A SCALABLE AND EFFICIENT METHOD FOR SALIENT REGION DETECTION USING SAMPLED TEMPLATE COLLATION

Andreas Holzbach, Gordon Cheng

Institute for Cognitive Systems, Technische Universität München, Germany.[1]

ABSTRACT

We propose a fast method for salient region detection which aims at providing a computationally efficient method for online image processing. It is scalable and can be adjusted on the run to adapt to different computational requirements, which makes it a perfect candidate for time crucial applications. In our approach, we apply a template sampling over the image and compare these templates with each other by calculating a dissimilarity score. Templates with a low overall response are therefore likely to be part of a salient region in the image. This conceptually easy method is simple to implement and still outperforms state-of-the-art salient region detection systems (Our model's AUC(ROC) Score 0.794 - AIM 0.772).

Index Terms— Salient region detection, Visual Attention, Computational Attention

1. INTRODUCTION

Salient region detection is a broadly investigated research area, because it is concerns a wide field of scientific disciplines. Life sciences like psychology [1] or neuroscience [2] are interested in analyzing and predicting why salient regions are attractive to the human brain and how the neural processing is involved in this decision [3]. Numerous computational models have been proposed trying to model visual attention and predicting which areas will be favoured over others. Those saliency estimators can loosely be separated into biologically based, computational, or a combination of both which builds the majority of models [4]. Over the last decades visual attention has vastly been applied in the field of computer vision, because it can help in various problems like visual tracking[5], image segmentation[6] or object recognition[7].

2. RELATED WORK

Many of the approaches to visual saliency are computational expensive and complex (see comparison in [8]), making those models less suitable in real-world scenarios. More recent systems, however, are more focused on compuational performance, like [9] or Cheng et al.'s work in [8]. Latter compare different models by their computation times for building a saliency map and propose a fast model of their own based on regional contrast. The most related model in regard to our template sampling approach is Erdem and Erdem's work in [10]. They compare covariances of non-overlapping neighboured image regions to compute the saliency map.

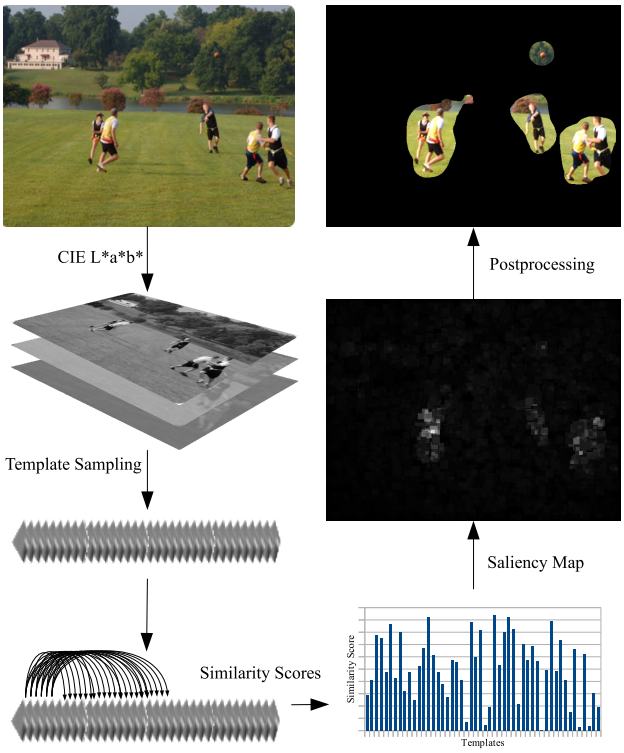


Fig. 1: Model Overview.

3. SAMPLED TEMPLATE COLLATION

Our model calculates the saliency map by sampling templates randomly over the image. Each template is then compared to the other templates by calculating a dissimilarity score. Higher scores mean lower similarity, lower responses higher similarity. Templates with a higher overall dissimilarity score therefore originate from areas in the image which stick out

[1] Find the author's e-mail addresses on www.ics.ei.tum.de. This work was supported by the DFG cluster CoTeSys and by BMBF through the Bernstein Center of Computational Neuroscience BCCN Munich.

from the rest. We consider these areas salient and use the templates' dissimilarity score to generate our saliency maps. See figure 1 for an overview of the model.

3.1. Sampling

First we sample templates from random positions on the image. For the evaluation we used templates of three different sizes (8,16,24). The different sizes account for the different dimensions a salient region might have. Using only one single template size however, doesn't affect the AUC(area under curve) of the receiver operator characteristics (ROC) score. We experienced only about 0.02% difference in the AUC score when using only one size. The number of sampled templates can be adjusted according to computational or accuracy requirements. Less templates can be calculated faster and are useful for generating single fixation points, more templates give a finer resolution and a more accurate and complete saliency map.

3.2. Collation Calculation

After the sampling process, each template T is compared with each other template of the same size. This leads to a complexity of $O(\frac{1}{2}n(n+1))$, as long as the used dissimilarity score is a commutative function, so that $f(T_1, T_2) = f(T_2, T_1)$. The complexity can be reduced by introducing a distance threshold (see 3.2.2). Different characteristics can be used to calculate the difference between the templates. For our evaluation we used *color space, distance and entropy*. The model can easily be extended to take different and more complex measures into account, like for example the correlation coefficient or a higher weight for templates which might contain faces using simple template matching.

3.2.1. Color Space

Although, different color spaces were tested, CIE Lab provided us with the most consistent results. We use a L_2 norm to calculate the difference of lightness L and color-opponent dimensions a and b between two templates T_1 and T_2.

$$l = ||T_{1_L} - T_{2_L}||_{L_2} = \sqrt{\sum (T_{1_L} - T_{2_L})^2} \tag{1}$$

$$a = ||T_{1_a} - T_{2_a}||_{L_2} = \sqrt{\sum (T_{1_a} - T_{2_a})^2} \tag{2}$$

$$b = ||T_{1_b} - T_{2_b}||_{L_2} = \sqrt{\sum (T_{1_b} - T_{2_b})^2} \tag{3}$$

3.2.2. Distance Weight

We include a distance weight to the dissimilarity score to account for local salient areas. Templates which are closer together have a higher weight than templates which are e.g. on the opposite side of the image. We compute the distance weight w by

$$w = 1 - \frac{d(T_1, T_2)}{max(d)} \tag{4}$$

with $d(T_1, T_2)$ being the pixel distance between template T_1 and T_2 and $max(d)$ being the maximum possible distance, which is the diagonal of the image. We set the distance weight to zero, if $d(T_1, T_2)$ is above a certain threshold (in our case half the maximum distance), this greatly improves computational performance while having no impact on the overall AUC(ROC) score. The complexity can now be approximated by assuming that we calculate the k nearest neighbours of each template which has complexity $O(n \log n)$ and the number of dissimilarity score calculations becomes $n * k$. The complexities combined are

$$\begin{aligned} &O(n \log n) + O(n * k) = O(n \log n) + O(n) \\ &= O(max(n \log n, n)). \end{aligned} \tag{5}$$

The inequation $n \log n > n$ is true for all $n > 2$. As the number of sampled templates will always be larger than 2 for a working system, we can say that the overall complexity is $O(n \log n)$.

3.2.3. Entropy

There exist numerous visual attention models which are built on information theoretic foundation to find the most salient areas [11, 12, 13]. We integrate the self-information of a template in our model by using:

$$H(X) = -\sum_{m=1}^{M} p_m \log p_m \tag{6}$$

with p_m being the relative frequency of brightness value m within the template. Using entropy we gain slightly better results (see table 2), as areas which would be salient because of their lightness and color uniqueness - e.g. a small area of a blue sky in the top of an image are not salient to a human subject.

We finally calculate the overall dissimilarity score s by calculating:

$$s = l(a + b) * w * H(T_1)H(T_2) \tag{7}$$

4. EVALUATION

The model was evaluated using two open accessible saliency benchmark databases and we measured the computational performance for online processing.

Table 1: Results of Judd's et al. saliency benchmark dataset.

Model	ROC	Similarity
GBVS[15]	0.801	0.472
Our Model /w CB	**0.794**	**0.477**
Multi-Resolution AIM[16]	0.772	0.471
Center Based	0.783	0.451
Our Model /wo CB	**0.687**	**0.357**
Torralba[17]	0.684	0.343
Itti & Koch[18]	0.562	0.284
Chance	0.503	0.327

4.1. Saliency Benchmarks

We tested our approach on Judd's et al. [14] saliency benchmark database[1]. The database contains 300 natural images with eye tracking data from 39 observers. Including a center bias our model performs significantly better than without a center bias (see table 1). The best results were achieved calculating the saliency map with $0.6 * centermap + 0.4 * our\ model$. The center map is a symmetric Gaussian streched to fit the apect ratio of the image. The factors were optimized using a different training set - see [14] for more details on center map and the optimization. Without a center bias our model still outperforms standard models like Itti & Koch (see table 1). See figure 2 for a comparison of images and saliency maps for several models.

We also tested our model with the ImgSal database[2] [19], which contains 235 color images, divided into six different categories ordered by their salient region size. We achieve an overall AUC(ROC) score of about 82% using a center bias, see table 2.

Our model outperforms state-of-the-art models like Multi-Resolution AIM [16] or long standing Itti et. al's [18]. There are models which outperform our system, like Judd et al. [14], which train a model using human fixation data, which incoporates the human's strong attention focus towards faces, persons or animals. We are not explicitely aiming at an adaption of the human fixation, but rather for generating a fixation point or region for salient areas. Our model however can be easily extended to bias templates with e.g. faces over templates with no faces using simple template matching.

4.2. Map Stability

We evaluated the effects of the sampling process on the stability of the saliency map and the salient point position. The more templates are sampled, the less the deviation between the maps and the higher the stability of a generated saliency map. We measure the deviation by calculating the L1-norm of two generated saliency maps of the same input image. The deviation of the most salient point is measured by the euclidean distance between the two points in the image. Both values are normalized, so that 100% deviation means the maximal possible deviation. From about 100 sampled templates, the deviation in the saliency map and the most salient points are constant with about $0.2 * 10^{-3}\%$ and 4% deviation, respectively.

[1] http://people.csail.mit.edu/tjudd/SaliencyBenchmark/

[2] http://www.cim.mcgill.ca/ lijian/database.htm

Table 2: Results of ImgSal saliency benchmark dataset with and without template entropy (TE); without center bias (CB) and without smoothing (Sm.).

	AUC Score (ROC)			
Category	/w TE	/wo TE	/wo CB	/wo Sm.
Large	0.819	0.818	0.793	0.707
Intermediate	0.813	0.799	0.780	0.670
Small	0.817	0.790	0.772	0.669
Cluttered	0.808	0.808	0.780	0.677
Repeating Distr.	0.848	0.844	0.816	0.715
Large & Small	0.826	0.809	0.798	0.706
Overall	**0.818**	**0.805**	**0.785**	**0.690**

4.3. Computational Performance

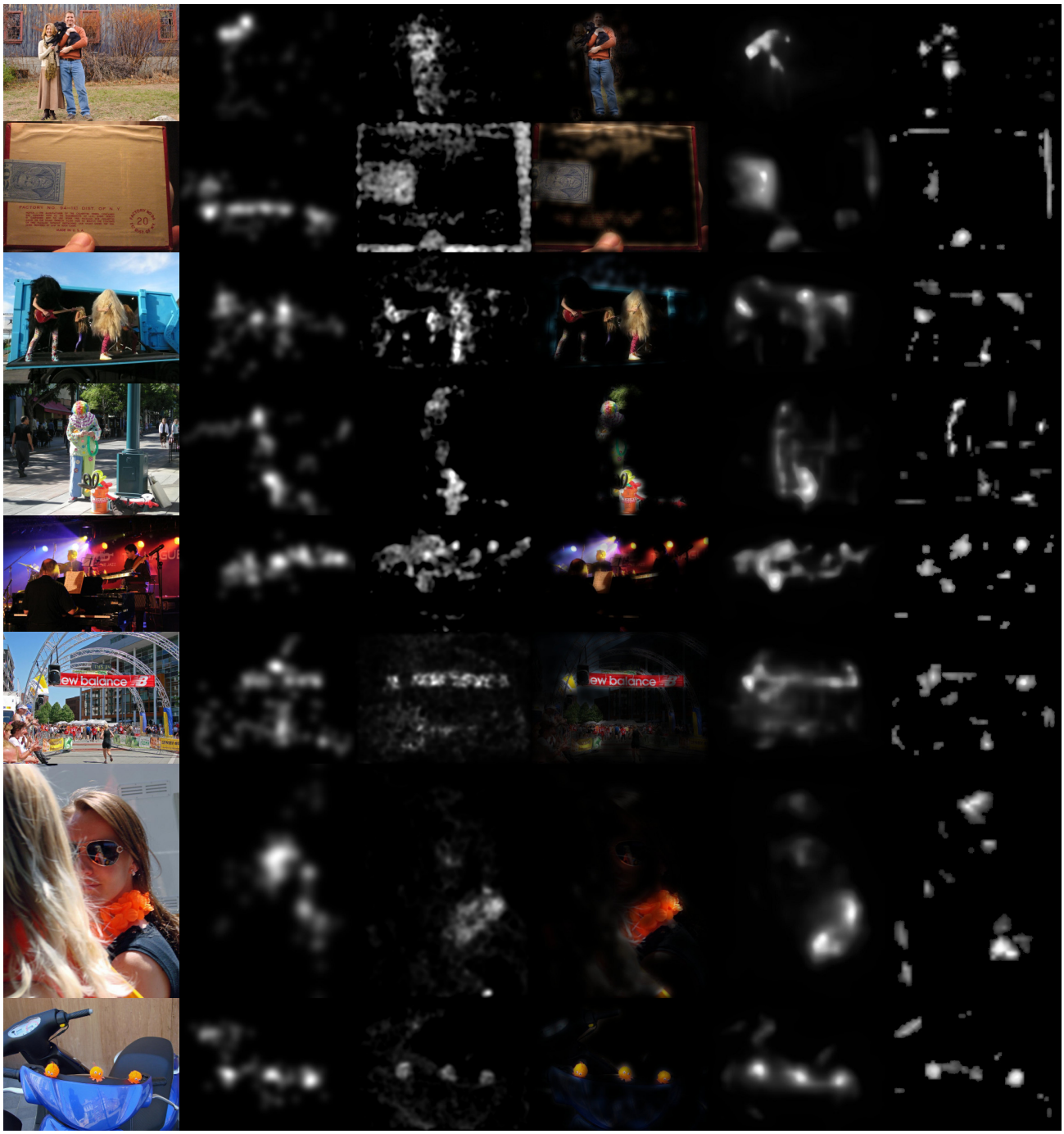
Our model's main aspect is the sampling process which has the major benefit, that it can be adjusted online. To estimate the computational speed of our performance we adaptively change the number of sampled templates to match a standard camera image frequency of about 30 fps at 640x480 pixels. If the processing is slower than 30 fps, less templates are sampled; if faster, more are sampled. This can of course be adjusted to personal requirements. We tested this setting on an intel i7 with 3.4 GHz and were able to sample about 130 templates using one core and about 440 templates using four cores for every camera frame captured at 30 Hz. For a video of the running system please have a look at our webpage[3]. It shows that salient areas are already detected using only the 130 templates.

We perform about as good as GBVS in regard to the ROC Score, but have a much lower complexity of $O(n \log n)$ compared to $O(n^4 K)$ (see [15] for details).

5. CONCLUSIONS

We presented a computationally efficient real-time capable method for salient region detection using sampled template collation. It is well suited for online processing, which is useful for generating a salient fixation point for preprocessing image data, but can also be applied to create whole saliency maps. We already showed good results in the saliency benchmarks but we believe that these results can still be improved by extending the templates dissimilarity measures.

[3] http://tinyurl.com/icip2014ICS



(a) Original Image (b) Human Fixation Map (c) Our Model (d) Modulated (e) GBVS[15] (f) Itti[18]

Fig. 2: Sample images and saliency maps for several models. The saliency maps shown of our model were created without center bias and were postprocessed using Gaussian Blur. The other images are taken from the MIT saliency benchmark database[14]

6. REFERENCES

[1] Koen Lamberts and Rob Goldstone, *Handbook of cognition*, Sage, 2004.

[2] Sabine Kastner Ungerleider and Leslie G, "Mechanisms of visual attention in the human cortex," *Annual review of neuroscience*, vol. 23, no. 1, pp. 315–341, 2000.

[3] Claus Bundesen, Thomas Habekost, and Søren Kyllingsbæk, "A neural theory of visual attention: bridging cognition and neurophysiology.," *Psychological review*, vol. 112, no. 2, pp. 291, 2005.

[4] Ali Borji and Laurent Itti, "State-of-the-art in visual attention modeling," 2013.

[5] Simone Frintrop, "General object tracking with a component-based target descriptor," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4531–4536.

[6] Ajay Mishra, Yiannis Aloimonos, and Cheong Loong Fah, "Active segmentation with fixation," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 468–475.

[7] Sunhyoung Han and Nuno Vasconcelos, "Biologically plausible saliency mechanisms improve feedforward object recognition," *Vision research*, vol. 50, no. 22, pp. 2295–2307, 2010.

[8] Ming-Ming Cheng, Guo-Xin Zhang, N. Mitra, Xiaolei Huang, and Shi-Min Hu, "Global contrast based salient region detection," in *IEEE CVPR*, 2011, pp. 409–416.

[9] Hongyu Li Lin Zhang, Zhongyi Gu, "Sdsp: A novel saliency detection method by combining simple priors," in *IEEE International Conference on Image Processing (ICIP 2013)*, 2013.

[10] Aykut Erdem Erkut Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *Journal of Vision*, vol. 13, no. 4, pp. 1–20, 2013.

[11] Neil Bruce and John Tsotsos, "Saliency based on information maximization," in *Advances in neural information processing systems*, 2005, pp. 155–162.

[12] Yuewei Lin, Bin Fang, and Yuanyan Tang, "A computational model for saliency maps by using local entropy.," in *AAAI*, 2010.

[13] Nadia Tamayo and V Javier Traver, "Entropy-based saliency computation in log-polar images.," in *VISAPP (1)*, 2008, pp. 501–506.

[14] Tilke Judd, Fredo Durand, and Antonio Torralba, "A Benchmark of Computational Models of Saliency to Predict Human Fixations A Benchmark of Computational Models of Saliency to Predict Human Fixations," 2012.

[15] Jonathan Harel, Christof Koch, and Pietro Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2006, pp. 545–552.

[16] Siddharth Advani, John Sustersic, Kevin Irick, and Vijaykrishnan Narayanan, "A multi-resolution saliency framework to drive foveation," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2596–2600.

[17] Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson, "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search.," *Psychological review*, vol. 113, no. 4, pp. 766, 2006.

[18] Laurent Itti and Christof Koch, "Computational modelling of visual attention," *Nature reviews neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[19] Jian Li, Martin D Levine, Xiangjing An, Xin Xu, and Hangen He, "Visual saliency based on scale-space analysis in the frequency domain," 2013.