with the following symmetric relation:

$$y_{<s,t>} = \underbrace{\left(|\, V_s^{\textcircled{\scriptsize D}} - V_t^{\textcircled{\scriptsize D}} |_1 + |\, V_s^{\blacktriangleright} - V_t^{\blacktriangleright} |_1 \right)}_{\text{Textural Features}} + \underbrace{\rho_c |\, V_s^{\bullet} - V_t^{\bullet} |_1}_{\text{Color Features}} \tag{1}$$

where $|.|_1$ is the L_1 norm and ρ_c is the weighting factor (see Section 3.1) between the color and spectral feature measures.

2.2. Likelihood Distributions

To use the observation measure $y_{<s,t>}$ (see Equation (1)) in a Bayesian settings, we must, before all, estimate the (marginal or conditional) likelihood distributions of $Y_{<s,t>}$, namely $P_{Y_{<s,t>}|X_{<s,t>}}$, in the two possible cases: identical pixel-pairwise label, $x_{<s,t>}=id$, or not, $x_{<s,t>}=di$.

2.2.1. Likelihood in the Identical Pixel-Pairwise Label Case

In our experiments, we noticed that if $x_{<s,t>}=id$, $P_{Y_{<s,t>}|X_{<s,t>}}$ was well approximated, for a given s, t, by an exponential distribution $p_{id} = \mathcal{E}(.;\lambda_{\text{id}})$ with shape (or inverse rate) parameter λ_{id}, i.e.,

$$\begin{aligned} P_{id}(y_{<s,t>}) &= P_{Y_{<s,t>}|X_{<s,t>}}(y_{<s,t>|x_{<s,t>}=id}) \\ &= \frac{\exp(-y_{<s,t>}/\lambda_{\text{id}})}{\lambda_{\text{id}}} \cdot H(y_{<s,t>}) \end{aligned} \tag{2}$$

with $\lambda_{id} > 0$ and the right-continuous Heaviside step function $H(x)$ where $H(0)=1$ (which makes the distribution supported on the interval $[0 \ \ \infty]$).

This approximation can be justified and understood if we notice that, for a site pair $<s,t>$, either fully included in a *salient* region or entirely within a *nonsalient* area (i.e., $x_{<s,t>}=id$), the computation of $y_{<s,t>}$ (see Equation (1)) is, in fact, a (weighted) sum of three n-order spatial gradient norms (n is the distance in pixel between s and t) in the textural sense (i.e., a difference, in the L_1 norm sense, between pairwise textural feature vectors). The gradient norm, meanwhile, built either from grayscale, color levels or from two texture feature vectors of an image, is known to be well approximated by a simple exponential distribution [63,69] or its possible variants (e.g., its generalized version [70], truncated variant [71,72] or finally a long-tail version with a shape and scale factor [60,61]).

2.2.2. Likelihood in the Different Pixel-Pairwise Label Case

In the case of $x_{<s,t>}=di$ (different pixel-pairwise labels), we empirically noticed that a normal distribution $P_{di} = \mathcal{N}(.;\mu_{\text{di}},\sigma_{\text{di}}^2)$ was well adapted to describe the measure $y_{<s,t>}$:

$$\begin{aligned} P_{di}(y_{<s,t>}) &= P_{Y_{<s,t>}|X_{<s,t>}}(y_{<s,t>|x_{<s,t>}=di}) \\ &= \frac{1}{\sqrt{2\pi\sigma_{\text{di}}^2}} \exp\left(-\frac{(y_{<s,t>} - \mu_{\text{di}})^2}{2\sigma_{\text{di}}^2} \right) \end{aligned} \tag{3}$$

Let us note that the Gaussian shape of this distribution is consistent with the central limit theorem saying that the Gaussian distribution is an attractor for the conditional random variable $Y_{<s,t>}|X_{<s,t>}$, whose realizations result from the sum of many i.i.d. (independent and identically distributed) random variables (in our case resulting from the difference of different spectral and spatial feature vectors for different (*salient* and *nonsalient*) texture feature vector pairs). Let us also add that the parameters of these two distributions, namely, $(\lambda_{\text{id}}, \mu_{\text{di}}, \sigma_{\text{di}})$ closely depends on the statistics of the salient or nonsalient region contained in each input image. The estimation of these parameters (see Section 2.4) is crucial in order to provide a nearly parameter-free—hence dataset independent—unsupervised saliency map estimation procedure.

2.3. Posterior Distribution

Let us now assume the independence of the pairwise data $Y_{<s,t>}$, conditionally on the pairwise labeling process $X_{<s,t>}$ relatively to the considered subsample $\mathcal{G}_s$ of pairs of pixels defined in Section 2.1 (see Figure 1), we have:

$$\mathrm{P}_{\mathbf{Y}_{<s,t>}|\mathbf{X}_{<s,t>}}(.) = \prod_{s\in S}\prod_{\substack{<s,t>\\ t\in\mathcal{G}_s}} \mathrm{P}_{Y_{<s,t>}|X_{<s,t>}}(y_{<s,t>}|x_{<s,t>}) \tag{4}$$

Moreover, if we assume that the distribution of $\mathbf{X}$ is Markovian and stationary and we specify a suitable prior distribution for the labeling process $\mathbf{X}$ and we agree that the saliency map x explicitely defines $x_{<s,t>}$, using likelihood (Section (2.2)), the joint distribution of $(\mathbf{X}, \mathbf{X}_{<s,t>}, \mathbf{Y}_{<s,t>})$ becomes:

$$\mathrm{P}_{\mathbf{X},\mathbf{X}_{<s,t>},\mathbf{Y}_{<s,t>}}(.) = \mathrm{P}_{\mathbf{X}}(.)\cdot\mathrm{P}_{\mathbf{Y}_{<s,t>}|\mathbf{X}_{<s,t>}}(.) \tag{5}$$

We obtain for the posterior distribution:

$$\mathrm{P}_{\mathbf{X},\mathbf{X}_{<s,t>}|\mathbf{Y}_{<s,t>}}(.) = \frac{\mathrm{P}_{\mathbf{X},\mathbf{Y}_{<s,t>}}(.)}{\mathrm{P}_{\mathbf{Y}_{<s,t>}}(.)} \propto \mathrm{P}_{\mathbf{X},\mathbf{Y}_{<s,t>}}(.) \tag{6}$$

We finally get:

$$\mathrm{P}_{\mathbf{X}|\mathbf{Y}_{<s,t>}}(.) \propto \prod_{s\in S}\prod_{\substack{<s,t>\\ t\in\mathcal{G}_s}} \mathrm{P}_{Y_{<s,t>}|X_{<s,t>}}(.)\cdot\mathrm{P}_{\mathbf{X}}(x) \tag{7}$$

We encoded a second-order isotropic Potts prior model related to the 8 closest neighbors of s, with equal potential value β for the various *cliques* $\langle s,v\rangle$ configurations (i.e., vertical, horizontal, left and right diagonal) of η_s, thus a model favoring homogeneous regions of the same class (*no-saliency* or *saliency* label) for $\hat{x}$, i.e., $\mathrm{P}_{\mathbf{X}}(x) \propto -\exp\{\sum_{\langle s,v\rangle\in\eta_s}[1-\delta(x_s,x_v)]\}$ [73], where $\hat{x}$, the saliency map to be computed, is related to the following corresponding posterior probability:

$$\begin{aligned}\mathrm{P}_{\mathbf{X}|\mathbf{Y}_{<s,t>}}(.) \quad \propto \quad & \prod_{s\in S}\Bigg\{\prod_{\substack{<s,t>\\ t\in\mathcal{G}_s}} \mathrm{P}_{Y_{<s,t>}|X_{<s,t>}}(.) \\ & \underbrace{\cdot\exp-\Big\{\frac{1}{|\mathcal{G}_s|}\cdot\sum_{\substack{\langle s,v\rangle\\ v\in\eta_s}}[1-\delta(x_s,x_v)]\Big\}\Bigg\}}_{P_{X_s}(x_s)}\end{aligned} \tag{8}$$

where δ is the delta Kronecker function and $|\mathcal{G}_s|$ is here the cardinality of the graph $\mathcal{G}_s$ (i.e., in our case $|\mathcal{G}_s| = 8$, see Figure 1).

2.4. Iterative Conditional Estimation

2.4.1. Principle

In our automatic Markovian segmentation model, we have first to estimate (*estimation step*) the parameter vector $\mathbf{\Theta}_{y_{<s,t>}}$ defining, respectively, the likelihood distributions $\mathrm{P}_{Y_{<s,t>}|X_{<s,t>}}(y_{<s,t>}|x_{<s,t>})$ for each of the two class labels $x_{<s,t>}$ of $y_{<s,t>}$ (see Equations (2) and (3)), i.e., the parameter vector $\mathbf{\Theta}_{y_{<s,t>}}(\lambda_{id}, \mu_{di}, \sigma_{di})$ encoding the shape parameter of the exponential distribution $P_{id}(y_{<s,t>})$ and the mean and variance parameters of the normal law $P_{di}(y_{<s,t>})$.

In our application, this estimation stage can be challenging for three main reasons; First of all, we find ourselves in the particular case of a mixture of different distributions (exponential and Gaussian). Second, these two distributions are also often strongly mixed

(see Figure 2). Finally, this mixture also has very different mixture proportions; generally, the *di* class is very underweighted because this class is related to the reduced number of labels or transitions per pixel pairs existing between the class *salient* and *non-salient* with respect to the considered graph (see Figure 1).

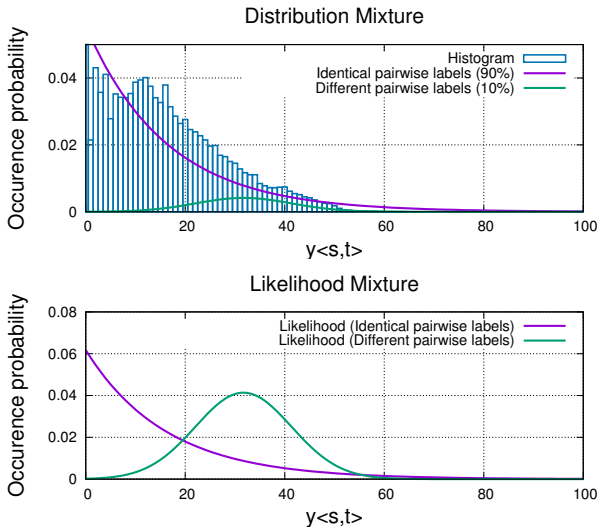

Figure 2. From top to bottom. Top: histogram of $y_{\langle s,t\rangle}$ obtained on the first frame of the DB-ECSSD database [21,68] (see Figure 1) and distribution mixture estimated by the ICE procedure (see Section 2.4) with an exponential distribution for the id_{entical} class label and a Gaussian distribution for the di_{fferent} pixel-pairwise label. Bottom: likelihood mixture composed of the two previous conditional likelihood distributions $P_{id}(y_{\langle s,t\rangle})$ and $P_{di}(y_{\langle s,t\rangle})$ (without weighting proportion) used in the likelihood part of the posterior density of our MRF SD model.

To this end, we resorted to the ICE [51,52,74] iterative procedure, which we used here in the particular case of our pixel-pair modeling and which was easily able to cope with different distributions and which experimentally turned out to be more efficient than the classical expectation maximization (EM) [75] algorithm or its stochastic version, the stochastic EM (SEM) [76]. This efficiency comes from the fact that the ICE estimation algorithm [51,52] can easily be understood as a direct improvement of the EM algorithm, and more precisely as both its stochastic and Markovian versions (since it is also constrained by the distribution of $\mathbf{X}$ assumed to be Markovian).

The ICE procedure first requires to find an estimator $\hat{\boldsymbol{\Theta}}_{y_{\langle s,t\rangle}} = \boldsymbol{\Theta}(x_{\langle s,t\rangle}, y_{\langle s,t\rangle})$ providing an estimate of $\boldsymbol{\Theta}_{y_{\langle s,t\rangle}}$ based on the complete data configuration $(x_{\langle s,t\rangle}, y_{\langle s,t\rangle})$. The random field $\mathbf{X}_{\langle s,t\rangle}$ being unobservable, the iterative ICE procedure thus defines the parameter $\boldsymbol{\Theta}^{[k+1]}_{y_{\langle s,t\rangle}}$, at step $[k+1]$, as the conditional expectations of $\hat{\boldsymbol{\Theta}}_{y_{\langle s,t\rangle}}$ given $\mathbf{Y}_{\langle s,t\rangle}=\{Y_{\langle s,t\rangle}\}$ and the current parameter $\boldsymbol{\Theta}^{[k]}_{y_{\langle s,t\rangle}}$.

The good behavior of this fixed point for the estimation of $\boldsymbol{\Theta}_{y_{\langle s,t\rangle}}$ in the sense of the mean squared error was demonstrated in the simple case [52] and in many past experiments.

By denoting $\mathbb{E}_k$ the conditional expectation based on $\boldsymbol{\Theta}^{[k]}_{y_{\langle s,t\rangle}}$, this estimation procedure is described as follows:

- We start from an initial value $\boldsymbol{\Theta}^{[0]}_{y_{\langle s,t\rangle}}$.
- $\boldsymbol{\Theta}^{[k+1]}_{y_{\langle s,t\rangle}}$ is computed from $\boldsymbol{\Theta}^{[k]}_{y_{\langle s,t\rangle}}$ and from $y_{\langle s,t\rangle}$ using:

$$\boldsymbol{\Theta}^{[k+1]}_{y_{\langle s,t\rangle}} = \mathbb{E}_k\left[\hat{\boldsymbol{\Theta}}_{y_{\langle s,t\rangle}}(x_{\langle s,t\rangle}, y_{\langle s,t\rangle}) \mid \mathbf{Y}_{\langle s,t\rangle}=\{Y_{\langle s,t\rangle}\}\right]$$

The computation of this expectation is impossible in practice, but we can approach it thanks to the law of large numbers [51]:

$$\mathbf{\Theta}_{y_{<s,t>}}^{[k+1]} = \frac{1}{n}\left[\hat{\mathbf{\Theta}}_{y_{<s,t>}}(x_{<s,t>}^{(1)}, y_{<s,t>}) + \cdots + \hat{\mathbf{\Theta}}_{y_{<s,t>}}(x_{<s,t>}^{(n)}, y)\right]$$

where $x_{<s,t>}^{(i)}$, $i = 1,\ldots,n$ are realizations drawn from the posterior $\mathbf{P}_{\mathbf{X}_{<s,t>}|\mathbf{Y}_{<s,t>},\mathbf{\Theta}}(x_{<s,t>}|y_{<s,t>}, \Theta_{y_{<s,t>}}^{[k]})$.

In our application, since x completely defines $x_{<s,t>}$ without ambiguity, these realizations can be drawn from the posterior distribution $\mathbf{P}_{\mathbf{X}|\mathbf{Y}_{<s,t>},\mathbf{\Theta}}(x|y_{<s,t>}, \Theta_{y_{<s,t>}}^{[k]})$ (see Section 2.3 and Equation (7)).

We noticed that $n = 1$ gave good results (while ultimately minimizing the computational cost of the iterative procedure) as was also found in several other experiments. (In fact, this is certainly due to the ergodic property of any image, which makes the ensemble average equivalent to a spatial average in the case of a random variable modeling the data or the (modified) color levels of an image [51]). From this observation, we therefore kept this value $n = 1$ in the rest of our experiments.

It was the case in our unsupervised Markovian saliency model, and we actually chose $n = 1$ in our experiments.

2.4.2. Ml Estimators for the ICE

For the Gaussian law, an ML estimate of $(\mu_{\text{di}}, \sigma_{\text{di}}^2)$, based on the complete data configuration, can be easily given by the empirical mean and variance statistics. For example, If $N_{di} \stackrel{\triangle}{=} \#\{x_{<s,t>} = di\}$, one gets:

$$\hat{\mu}_{\text{di}}(x_{<s,t>}, y_{<s,t>}) = \frac{\sum_{x_{<s,t>}=\text{di}} y_{<s,t>}}{N_{\text{di}}} \tag{9}$$

$$\hat{\sigma}_{\text{di}}^2(x_{<s,t>}, y_{<s,t>}) = \frac{\sum_{x_{<s,t>}=\text{di}} (y_{<s,t>} - \hat{\mu}_{\text{di}})^2}{(N_{\text{di}} - 1)} \tag{10}$$

For the exponential law, if $N_{id} \stackrel{\triangle}{=} \#\{x_{<s,t>} = id\}$, an ML estimate of the shape parameter is:

$$\hat{\lambda}_{\text{id}}(x_{<s,t>}, y_{<s,t>}) = \frac{\sum_{x_{<s,t>}=\text{id}} y_{<s,t>}}{N_{\text{id}}} \tag{11}$$

In our Bayesian SD framework, we do not need to estimate the proportion of each class. Nevertheless, the mixing proportion can be easily estimated within this procedure with the empirical frequency estimator $\pi_{\text{id}} = N_{\text{id}}/(N_{\text{id}} + N_{\text{di}})$ (and $\pi_{\text{di}} = N_{\text{di}}/(N_{\text{id}} + N_{\text{di}})$).

2.4.3. ICE Algorithm

$\mathbf{\Theta}_{y_{<s,t>}}(\lambda_{id}, \mu_{di}, \sigma_{di})$ is therefore estimated with the ICE algorithm as follows:

• *Parameter Initialization*:

We start with a randomly initialized saliency map x with two classes (*salient*/*nonsalient*) and from $\mathbf{\Theta}_{y_{<s,t>}}^{[0]} = (\lambda_{id}^{[0]}, \mu_{di}^{[0]}, \sigma_{di}^{[0]})$.

• **ICE procedure**: $\mathbf{\Theta}_{y_{<s,t>}}^{[k+1]}$ is then calculated from $\mathbf{\Theta}_{y_{<s,t>}}^{[k]}$ in the following way:

(1) *Stochastic step*: using the Gibbs sampler, one realization x of the saliency map is simulated according to the posterior distribution $\mathbf{P}_{\mathbf{X}/\mathbf{Y}_{<s,t>}}(x/y_{<s,t>})$, with parameter vector $\mathbf{\Theta}_{y_{<s,t>}}^{[k]}$.

More precisely, for each site s (lexicographically), we sample x_s with the local version of Equation (7), i.e.,

$$\mathrm{P}_{X_s|Y_{<s,t>}}(.) \propto \prod_{\substack{<s,t> \\ t \in \mathcal{G}_s}} \mathrm{P}_{Y_{<s,t>}|X_{<s,t>}}(.) \cdot \mathrm{P}_{X_s}(x_s) \tag{12}$$

with $P_{Y_{<s,t>}|X_{<s,t>}}$ an exponential law for $x_{<s,t>} = id$ with parameter λ_{id} (see Section 2.2.1);
with $P_{Y_{<s,t>}|X_{<s,t>}}$ a Gaussian law for $x_{<s,t>} = di$ with parameter (μ_{di}, σ_{di}) (see Section 2.2.2).

(2) *Estimation step*: the parameter vector $\boldsymbol{\Theta}_{y_{<s,t>}}^{[k+1]}$ is estimated with the ML estimator of the "complete data" (see Equations (9)–(11)).

(3) Repeat until a stopping criterion is met or until convergence is achieved, i.e., if $\boldsymbol{\Theta}_{y_{<s,t>}}^{[k+1]} \not\approx \boldsymbol{\Theta}_{y_{<s,t>}}^{[k]}$ (i.e., if the 1-norm of the difference between these two parameter vectors is below a threshold or after a maximum number of iterations), we return to the stochastic step.

In order to always ensure a good convergence of the ICE procedure, even in the presence of strongly mixed mixture distributions with unbalanced mixing proportions (as shown in Figure 1), we start this iterative procedure with $\boldsymbol{\Theta}_{<s,t>}^{[0]} = (\lambda_{\mathbf{id}}^{[0]}, \mu_{\mathrm{di}}^{[0]}, \sigma_{\mathrm{di}}^{[0]})$, with $\lambda_{\mathrm{id}}^{[0]} = \sigma_{\mathrm{di}}^{[0]} = 10$ and $\mu_{\mathrm{di}}^{[0]}$ as the average of the 10% highest values of $y_{<s,t>}$ in order to model the fact that the mean of the $(x_{<s,t>} =)\, di$ class is generally associated to the mean of the highest values of $y_{<s,t>}$. We finally use the *stochastic step* with a Gibbs sampler with a temperature (with a temperature factor T, we recall that the local posterior distribution is $\frac{1}{Z_s} \exp\left\{\frac{1}{T} \log P_{X_s|Y_{<s,t>}}(.)\right\}$) equal to 0.15 (empirically set after a couple of trials and errors) in order to allow a fast convergence and to reduce the number of explored solutions around the initialization values.

2.5. Saliency Map Generation Step

Once the estimation step is completed and based on the value of these crucial parameters, we now have to generate the saliency measure map with range values between 0 and 1 as efficiently as possible. To this end, the maximum a posteriori MAP [73] estimator, namely, the one that searches the configuration $\hat{x}$ such that $\hat{x} = \arg\max_{x \in \Omega} P_{\mathbf{X}|\mathbf{Y}_{<s,t>}}(.)$ (with the configuration set $\Omega = \Lambda^N$) would allow to find, in our application, the most probable, in the MAP sense, binary saliency map $\hat{x}$ given the image data. Equivalently, this strategy would search to find the global minimum of the negative log-posterior. Nevertheless, this strategy would allow us to finally estimate a binary saliency map and not a saliency measure map with range values between 0 and 1.

If a good initialization is not available, the ICM [73] algorithm will be ineffective and will be stuck in a local minima (i.e., give a poor suboptimal binary saliency map solution). The only way to avoid suboptimal local minima is to use a simulated annealing scheme, which is very computational demanding [63,77], especially for such an energy function to be optimized.

In our case, there is an interesting alternative which makes it possible to circumvent the global solution of the MAP estimator [78] associated with a binary or hard classification map (the MAP) and which is well suited to provide (rather quickly) either a binary saliency map or (as rather required here) a soft saliency measurement map; it is a more local Bayesian estimator that associates to each couple of sites s, t, the value of $X_{s,t}$ that is the most probable given the image data. It is referred to as the "marginal posterior mode" (MPM) estimator [53]. It relies, as the ICE procedure (see Section 2.4), on a sitewise posterior, i.e., the local version, for each site, of the posterior (Equation (12)). Mathematically, the MPM estimator is the one that searches the configuration $\hat{x}$ such that $\hat{x} = \arg\max_{x \in \Lambda} P_{X_s|Y_{<s,t>}}(.)$.

Algorithmically speaking, it consists in sampling N_{MPM} realizations of the random field X with the local version of the posterior or equivalently, this means repeating the stochastic step of the ICE procedure (see Section 2.4.3 and point 1) in order to simulate N_{MPM} samples of the binary saliency map. After these sampling steps, the proportion of *salient* labels for each pixel simply gives us the soft measure of salience for this pixel and thus the soft saliency map. Moreover, the median label for each site can give us the binary saliency map, in the MPM criterion sense.

2.5.1. Additional Location-Based Prior

In addition to the classical and standard isotropic Pott type prior $P_{X_s}(x_s)$ used in the posterior (see Equation (8)), which aims to favor homogeneous *salient* or *nonsalient* regions in each of the N_{MPM} binary saliency map samples generated by the MPM algorithm and which finally induces a regularizing effect on the soft saliency measure map, we added another prior. This additional prior (see Figure 3) was independent of the visual features and reflected prior knowledge of where the salient object/region was likely to appear, i.e., most likely in the center of the image [18,21,24,29]. In our case, this prior was modeled by the following additional prior distribution:

$$P_{X_{s:c}}(x_{s:c}) = \exp -\left\{ \delta(x_s, e_1) \cdot \left[-1 + 2 \max\left(\left| \frac{s_{\text{row}} - \text{hgth}/2}{\text{hgth}/2} \right|, \left| \frac{s_{\text{col}} - \text{wdth}/2}{\text{wdth}/2} \right| \right) \right] \right\} \tag{13}$$

where s_{row} and s_{col} designates, respectively, the row and column coordinates of pixel s. hgth and wdth are the height and width of the image and we recall that e_1 is the label associated to the class *salient*. With this additional prior, the prior part of the posterior becomes $\mathrm{P}_{\boldsymbol{X}}(x) = P_{X_s}(x_s) \cdot P_{X_{s:c}}(x_{s:c})$ (see Equations (7) and (8)).

2.5.2. Region-Based Constraint

We finally added a region constraint which aimed to better preserve the boundary of salient objects and regions. This kind of constraint exploits the concept of superpixel and has already been used successfully, in different ways, in several saliency SD models [79], especially in [18,19,29,30,79–81]. Superpixel algorithms produce an oversegmentation of an image in which each oversegmented region (called superpixel) brings together a group of pixels (forming a perceptually meaningful atomic region) that can be used to replace the rigid structure of the pixel grid in images. In our application, we resorted to the superpixel algorithm proposed by Felzenswalb [82] with the default values (settings) suggested by the authors. To this end, we first computed, before the MPM, a superpixel map, and at the end of each sampling (among the existing N_{MPM} samples) of the binary saliency map achieved by the MPM sampler, we simply counted the number of *salient* labels contained in each superpixel; if this one was less than half the number of pixels it contained, then we associated to all these pixels the *nonsalient* label.

We individually discuss and quantify the effectiveness of these two prior constraints in the following section.

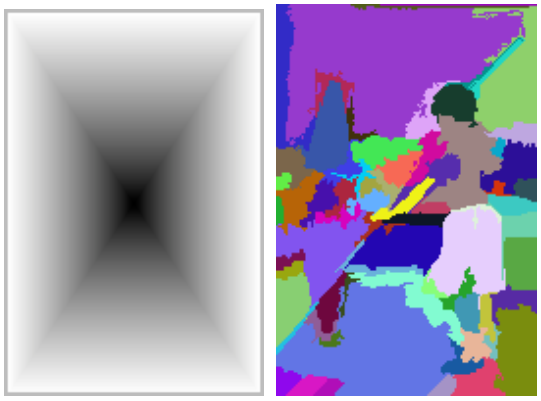

Figure 3. Visualization of the two prior constraints. Left: center prior with the term $[-1 + 2\max(.,.)]$ of Equation (13) (with convention -1: black and 1: white). Right: region prior given by a superpixel segmentation [82].

3. Experimental Results

3.1. Setup

First, we resized all images so that max(height,width) of the image was 200 pixels. In the following, all parameters were set with respect to this basic image size. Due to its strong correlation with the human visual perception, the perceptual CIE-Lab color space was herein used in this application. Experimentally speaking, it also turned out to be more effective in our application than the RGB color space.

The internal parameters of the model that remained to be set were the size of the squared window N_w (size of the graph $\mathcal{G}_s$), the size of the squared color window N_c (see Figure 1) and ρ_c, the weighting factor between the color and spectral feature measures (see Equation (1)).

In our experiment, we fine-tuned the value of these three internal parameters by trying to maximize the F measure on a subset of 10 randomized images from the DSCS data set by doing a local discrete grid search routine, with a fixed step size, on the parameter space and in the feasible ranges of parameter values (namely, $N_w \in$ [10–100] (step size: 10), $N_c \in$ [3–12] (step size: 2) and $\rho_c \in$ [1–10] (step size: 1). We found $N_w = 50$, $N_c = 5$ and $\rho_c = 6$. Let us note that the number of MPM iterations was not critical since in fact the higher, the better the results (but the higher the computational time). In our application, the number of MPM iteration was set to 300 in order to ensure a computational time around 4 seconds per image. Finally, the MPM sampler was initialized, at the start of the sampling process, not by a random label image but by an image containing class labels *change* in a central square with a surface area half that of the image on a background of class labels *no-change*. This strategy allowed us to improve the SD results compared to a conventional random initialization.

The efficiency of the MPM sampler, in term of F measure, is also quantified in Section 3.4 for a random initialization, for different number of iterations and for the different location priors considered in Sections 2.5.1 and 2.5.2.

3.2. Dataset Description

We validated our algorithm on the Extended Complex scene saliency dataset (ECSSD) built by Shi et al. [21,68]. This database is composed of 1000 images, including many semantically meaningful but structurally complex images for evaluation and containing various categories, such as natural objects (vegetables, flowers, mammals, insects and human) along with man-made objects (cups, vehicles, clocks, etc.). The backgrounds of many of these images are not uniform but possibly contains a small-scale structure or are composed of several parts, and some of the salient objects or regions do not have a sharp, clear boundary and/or an obvious difference with the background (which sometimes changes in illumination intensity) and are sometimes transparent. In addition, multiple salient objects possibly exist in one image, while a part of or all of them are regarded as salient as decided by an expert. A rigorous protocol was employed using several experts and several manually drawn saliency maps combined with each other in order to find the binary mask or ground truths that minimized the intersubject label consistency in the majority vote sense. More details about the database and the protocol used to construct the ground truths can be found in [68]. We also tested our method on the saliency dataset MSRA-5000 (or MSRA-5K) [23]. Images of MSRA-5000 are relatively more uniform and therefore the SD estimation is somewhat less difficult than on the ECSSD dataset [68].

3.3. Quantitative Measure

Our quantitative evaluations and experiment followed the setting described in [11,15,68,79]. We first plotted the precision–recall curve for the set of saliency measure maps obtained by our model and compared the obtained curve against to other methods (see Figure 4a). In addition, since in many applications, a high precision is required, we

thus also estimated the F_β-measure, as a function of each possible threshold (within the range $[0, 255]$), as proposed in [68,79]:

$$F_\beta = \frac{(1+\beta^2) \cdot \text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}} \tag{14}$$

in which thresholding was applied, and β^2 was set to 0.3 as suggested in [11,68,79] (the reason for weighting precision more than recall is that recall rate is not as important as precision [23,79] (since a 100% recall can be easily achieved by setting the whole region foreground)) and the obtained F_β-measure was plotted and compared to other methods (see Figure 4b).

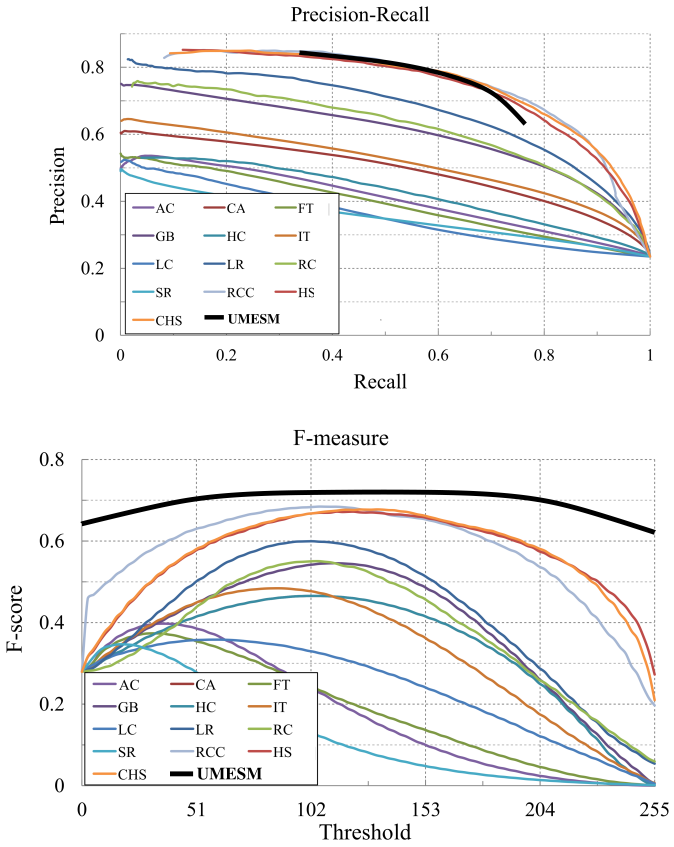

Figure 4. Quantitative comparison on ECSSD. From left to right: precision–recall curve for each possible threshold (within the range $[0, 255]$) and F_β-measure (see Equation (14) and [68]) with $\beta^2 = 0.3$ as a function of the threshold.

Moreover, we further investigated the performance of the *mean absolute error* (MAE) following [68,79], which measures the quality of the weighted continuous saliency map which may be of higher importance than the binary mask itself. More precisely, the MAE measures the mean absolute error between the soft saliency measure map x and the binary ground truth x_G, both normalized in the range $[0, 1]$. The MAE score is defined as follows:

$$\text{MAE} = \frac{1}{(\text{hgth} \times \text{wdth})} \sum_{i=1}^{\text{hgth}} \sum_{j=1}^{\text{wdth}} \left| x(i,j) - x_G(i,j) \right| \tag{15}$$

with i and j designating, respectively, the row and column coordinates of x or x_G.

3.4. Results and Comparisons

First, we tested (see Figure 5) the efficiency of the MPM-based SD estimator, in terms of precision–recall measures, for: 1. different number of iterations (namely, 2, 10, 100, 300 and 3000); 2. for a random initialization of the MPM (see Section 3.1); 3. by considering just the color features (see Equation (1)); 4. without considering the location-based prior (see Section 2.5.1); or finally, 5. without considering the region-based constraint (see Section 2.5.2). In all these cases, the other internal parameters and the proposed model were kept unchanged. From these tests, we can conclude, in term of F_β-measure, that the higher the number of iterations of the MPM, the better the results (but the higher the computation time too). The textural feature, the region-based constraint or a random initialization for the MPM slightly improved the results by only 2.1%, 1.2% and 1.7% (in term of F_β measure), respectively, whereas the location-based prior was important to achieve better SD results with the highest F_β-measure.

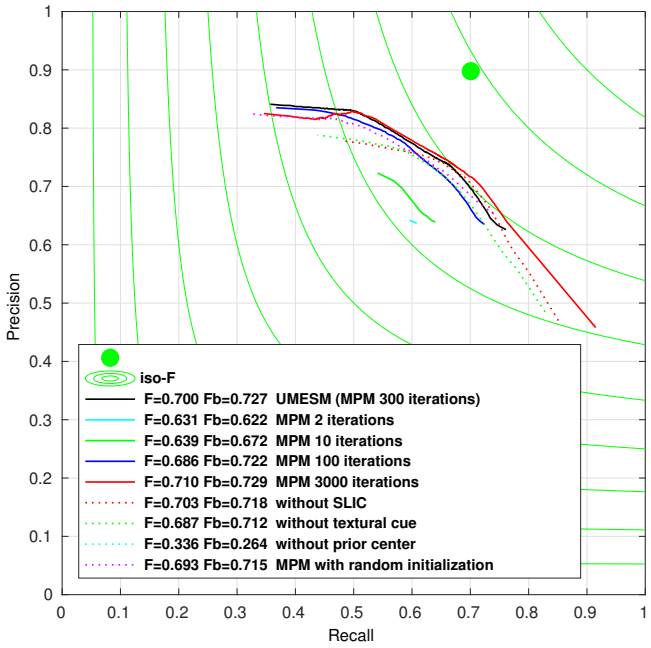

Figure 5. Precision–recall curves (and optimal F and F_β-measures (see Equation (14))) for different variations of our UMESM (unsupervised Markovian estimation of saliency map) model on the ECSSD [68] dataset.

Moreover, Figure 6 shows the distribution of the F_β, F and MAE measures given by our model (saliency map estimation with pixel-pairwise-based MRF) model on the 1000 images of the ECSSD dataset.

We evaluated our model on the ECSSD and MSRA-5000 datasets and compared our results in terms of the average MAE measure (see Equation (15)), with local schemes IT [3], GB [5], AC [9] and global schemes LC [14], FT [11], CA [16], HC [15], RC [15], RCC [20], LR [17], SR [7] and CHS [68] (see Table 1). Figure 4 shows the plot of the precision–recall curve and the F_β measure (see Equation (14)), as a function of the threshold, between these different SD methods comparatively to our model on the ECSSD dataset.

Table 1. Quantitative comparison for MAE on ECSSD [68] (first column) and MSRA-5000 (second column) datasets.

AC [9]	CA [16]	FT [11]	GB [5]	HC [15]	IT [3]	LC [14]	LR [17]	SR [7]	RC [15]	RCC [20]	HS [68]	CHS [68]	UMESM
0.264	0.310	0.270	0.282	0.326	0.290	0.294	0.267	0.264	0.301	0.187	0.224	0.227	**0.150**
0.228	0.250	0.230	0.243	0.239	0.248	0.245	0.215	0.225	0.264	0.140	0.153	0.150	**0.108**

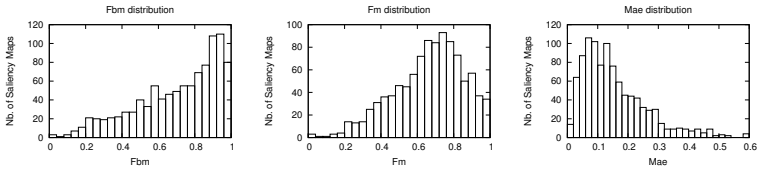

Figure 6. From left to right, distribution of the F_β, F and MAE measures given by our UMESM (unsupervised Markovian estimation of saliency map) model on the 1000 images of the ECSSD dataset.

3.5. Discussion

Our current implementation took on average 4.2 s to process one image with resolution 400×300 (the 1000 images in the ESCCD database were processed in 70 min) in the benchmark data on a 3.33 GHz Intel i7 CPU (6675.25 bogomips and 8 GB memory) with an unoptimized C++ code running on Linux. Up to 80% of this total time was dedicated to the stochastic saliency map generation step achieved by the MPM sampler (see Section 2.5), which could be easily parallelized using a GPU implementation as described in [83] in order to reduce the computation time by a factor greater than 100.

Firstly, we can notice that the precision–recall curve corresponding to our model was comparable to the best existing state-of-the-art CD algorithms (see Figure 4a) on the ECSSD dataset.

Secondly, we can also notice that our F_β-measure curve (Equation (14) with $\beta = 1$), as a function of the threshold, was the highest and flattest. This property allowed us to obtain the best F_β-measure ($F_\beta = 0.727$) but also to obtain, by very far, the best MAE score, i.e., the lowest mean absolute error between the continuous saliency map and the binary ground truth (both normalized in the range $[0, 1]$) compared to all other proposed SD methods (see Table 1). Indeed, the MAE score was proportional to the area under the plotted F_β-measure curve (as a function of the threshold) and to the line of equation $y = F_\beta$-measure $= 1$ in Figure 4b.

It is also worth mentioning that our F_β curve had a very different shape from all the others (very flat even at the ends) which showed us (in addition to showing that the F_β measure of our method was better than the other existing state-of-the-art algorithms whatever the value of the threshold (a curve constantly above)), two important things. First, our proposed SD method was very different conceptually and in terms of modeling than the other proposed methods. Second, and more importantly, the fact that the F_β measure curve, as a function of this threshold, associated with our model, was also much flatter than the other curves, showed us that our model was less sensitive to the threshold necessary to convert the saliency probability map into a binary saliency map. Graphically speaking, this was indicated by lesser grayscale variations in the estimated soft saliency map (see Figure 7). Perhaps for this reason, a greater confidence could be placed in our SD estimation result, or conversely, there may be less ambiguity in our model for accurately detecting and locating salient areas compared to other methods. In addition, the F_β score was better in our case, whereas our precision–recall curve remained comparable to the three best existing state-of-the-art CD methods; this meant that according to (Equation (14), since the F_β score was favored by a better precision measure, our model allowed us to obtain the best precision measure.

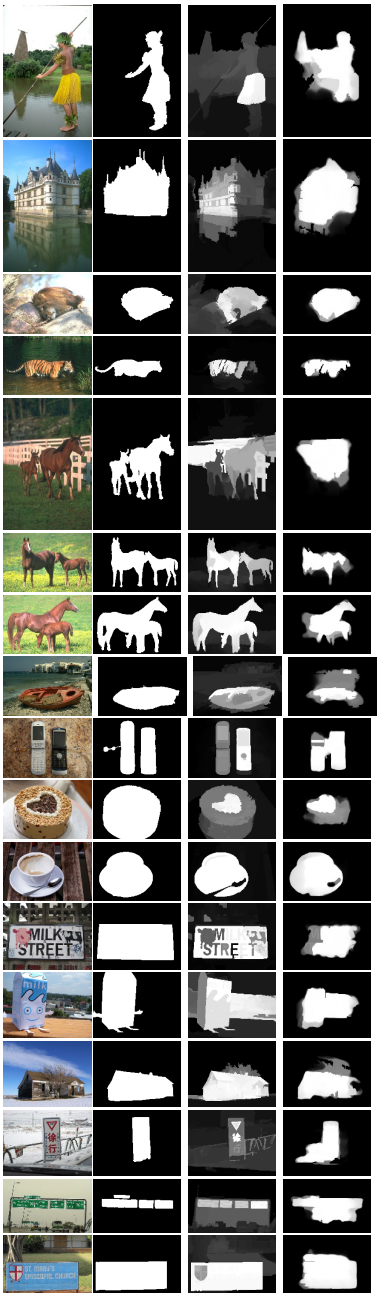

Figure 7. Visual comparison between our model and the state-of-the art CHS saliency model presented in [68] on the first 8 images and last 8 images of the ECSSD dataset. From left to right: ECSSD image, ground-truth salient binary mask, CHS saliency map [68] and our UMESM saliency result.

Concerning the MAE score, it is also worth recalling that this evaluation metric is actually quite different from the two F-based evaluation measures. Indeed, unlike the MAE score, the two F-based measures do not take into account the true negative saliency assignments [79], (i.e., the pixels correctly marked as *nonsalient*, which are, in fact, related to a quite large proportion of pixels in an image). A good MAE score combined with a good F-based measure rewards an SD method that successfully assigns saliency assignments to salient pixels and, unlike the methods with a lower MAE score, does not fail to correctly detect all the nonsalient regions existing in an image, as it should be.

For the MSRA-5000 dataset, the maximal F_β we obtained was $F_\beta = 0.790$ (for a threshold of 136) and $F = 0.757$ ($\beta = 1$, for a threshold of 51), which was a very competitive measure and also the best MAE measure (for the same reason above mentioned, because the F_β measure curve, as a function of the threshold, reached a maximum as high as the

best curves while remaining flatter than the other curves, thus inducing a lower surface area under the curve).

The estimation of each saliency map from the MSRA-5000 dataset depended on the image size but took, on average, 4.42 s using unoptimized C++ code running on a single core of an Intel i7 CPU with 3.33 GHz and 6675.25 bogomips.

In the future, we will study the fuzzy Markov chain modeling or the combination of a hidden Markov model and a fuzzy logic reasoning framework that will combine, on the one hand, the advantages of a Bayesian statistical analysis of the data with, on the other hand, the inaccuracies and uncertainties of the data on this difficult and imprecisely defined saliency map detection problem [84,85].

4. Conclusions

In this paper, we presented an unsupervised Markovian model for the saliency map estimation problem. This model was based on an original pixel-pair modeling and a likelihood model that used a parametric mixture of two different class-conditional likelihood distributions whose parameters were adaptively and previously estimated, according to a criterion that mixed maximum likelihood and least squares, for each image, with the iterative conditional estimation algorithm. Once the estimation step was completed, the MPM estimate solution of our model was particularly well-suited to our problem since it allowed us to obtain, by minimizing the Bayesian risk associated to the expected number of mis-(saliency) detection error, either a reliable binary saliency map or its (soft) probabilistic version. This unsupervised data-driven Markovian framework adapted our saliency estimation model to the specific characteristics of each image or/of each dataset in a widely unsupervised manner. Experimental results showed that the proposed algorithm performed very well against state-of-the-art methods for different and complementary measures of good saliency detection and was particularly stable across a wide variety of images. Moreover, the average F-measure curve, as a function of the threshold, associated with our model, also appeared much flatter than the curves associated to other state-of-the-art algorithms, which showed that our model was less sensitive to the threshold necessary to convert the saliency probability map into a binary saliency map comparatively to the other SD methods proposed in the literature. In addition, let us add that our model is perfectible by increasing the number of MPM iterations or by adding other informative prior knowledge (possibly via soft or hard constraints and may be expressed at different levels of abstraction such as pixel, segment, region, pairwise region, etc.), which can be easily integrated in the MPM random simulation scheme.

Funding: This research was funded by individual discovery grant RGPIN-2022-03654.

Data Availability Statement: The data that support the findings of this study are openly available at http://www.iro.umontreal.ca/~mignotte/ResearchMaterial/index.html (accessed on 31 January 2023).

Acknowledgments: The author would like to thank the NSERC (Natural Sciences and Engineering Research Council of Canada) for having supported this research work via the individual discovery grant program (RGPIN-2022-03654).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xia, C.; Li, X.; Zhao, L. Infrared Small Target Detection via Modified Random Walks. *Remote Sens.* **2018**, *10*, 2004. [CrossRef]
2. Itti, L.; Koch, C. Computational Modelling of Visual Attention. *Nat. Rev. Neurosci.* **2001**, *2*, 194–203. [CrossRef] [PubMed]
3. Itti, L.; Koch, C.; Niebur, E. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1254–1259. [CrossRef]
4. Ma, Y.F.; Zhang, H.J. Contrast-based Image Attention Analysis by Using Fuzzy Growing. In Proceedings of the Eleventh ACM International Conference on Multimedia, MULTIMEDIA'03, Berkeley CA, USA, 2–8 November 2003; ACM: New York, NY, USA, 2003; pp. 374–381.

5. Harel, J.; Koch, C.; Perona, P. Graph-Based Visual Saliency. In Proceedings of the 19th International Conference on Neural Information Processing Systems, NIPS'06, Vancouver, BC, Canada, 4–7 December 2006; MIT Press: Cambridge, MA, USA, 2006; pp. 545–552.
6. Harel, J.; Koch, C.; Perona, P. Graph-Based Visual Saliency. In *Advances in Neural Information Processing Systems 19*; Schölkopf, B., Platt, J.C., Hoffman, T., Eds.; MIT Press: Cambridge, MA, USA, 2007; pp. 545–552.
7. Hou, X.; Zhang, L. Saliency detection: A spectral residual approach. In Proceedings of the Computer Vision and Pattern Recognition, CVPR'07, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
8. Achanta, R.; Estrada, F.; Wils, P.; Susstrunk, S. Salient Region Detection and Segmentation. In *Computer Vision Systems*; Gasteratos, A., Vincze, M., Tsotsos, J.K., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 66–75.
9. Achanta, R.; Estrada, F.; Wils, P.; Susstrunk, S. Salient Region Detection and Segmentation. In Proceedings of the 6th International Conference on Computer Vision Systems, Santorini, Greece, 12–14 May 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 66–75.
10. Guo, C.; Ma, Q.; Zhang, L. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008.
11. Achanta, R.; Hemami, S.S.; Estrada, F.J.; Susstrunk, S. Frequency-tuned salient region detection. In Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Miami, FL, USA, 20–25 June 2009; pp. 1597–1604.
12. Klein, D.; Frintrop, S. Center-surround divergence of feature statistics for salient object detection. In Proceedings of the International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, 6–13 November 2011; pp. 2214–2219.
13. Zhang, J.; Ma, S.; Sameki, M.; Sclaroff, S.; Betke, M.; Lin, Z.; Shen, X.; Price, B.; Radomír, R. Salient Object Subitizing. *Int. J. Comput. Vision* **2017**, *124*, 169–186. [CrossRef]
14. Zhai, Y.; Shah, M. Visual Attention Detection in Video Sequences Using Spatiotemporal Cues. In Proceedings of the 14th ACM International Conference on Multimedia, Santa Barbara, CA, USA, 23–27 October 2006; ACM: New York, NY, USA, 2006; pp. 815–824.
15. Cheng, M.M.; Zhang, G.X.; Mitra, N.J.; Huang, X.; Hu, S.M. Global Contrast Based Salient Region Detection. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 20–25 June 2011; IEEE Computer Society: Washington, DC, USA, 2011; pp. 409–416.
16. Tal, A.; Zelnik-Manor, L.; Goferman, S. Context-Aware Saliency Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1915–1926.
17. Shen, X.; Wu, Y. A Unified Approach to Salient Object Detection via Low Rank Matrix Recovery. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; IEEE Computer Society: Washington, DC, USA, 2012; pp. 853–860.
18. Jiang, H.; Wang, J.; Yuan, Z.; Wu, Y.; Zheng, N.; Li, S. Salient Object Detection: A Discriminative Regional Feature Integration Approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013.
19. Liu, Z.; Zou, W.; Le Meur, O. Saliency Tree: A Novel Saliency Detection Framework. *IEEE Trans. Image Process.* **2014**, *23*, 1937–1952.
20. Cheng, M.; Mitra, N.J.; Huang, X.; Torr, P.H.S.; Hu, S. Global Contrast Based Salient Region Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 569–582. [CrossRef]
21. Yan, Q.; Xu, L.; Shi, J.; Jia, J. Hierarchical Saliency Detection. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; IEEE Computer Society: Washington, DC, USA, 2013; pp. 1155–1162.
22. Aytekin, C.; Kiranyaz, S.; Gabbouj, M. Automatic Object Segmentation by Quantum Cuts. In Proceedings of the 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweeden, 24–28 August 2014; pp. 112–117.
23. Liu, T.; Yuan, Z.; Sun, J.; Wang, J.; Zheng, N.; Tang, X.; Shum, H. Learning to Detect a Salient Object. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 353–367.
24. Zhang, L.; Tong, M.H.; Marks, T.K.; Shan, H.; Cottrell, G.W. Sun: A Bayesian framework for saliency using natural statistics. *J. Vis.* **2008**, *8*, 32. [CrossRef]
25. Huang, F.; Qi, J.; Lu, H.; Zhang, L.; Ruan, X. Salient Object Detection via Multiple Instance Learning. *IEEE Trans. Image Process.* **2017**, *26*, 1911–1922. [CrossRef]
26. Yan, Y.; Ren, J.; Sun, G.; Zhao, H.; Han, J.; Li, X.; Marshall, S.; Zhan, J. Unsupervised image saliency detection with Gestalt-laws guided optimization and visual attention based refinement. *Pattern Recognit.* **2018**, *79*, 65–78. [CrossRef]
27. Peng, H.; Li, B.; Ling, H.; Hu, W.; Xiong, W.; Maybank, S.J. Salient Object Detection via Structured Matrix Decomposition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 818–832. [CrossRef]
28. Ye, L.; Liu, Z.; Li, L.; Shen, L.; Bai, C.; Wang, Y. Salient Object Segmentation via Effective Integration of Saliency and Objectness. *IEEE Trans. Multimed.* **2017**, *19*, 1742–1756. [CrossRef]
29. Zhu, W.; Liang, S.; Wei, Y.; Sun, J. Saliency Optimization from Robust Background Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
30. Li, X.; Lu, H.; Zhang, L.; Ruan, X.; Yang, M.H. Saliency Detection via Dense and Sparse Reconstruction. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013.
31. Xie, Y.; Lu, H.; Yang, M. Bayesian Saliency via Low and Mid Level Cues. *Trans. Image Process.* **2013**, *22*, 1689–1698.

32. Feng, L.; Li, H.; Cheng, D.; Zhang, W.; Xiao, C. An Improved Saliency Detection Algorithm Based on Edge Boxes and Bayesian Model. *Trait. Signal* **2022**, *39*, 59–70. [CrossRef]
33. Mai, L.; Niu, Y.; Liu, F. Saliency Aggregation: A Data-Driven Approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013.
34. Rahtu, E.; Kannala, J.; Salo, M.; Heikkilä, J. Segmenting Salient Objects from Images and Videos. In Proceedings of the Computer Vision—ECCV 2010, Heraklion, Greece, 5–11 September 2010; Daniilidis, K., Maragos, P., Paragios, N., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 366–379.
35. Fu, K.; Gu, I.Y.; Yang, J. Saliency Detection by Fully Learning a Continuous Conditional Random Field. *IEEE Trans. Multimed.* **2017**, *19*, 1531–1544. [CrossRef]
36. Qiu, W.; Gao, X.; Han, B. A superpixel-based CRF saliency detection approach. *Neurocomputing* **2017**, *244*, 19–32. [CrossRef]
37. Junwei, H.; Ngi, N.K.; Mingjing, L.; HongJiang, Z. Unsupervised extraction of visual attention objects in color images. *IEEE Trans. Circuits Syst. Video Techn.* **2006**, *16*, 141–145.
38. da Fontoura Costa, L. Visual Saliency and Attention as Random Walks on Complex Networks. *arXiv* **2006**, arXiv:0603025. [CrossRef]
39. Gopalakrishnan, V.; Hu, Y.; Rajan, D. Random walks on graphs for salient object detection in images. *IEEE Trans. Image Process.* **2010**, *19*, 3232–3242. [CrossRef]
40. Wang, W.; Wang, Y.; Huang, Q.; Gao, W. Measuring visual saliency by Site Entropy Rate. In Proceedings of the Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13–18 June 2010; pp. 2368–2375.
41. Jiang, B.; Zhang, L.; Lu, H.; Yang, C.; Yang, M.H. Saliency Detection via Absorbing Markov Chain. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013.
42. Yang, J.; Yang, M. Top-Down Visual Saliency via Joint CRF and Dictionary Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 576–588. [CrossRef] [PubMed]
43. Li, C.; Yuan, Y.; Cai, W.; Xia, Y.; Feng, F.D.D. Robust Saliency Detection via Regularized Random Walks Ranking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
44. Yuan, Y.; Li, C.; Kim, J.; Cai, W.; Feng, D.D. Reversion Correction and Regularized Random Walk Ranking for Saliency Detection. *IEEE Trans. Image Process.* **2018**, *27*, 1311–1322. [CrossRef] [PubMed]
45. Jiang, F.; Kong, B.; Li, J.; Dashtipour, K.; Gogate, M. Robust Visual Saliency Optimization Based on Bidirectional Markov Chains. *Cogn. Comput.* **2021**, *13*, 69–80. [CrossRef]
46. Singh, V.; Kumar, N. CoBRa: Convex hull based random walks for salient object detection. *Multimed. Tools Appl.* **2022**, *81*, 30283–30303. [CrossRef]
47. Tang, W.; Wang, Z.; Zhai, J.; Yang, Z. Salient Object Detection via Two-Stage Absorbing Markov Chain Based on Background and Foreground. *J. Vis. Commun. Image Represent.* **2019**, *71*, 102727. [CrossRef]
48. Pengfei, L.; Xiaosheng, Y.; Jianning, C.; Chengdong, W. Saliency Detection via Absorbing Markov Chain with Multi-level Cues. *Ieice Trans. Fundam. Electron. Commun. Comput. Sci.* **2021**, *105*, 1010–1014.
49. Zhang, L.; Ai, J.; Jiang, B.; Lu, H.; Li, X. Saliency Detection via Absorbing Markov Chain With Learnt Transition Probability. *IEEE Trans. Image Process.* **2018**, *27*, 987–998. [CrossRef]
50. Wu, J.; Han, G.; Liu, P.; Yang, H.; Luo, H.; Li, Q. Saliency Detection with Bilateral Absorbing Markov Chain Guided by Depth Information. *Sensors* **2021**, *21*, 838. [CrossRef]
51. Salzenstein, F.; Pieczynski, W. Parameter Estimation in Hidden Fuzzy Markov Random Fields and Image Segmentation. *Cvgip Graph. Model Image Process.* **1997**, *59*, 205–220. [CrossRef]
52. Pieczynski, W. Convergence of the Iterative Conditional Estimation and Application to Mixture Proportion Identification. In Proceedings of the IEEE/SP 14th Workshop on Statistical Signal Processing, Madison, WI, USA, 26–29 August 2007; pp. 49–53.
53. Marroquin, J.; Mitter, S.; Poggio, T. Probabilistic Solution of ill-Posed problems in Computation Vision. *J. Am. Stat. Assoc.* **1987**, *82*, 76–89. [CrossRef]
54. Gimel'farb, G.L. Texture modeling by multiple pairwise pixel interactions. *IEEE Trans. Pattern Anal. Mach. Intell.* **1996**, *18*, 1110–1114. [CrossRef]
55. Mignotte, M. MDS-based multiresolution nonlinear dimensionality reduction model for color image segmentation. *IEEE Trans. Neural Netw.* **2011**, *22*, 447–460. [CrossRef]
56. Mignotte, M. MDS-based segmentation model for the fusion of contour and texture cues in natural images. *Comput. Vis. Image Underst.* **2012**, *116*, 981–990. [CrossRef]
57. Mignotte, M. A multiresolution Markovian fusion model for the color visualization of hyperspectral images. *IEEE Trans. Geosci. Remote. Sens.* **2010**, *48*, 4236–4247. [CrossRef]
58. Mignotte, M. A bi-criteria optimization approach based dimensionality reduction model for the color display of hyperspectral images. *IEEE Trans. Geosci. Remote. Sens.* **2012**, *50*, 501–513. [CrossRef]
59. Moevus, A.; Mignotte, M.; de Guise, J.; Meunier, J. A perceptual map for gait symmetry quantification and pathology detection. *Biomed. Eng. OnLine (BMEO)* **2015**, *14*, 1–24. [CrossRef]
60. Mignotte, M. An energy based model for the image edge histogram specification problem. *IEEE Trans. Image Process.* **2012**, *21*, 379–386. [CrossRef]

61. Mignotte, M. Non-local pairwise energy based model for the HDR image compression problem. *J. Electron. Imaging* **2012**, *21* , 99. [CrossRef]
62. Mignotte, M. A Label Field Fusion Bayesian Model and Its Penalized Maximum Rand Estimator For Image Segmentation. *IEEE Trans. Image Process.* **2010**, *19*, 1610–1624. [CrossRef]
63. Touati, R.; Mignotte, M.; Dahmane, M. Multimodal change detection in remote sensing images using an unsupervised pixel pairwise-based Markov Random Field model. *IEEE Trans. Image Process.* **2019**, *29*, 757–767. [CrossRef]
64. Mignotte, M.; Collet, C.; Pérez, P.; Bouthemy, P. Sonar image segmentation using an unsupervised hierarchical MRF model. *IEEE Trans. Image Process.* **2000**, *9*, 1216–1231. [CrossRef] [PubMed]
65. Mignotte, M.; Meunier, J.; Soucy, J.P. DCT-based complexity regularization for EM tomographic reconstruction. *IEEE Trans. Biomed. Eng.* **2008**, *55*, 801–805. [CrossRef] [PubMed]
66. Mignotte, M. Fusion of regularization terms for image restoration. *J. Electron. Imaging* **2010**, *19*, 333004. [CrossRef]
67. Ooura, T. General Purpose FFT (Fast Fourier/Cosine/Sine Transform) Package. Available online: http://momonga.t.u-tokyo.ac.jp/~ooura/fft.html (accessed on 31 January 2023).
68. Shi, J.; Yan, Q.; Xu, L.; Jia, J. Hierarchical Image Saliency Detection on Extended CSSD. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 717–729. [CrossRef]
69. Pérez, P.; Blake, A.; Gangnet, M. JetStream: probabilistic contour extraction with particles. In Proceedings of the IEEE International Conference on Computer Vision, ICCV'01, Vancouver, BC, Canada, 7–14 July 2001.
70. Widynski, N.; Mignotte, M. A multiscale particle filter framework for contour detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1922–1935. [CrossRef]
71. Destrempes, F.; Mignotte, M. A statistical model for contours in images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 626–638. [CrossRef]
72. Destrempes, F.; Mignotte, M. Localization of shapes using statistical models and stochastic optimization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1603–1615. [CrossRef]
73. Besag, J. On the Statistical Analysis of Dirty Pictures. *J. R. Stat. Soc.* **1986**, *B-48*, 259–302.
74. Delignon, Y.; Marzouki, A.; Pieczynski, W. Estimation of Generalized Mixture and Its Application in Image Segmentation. *IEEE Trans. Image Process.* **1997**, *6*, 1364–137. [CrossRef]
75. Dempster, A.; Laird, N.; Rubin, D. Maximum likelihood from incomplete data via the *EM* algorithm. *R. Stat. Soc.* **1976**, *39*, 1–22.
76. Masson, P.; Pieczynski, W. SEM algorithm and unsupervised statistical segmentation of satellite images. *IEEE Trans. Geosci. Remote. Sens.* **1993**, *31*, 618–633. [CrossRef]
77. Geman, S.; Geman, D. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. Pattern Recognit.* **1984**, *6*, 721–741. [CrossRef]
78. Perez, P. *Markov Random Fields and Images*; CWI Quarterly: Amsterdam, The Netherlands, 1998; pp. 413–437.
79. Borji, A.; Cheng, M.; Jiang, H.; Li, J. Salient Object Detection: A Benchmark. *IEEE Trans. Image Process.* **2015**, *24*, 5706–5722. [CrossRef]
80. Wang, J.; Jiang, H.; Yuan, Z.; Cheng, M.M.; Hu, X.; Zheng, N. Salient Object Detection: A Discriminative Regional Feature Integration Approach. *Int. J. Comput. Vision* **2017**, *123*, 251–268. [CrossRef]
81. Liu, G.; Yang, J. Exploiting Color Volume and Color Difference for Salient Region Detection. *IEEE Trans. Image Process.* **2019**, *28*, 6–16. . [CrossRef]
82. Felzenszwalb, P.; Huttenlocher, D. Efficient Graph-Based Image Segmentation. *Int. J. Comput. Vision* **2004**, *59*, 167–181. [CrossRef]
83. Jodoin, P.M.; Mignotte, M. Markovian segmentation and parameter estimation on graphics hardware. *J. Electron. Imaging* **2006**, *15*, 033005. [CrossRef]
84. Mignotte, M.; Collet, C.; Pérez, P.; Bouthemy, P. Markov Random Field and fuzzy logic modeling in sonar imagery: application to the classification of underwater floor. *Comput. Vis. Image Underst.* **2000**, *79*, 4–24. [CrossRef]
85. Tang, Y.M.; Zhang, L.; Bao, G.; Ren, F.; Pedrycz, W. Symmetric implicational algorithm derived from intuitionistic fuzzy entropy. *Iran. J. Fuzzy Syst.* **2022**, *19*, 27–44.