Universitat Autònoma de Barcelona

Context, Motion and Semantic Information for Computational Saliency

Aymen Azaza

Universitat Autònoma
de Barcelona

Context, Motion and Semantic Information for Computational Saliency

A dissertation submitted by **Aymen Azaza** to the Universitat Autònoma de Barcelona in fulfilment of the degree of **Doctor of Philosophy** in **Informàtica**.

Bellaterra, 2018

Director	**Dr. Joost van de Weijer** Centre de Visió per Computador Universitat Autònoma de Barcelona
Director	**Dr. Ali Douik** NOCCS Laboratory, ENISO University of Sousse

Thesis committee	**Dr. Anis Sakly** National Engineering School of Monastir University of Monastir **Dr. Raimondo Schettini** DISCo (Department of Informatics, Systems and Communication) University of Milan-Bicocca **Dr. Xavier Giró** Image Processing Group Universitat Politecnica de Catalunya (UPC), Barcelona **Dr. Zied Lachiri** National Engineering School of Tunis University of Tunis **Dr. Ali Douik** NOCCS Laboratory, ENISO University of Sousse **Dr. Joost van de Weijer** Centre de Visió per Computador Universitat Autònoma de Barcelona

This document was typeset by the author using LaTeX 2ε.

The research described in this book was carried out at the Centre de Visió per Computador, Universitat Autònoma de Barcelona.

Dedicated to Jannet and Ahlem ...

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor *Dr. Joost van de Weijer* for the continuous support of my Ph.D study and research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. Without him this work could not have achieved its scientific level.

Also, I would like to thank also my advisor *DR. Ali Douik* for his support, advice and invaluable guidance since I worked with him during my bachelor study, master and my PhD, since 7 years ago.

I would like to thank the committee for giving me the opportunity to present the scientific outcome of the last four years of my research life as a member of the Computer Vision Center in Spain and The NOCCS Laboratory in Tunisia.

A very special gratitude is dedicated to all the members of the Learning and machine perception Team (*LAMP*) for their constructive suggestions and encouraging comments that considerably helped me to improve the quality of my research. I would like to thank every single one of them very much, Dankjewel *Joost*, Thanks *Andy*, mulțumesc *Bogdan*, moltíssimes gràcies *Marc*, *Laura*; and thanks *Lichao*, *Lu*, *Yaxing* and *Xialei*.

Many a thanks to all the CVC personnel for their kind support during my visits to the CVC. My stay in Barcelona was made amazing and exciting by numerous friends and colleagues who I would like to thank for their support. Many a thanks to Montse, Andy, Marc, Javad, Anguleos, Lu, Xialei, Lichao , Yaxing, Albert, Yash, Arash, Sounak, Fang(Jack), Claire. and other CVC friends. I would like to pay gratitude to my best friend Anguelos. Also thanks to Neji, Amine, Hachem, Hedi, Jihed, Daniel and Mohamed for their support and friendship.

Lastly, I would like to thank my family for their endless support and care. I would like to thank to the most important person in my life, my mother Jannet, for all her support and love. The immense support from my sisters Aida and Nadia and brothers Adel and Mohamed during these years means a lot to me. Thanks to all my friends who supported me during that time. I thank my fellow labmates in NOCCS Lab in Tunisia: Olfa, Mehrez and Rahma.

And finally, to the love of my life, *Ahlem*. I am very delighted that we are sharing this journey together.

Abstract

The main objective of this thesis is to highlight the *salient object* in an image or in a video sequence. We address three important — but in our opinion insufficiently investigated — aspects of saliency detection. Firstly, we start by extending previous research on saliency which explicitly models the information provided from the *context*. Then, we show the importance of explicit context modelling for saliency estimation. Several important works in saliency are based on the usage of object proposals. However, these methods focus on the saliency of the object proposal itself and ignore the context. To introduce context in such saliency approaches, we couple every object proposal with its direct context. This allows us to evaluate the importance of the immediate surround (context) for its saliency. We propose several saliency features which are computed from the context proposals including features based on omni-directional and horizontal context continuity. Secondly, we investigate the usage of top-down methods (*high-level semantic information*) for the task of saliency prediction since most computational methods are bottom-up or only include few semantic classes. We propose to consider a wider group of object classes. These objects represent important semantic information which we will exploit in our saliency prediction approach. Thirdly, we develop a method to detect *video saliency* by computing saliency from supervoxels and optical flow. In addition, we apply the context features developed in this thesis for video saliency detection. The method combines shape and motion features with our proposed context features. To summarize, we prove that extending object proposals with their direct context improves the task of saliency detection in both image and video data. Also the importance of the semantic information in saliency estimation is evaluated. Finally, we propose a new motion feature to detect saliency in video data. The three proposed novelties are evaluated on standard saliency benchmark datasets and are shown to improve with respect to state-of-the-art.

Key words: *visual saliency, computer vision, semantic segmentation, context proposals, object proposal, motion detection*

Resumen

El objetivo principal de esta tesis es resaltar el objeto más sobresaliente (salient) de una imagen o en una secuencia de video. Abordamos tres aspectos importantes — según nuestra opinión, no han sido suficientemente investigados — en la detección de saliencia. En primer lugar, comenzamos ampliando la investigación previa sobre saliency que modela explícitamente la información proporcionada desde el contexto. Luego, mostramos la importancia del modelado de contexto explícito para la estimación del saliency. Varios trabajos importantes en saliency se basan en el uso de "object proposal". Sin embargo, estos métodos se centran en el Saliency del "object proposal" e ignoran el contexto. Para introducir el contexto en tales enfoques de Saliency, unimos cada "object proposal" con su contexto directo. Esto nos permite evaluar la importancia del entorno inmediato (contexto) para calcular su Saliency. Proponemos varias características de Saliency, que se calculan a partir de los "object porposal", incluidas las funciones basadas en continuidad de contexto omnidireccional y horizontal. En segundo lugar, investigamos el uso de métodos top-down (información semántica de alto nivel) para la tarea de predicción de saliency, ya que la mayoría de los métodos computacionales son bottom-up o solo incluyen pocas clases semánticas. Proponemos considerar un grupo más amplio de clases de objetos. Estos objetos representan información semántica importante que explotaremos en nuestro enfoque de predicción de prominencias. En tercer lugar, desarrollamos un método para detectar la saliency de video mediante el cálculo de la saliencia de supervoxels y optical flow. Además, aplicamos las características de contexto desarrolladas en esta tesis para la detección de saliency en video. El método combina características de forma y movimiento con nuestras características de contexto. En resumen, demostramos que la extensión de "object proposal" con su contexto directo mejora la tarea de detección de saliency en datos de imágenes y video. También se evalúa la importancia de la información semántica en la estimación del saliency.

Finalmente, proponemos una nueva función de movimiento para detectar el salient en los datos de video. Las tres novedades propuestas se evalúan en conjuntos de datos de referencia de saliency estándar y se ha demostrado que mejoran con respecto al estado del arte.

Palabras clave: *percepción visual, visión por computador, segmentación semántica, propuestas de contexto, propuesta de objeto, detección de movimiento*

Résumé

L'objectif principal de cette thèse est de mettre en évidence *l'objet saillant* dans une image ou dans une séquence vidéo. Nous abordons trois aspects importants — mais à notre avis insuffisamment étudiés — pour la détection de saillance. Premièrement, nous commençons par étendre les recherches précédentes sur la saillance qui modélise explicitement les informations fournies par le *contexte*. Ensuite, nous montrons l'importance de la modélisation explicite de contexte pour l'estimation de la saillance visuelle. Plusieurs études de saillance sont basés sur l'utilisation de "object proposal". Cependant, ces méthodes se concentrent sur la saillance de l'"object proposal" elle-même et ignorent le contexte. Pour introduire le contexte dans de telles approches de saillance, nous couplons chaque "object proposal" avec son contexte direct. Cela nous permet d'évaluer l'importance de contour immédiat (contexte) pour la saillance. Nous proposons plusieurs caractéristiques saillantes qui sont calculées à partir de contexte, y compris les caractéristiques basées sur la continuité du contexte horizontal et omnidirectionnel. Deuxièmement, nous étudions l'utilisation de méthodes top-down (*informations sémantiques de haut niveau*) pour la tâche de prédiction de saillance puisque la plupart des méthodes de calcul sont bottom-up ou n'incluent que peu de classes sémantiques. Nous proposons de considérer un groupe plus large de classes d'objets. Ces objets représentent des informations sémantiques importantes que nous exploiterons dans notre approche de prédiction de saillance. Troisièmement, nous développons une méthode pour détecter la *saillance vidéo* en calculant la saillance des supervoxels et du flux optique. La méthode combine des caractéristiques de forme et de mouvement avec nos caractéristiques de contexte. Pour résumer, nous démontrons que l'extension des "object proposal" avec leur contexte direct améliore la tâche de détection de saillance dans les données image et vidéo. L'importance de l'information sémantique dans l'estimation de la saillance est également évaluée. Enfin, nous proposons une nouvelle caractéristique

de mouvement pour la détection de la saillance dans les données vidéo. Les trois nouveautés proposées sont évaluées sur des bases de données de saillance et montrent une amélioration par rapport à l'état de l'art.

Mot clés : *saillance visuelle, vision par ordinateur, segmentation sémantique, contexte 'proposal', 'objet proposal', détection de mouvement*

Contents

List of Figures

List of Tables

List of acronyms

GOP	Geodesic Object Proposals
MCG	Multiscale Combinatorial grouping
SS	Selective Search
CPMC	Constrained Parametric Min-Cuts
MDF	Multiscale Deep Features
DCL	Deep Contrast Learning
CNN	Convolutional Neural Networks
HVS	Human Visual System
CRF	Conditional random fields

1 Introduction

Images play an important role in our daily live. Many factors such as the social networks and smartphones, which become more integrated into our daily lives, make the usage of images more and more primordial. Social networks such as facebook, twitter, google or instagram use a great amount of digital images. Therefore visual communications make up an important part of our daily communications. The manual management of this large amount of pictures is quite difficult which increases the demand of application which can automatically understand these images with the aim to better manage them.

The aim of computer vision is to analyze and better understand the image content automatically. Recently, convolutional neural networks have revolutionized computer vision progress. The initial success of deep networks on image classification was followed by excellent results on other computer vision fields, showing time and again that the features learned with deep networks outperformed the existing hand-crafted features. The computer vision fields in which deep learning has been applied include object detection, semantic segmentation, image captioning, optical flow, and saliency detection. The later is the focus in this thesis.

Computational saliency plays an important role to automatically understand and analyze image content. It helps to highlight the most important parts or regions in images and videos. It has a wide range of applications such as image retargeting, image segmentation, image and video compression and automatic target detection. In this thesis, we aim to detect salient objects in images and videos using different approaches. First we extend object proposals with its direct context (context proposals) since the object is salient with respect to a background which is occluding by it, after that we exploit the recent advances in convolutional neural networks which were already shown important for object detection and semantic segmentation results. We develop a method to incorporate high-level image understanding

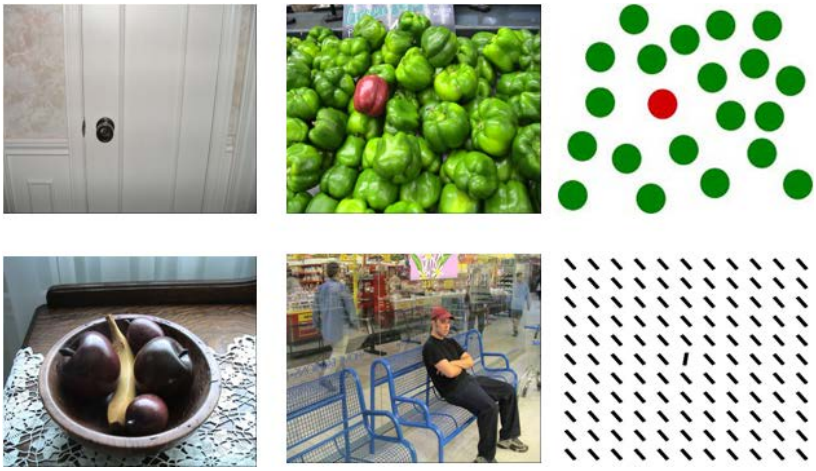


Figure 1.1 – Pop out effect, Example of images with one salient object.

for saliency detection using saliency features based on high-level semantic information including object detection and semantic segmentation. Finally, we compute saliency in video using motion features and the proposed context features.

1.1 Computational saliency

Visual saliency has been a fundamental problem in neuroscience, psychology, and computer vision for a long time. To rapidly extract important information from a scene, the human visual system allocates more attention to salient regions. Research on computational saliency focuses on designing algorithms which, similarly to human vision, predict which regions in a scene are salient. As definition, visual saliency is the perceptual quality that makes an object, person or pixel region stand out relative to their neighbors in order to grab our attention (see figure 1.1).

Computational saliency can be roughly divided in two main research branches. Firstly, it is originally defined as a task of predicting eye-fixations on images [123, 128]. Secondly, researchers also use the term to refer to salient object detection or salient region detection [52, 70]. Here the task

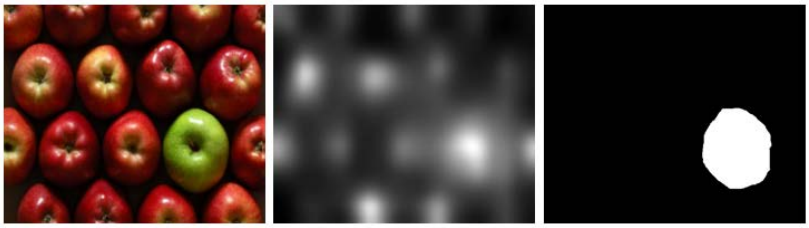


Figure 1.2 – Input image, eye fixation map and salient object ground truth.

is extended to identifying the region, containing the salient object, which is a binary segmentation task for salient object extraction. See for example figure1.2 where we show the outputs which are related with the two main approaches in computational saliency, namely an eye-fixation map and a salient object segmentation. We have identified three main research challenges in the field of computational saliency, which we will detail in this section.

1.1.1 Context for saliency

One of the main challenges of computational saliency is the question how to incorporate context information into the computational pipeline. As said above, saliency is defined as the property of objects to stand out with respect to their surrounding. Given this definition, it seems logical that context should play an explicit role in computational saliency models. However, in many of the computational saliency methods, research has focused on other aspects of saliency in images. For example in the study on a wide range of features important for computational saliency by Judd and Torralba [53] they studied the following features: intensity, orientation and color contrast, and distance to the center. They found that center prior was the most important saliency feature in many saliency datasets, which is probably caused by the choice of images in the dataset (people tend to put the object of interest in the middle of the picture). Surprisingly, context was not explicitly evaluated in this study. Many other important papers on computational saliency did not explicitly incorporate context [41, 45, 54].

Only several works have considered to explicitly model context. Jiang et

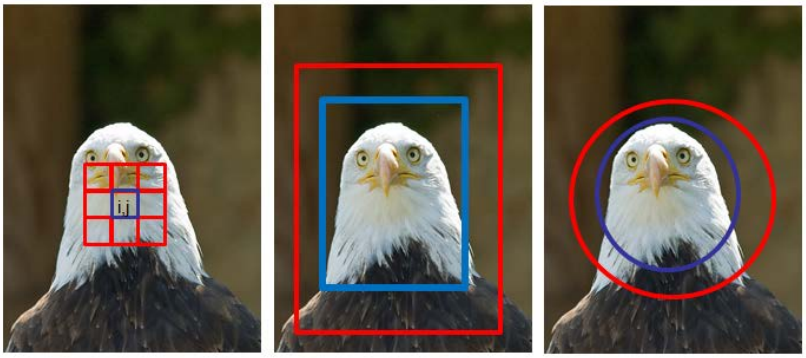


Figure 1.3 – Different types of context, left the local neighborhood of the pixel, middle rectangular surround and right center-surround structure. Example shows that the uniformity of the context is an important cue for saliency. It also shows that different context shapes could be considered for its computation.

al. [52] compute the rarity of a feature by comparing the contrast between a border around the image and the object proposal histogram. Goferman et al. [39] compute saliency using the difference of patches with all other patches in the image. Liu et al. [73] consider rectangular center-surround neighborhood in the images for saliency detection. In a recent work, Mairon and Ben Shahar [80] prove the importance of visual context in saliency prediction. Their work is based on the observation that an object is salient with respect to its surrounding. They show that excellent saliency results can be obtained by only considering the context of the salient object. A drawback of their method is that it is computationally very demanding which is due to the iterative nature of the computation. In conclusion, we believe that fast, scalable methods for incorporating context in computational saliency require further research. In Figure 1.3 we present an example which show the importance of context for saliency computation.

1.1.2 Top-down semantic saliency

The vast majority of research on computational saliency focuses on bottom-up saliency, where the saliency is computed from low-level features in images such as edges, colors, and luminance. However, from human vision research it is known that high-level information plays a crucial role in saliency assignment [20, 79]. For example, in the work of Cerf et al. [20] they evaluate the impact on eye movement of semantic classes such as faces and text. They found that under free-viewing conditions these regions were visited significantly more (over 10×) than other regions.

Based on these results, several papers have investigated the use of object detection to improve saliency detection. Judd et al. [53] propose a top-down algorithm to detect objects such as faces, people, text, body parts and animals. Einhauser et al. [26] prove that objects predict fixations better than early saliency, so they propose a model based on segmenting objects to predict salient regions. Yang et al. [131] propose a top-down approach based on learning a conditional random field and a dictionary. Given the recent improvement of object recognition due to the use of deep learning algorithms in the field, we think it would be interesting to evaluate these for the task of top-down saliency detection. An example of object detection and semantic segmentation results is shown in Figure 1.4.

1.1.3 Saliency in video

Itti et al [49] proposed one of the first computational saliency methods for images. Based on human vision research they proposed to use the features: luminance, orientation and color. Further human vision research has also shown that motion is an important cue for saliency [1, 132]. Abrams et al. [1] demonstrate the fact that motion is a strong cue in attracting attention. Yantis and Egeth [132] establish that motion can be selected by the attention mechanism.

When extending saliency research to the video data, a model of motion saliency needs to be proposed. Recently, several works have started investigating saliency in videos. Singh et al. [108] compute video saliency using color dissimilarity, motion difference, objectness measure, and boundary score feature. They use temporal superpixels to simulate attention to a set

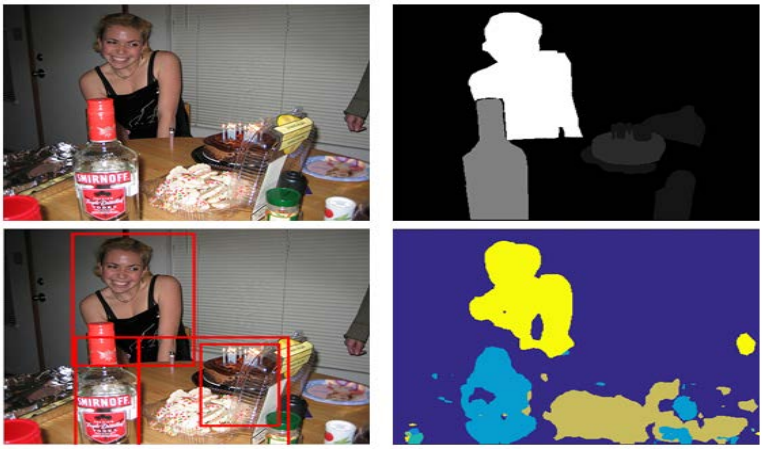


Figure 1.4 – Input image and salient object ground truth (top), we will exploit high-level information object detection (bottom left) and semantic segmentation (bottom right) in the task of saliency prediction.

of moving pixels. Lezama et al. [62] compute motion vectors from two consecutive frames. Then they group pixels that have coherent motion together. Wang et al. [124] propose a video saliency object segmentation approach based on geodesic distance where they use spatial edges and temporal motion boundaries as foreground indicators. Given the growing importance of video data in current society, we think that further research on saliency for video has many potential applications. We provide an example of video saliency in Figure 1.5.

1.2 Objectives and Approach

In this section we summarize the objectives of the thesis and outline the approach which we have used to address them.

1.2.1 Context proposals for saliency detection

In chapter 2, we introduce our saliency approach based on the usage of context proposals. Recent work in saliency detection literature shows that

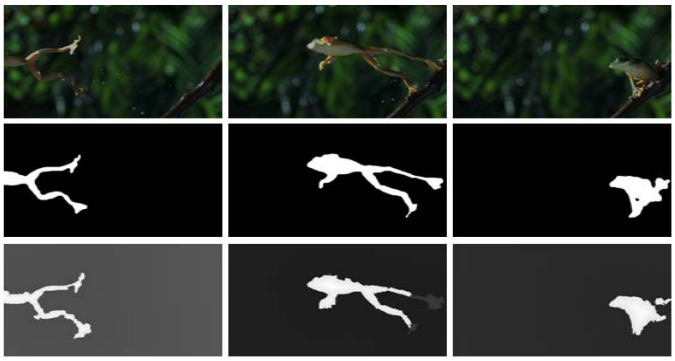


Figure 1.5 – Example of video saliency results. (top) input frames, (middle) ground truth and (bottom) video saliency results.

the direct context of the salient object provides important information on the saliency of the object (because the object is typically occluding a background) which leads us to investigate the usage of the context proposals for saliency prediction. By pairing each object proposal with its direct context proposal the computational cost can be significantly reduced. We generalize the concept of center-surround [49, 73] to arbitrary shaped object proposals by introducing context proposals. The circular or rectangular neighborhood [49, 73] do not represent the real saliency of the object very well, because part of the object is in the surround and part of the surround is in the object. This problem is addressed by our proposed context proposals.

For each object proposal we compute its direct context proposal which encompasses the object proposal and indicates the part of the image which describes its direct surrounding. To compute the saliency with respect to the context proposals, we use a similar approach as in [80]. For an object to be salient, it should be so with respect to the region described by the context proposal. As a consequence, the saliency of the object proposal is increased if the corresponding context-proposal is homogeneous in itself, and different with respect to the object segment. In [80] these observations on context-based saliency led to an iterative multi-scale accumulation procedure to compute the saliency maps. Here, however, we circumvent this iterative process by directly computing context proposals derived from ob-

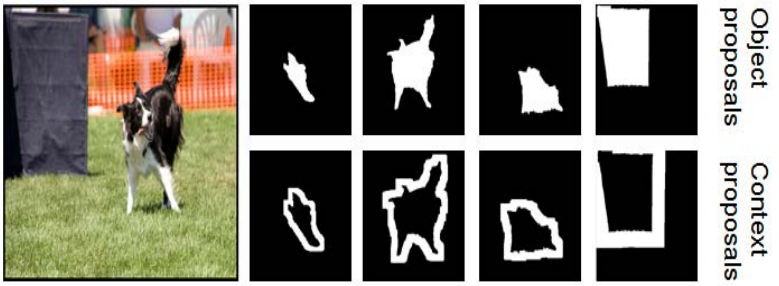


Figure 1.6 – Input image and (top row) examples of the object proposals and (bottom row) the proposed context proposals.

ject proposals, and subsequently computing the context saliency between the proposal and its context proposal. Our contribution is that we propose to couple a context proposal with each object proposal. In addition, we propose several context based features for saliency detection. An example of a set of a context proposals is shown in Figure 1.6.

1.2.2 Saliency from high-level information

In chapter 3, we present our saliency method based on the exploitation of high-level semantic information. Most computational methods are bottom-up or only include few semantic classes such as faces, and text [9, 53]. Even though much evidence of top-down saliency exists, currently most state-of-the-art methods do not explicitly incorporate it in the algorithm.

Given the recent advances in the quality of object detection and in semantic segmentation results([37, 75]) and the impressive performance gains based on deep networks, we propose to evaluate the impact that these much improved object detection and segmentation algorithms can have on saliency detection. For object detection we will consider the Fast-RCNN [37] and for semantic segmentation we use a fully convolutional network (FCN) which was proposed by Long et al. [75]. We propose several saliency features which are computed from object detection and semantic segmentation results combined with features based on object proposals. Furthermore, we

consider a wider group of twenty object classes and evaluate their impact on saliency prediction.

1.2.3 Context Proposals for Salient Object Segmentation in Videos

In chapter 4, we investigate the usage of context proposals to salient object segmentation in videos. Context proposals are derived from object proposals for video. Object proposals in video consist of a set of supervoxels grouped together and hypothesized to contain a single object [90]. Here we aim to segment the salient object from its background.

Given an input video frame, we use the spatial-temporal object proposal method of [90] to get a set of object proposals. For each object proposal we derive a set of shape features and motion features derived from the optical flow structure tensor for video saliency detection. This motion feature is based on the observation that the presence of two different motion vectors within a local neighborhood is an important feature for saliency. In addition, we apply the proposed context features which are based on context proposals for video saliency detection task.

1.3 Organization of the dissertation

This doctorate dissertation is centred on the three main subjects discussed above:

- **Context proposals for saliency detection** (chapter 2): One of the main challenges of computational saliency is the question how to incorporate context information into the computational pipeline. In this thesis we use object proposals to compute visual saliency in still images and video. We apply context proposals since the context provides important information on the saliency of the object. In our approach we propose to use object proposals methods [5], which are designed to directly provide a segmentation of the object, for the computation of context-based saliency. Since object proposals have the potential to correctly separate objects from surround, we hypothesize that considering their contrast can lead to a better saliency assessment than with other methods.

- **Saliency from high-level information** (chapter 3): Here, we investigate the usage of high-level semantic information including object detection and semantic segmentation combined with object proposal for the task of saliency prediction. Given the significant improvements of high-level object detection, semantic segmentation and the trend to use object proposal in saliency detection, we think it is timely to revisit top-down high-level features for saliency detection using the combination of object proposals with semantic information.

- **Context Proposals for Salient Object Segmentation in Videos** (chapter 4): We develop a method to detect saliency in video data. We compute saliency from supervoxels [90] and optical flow. In addition, we combine shape features with motion features. Using the results obtained from our method in the third part of thesis, we extend our method to detect saliency in video data using the context features. Our approach is tested on standard object recognition data sets. The results obtained clearly demonstrate the effectiveness of our approach.

Summary of published works

Parts of the materials presented in this doctorate dissertation have been published in/submitted to the following journals and conferences:

International Journal

- Aymen Azaza, Joost van de Weijer, Ali Douik and Marc Masana, "Context proposals for saliency Detection", Computer vision and image understanding (CVIU), accepted 2018.

- Aymen Azaza, Joost van de Weijer, Ali Douik and Javad zolfaghari , "Saliency from High-Level Semantic Image Features", IET image processing 2017 (Submitted).

- Rahma Kalboussi Aymen Azaza, Joost van de Weijer, Mehrez Abdellaoui and Ali Douik, "Object Proposals for Salient Object Segmentation in Videos", Multimedia Tools and Applications (MTA) 2018(submitted).

International Conferences

- Aymen Azaza, "Saliency detection using modified central map", Conference at University of Granada, Spain, November, 2014.
- Leila Kabbai, Aymen Azaza, Mehrez Abdellaoui and Ali Douik, "Image Matching Based on LBP and SIFT Descriptor", In the 12 th International Multi conference on Systems, Signals and Devices (SSD 2015), Mahdia, Tunisia, 2015.
- Aymen Azaza, Leila Kabbai, Mehrez Abdellaoui and Ali Douik, "Salient Regions Detection Method Inspired From Human Visual System Anatomy", In the 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP'2016), Monastir, Tunisia, 2016.
- Aymen Azaza and Ali Douik, "Saliency Detection based object proposal", 14th International Multi conference on Systems, Signals and Devices (SSD 2017), Marrakech, Morocco, 2017.
- Rahma Kalboussi Aymen Azaza, Mehrez Abdellaoui and Ali Douik, "Detecting video saliency via local motion estimation ", 14th International Conference on Computer Systems and Applications (AICCSA), 2017.

National Workshops

- Aymen Azaza, Joost van de Weijer and Ali Douik. "Object proposals for fast context based visual saliency ". In 10 th CVC workshop on the progress of research and development (CVCR& D), Bellaterra, Spain, July, 2015.

2 Context Proposals for Saliency Detection[1]

One of the fundamental properties of a salient object region is its contrast with the immediate context. The problem is that numerous object regions exist which potentially can all be salient. One way to prevent an exhaustive search over all object regions is by using object proposal algorithms. These return a limited set of regions which are most likely to contain an object. Several saliency estimation methods have used object proposals. However, they focus on the saliency of the proposal only, and the importance of its immediate context has not been evaluated.

In this chapter, we aim to improve salient object detection. Therefore, we extend object proposal methods with context proposals, which allow to incorporate the immediate context in the saliency computation. We propose several saliency features which are computed from the context proposals. In the experiments, we evaluate five object proposal methods for the task of saliency segmentation, and find that Multiscale Combinatorial Grouping outperforms the others. Furthermore, experiments show that the proposed context features improve performance, and that our method matches results on the FT datasets and obtains competitive results on three other datasets (PASCAL-S, MSRA-B and ECSSD).

2.1 Introduction

To rapidly extract important information from a scene, the human visual system allocates more attention to salient regions. Research on computational saliency focuses on designing algorithms which, similarly to human vision, predict which regions in a scene are salient. In computer vision, saliency is used both to refer to eye-fixation prediction [123, 128] as well as to salient object segmentation [52, 70]. It is the latter which is the focus of

[1] The material in this chapter has been accepted for publication in the Computer Vision and Image Understanding journal (CVIU) [6].

this chapter. Computational saliency has been used in applications such as image thumbnailing [83], compression [110], and image retrieval [117].

Object proposal methods have recently been introduced in saliency detection methods [70]. They were first proposed for object recognition, which was long dominated by sliding window approaches (see e.g. [30]). Object proposal methods reduce the number of candidate regions when compared to sliding window approaches [114]. They propose either a set of bounding boxes or image segments, which have a high probability of containing an object [43, 58]. Recently, these methods have been applied in saliency detection [33, 70, 119]. Object proposals especially help in obtaining exact boundaries of the salient objects [70]. In addition, they can reduce the computational costs of evaluating saliency based on a sliding window [73].

The saliency of an object is dependent on its context, i.e. an object is salient (or not) with respect to its context. If a visual feature, e.g. color, textures or orientation, of an object differs from that of its context it is considered salient. Traditionally, this has been modeled in saliency computation with the center-surround mechanism [35, 40], which approximates visual neurons. This mechanism divides the receptive field of neurons into two regions, namely the center and surround, thereby modeling the two primary types of ganglion cells in the retina. The first type is excited by a region in the center, and inhibited by a surround. The second type has the opposite arrangement and is excited from the surround and inhibited by a center. In computational saliency the center-surround mechanism has been implemented in different ways. For example, [49] model this by taking the difference between fine (center) and coarse scale (surround) representations of image features. Even though this has been shown to successfully model eye fixation data, for the task of salient object detection this approach is limited to the shapes of the filters used. It can only consider the differences between circle regions of different radii. This led [73] to consider center-surround between arbitrary rectangles in the images for salient object detection. In this work we will further generalize the concept of center-surround but now to arbitrarily shaped object proposals.

To generalize the concept of center-surround to arbitrary shaped object proposals we extend object proposals with context proposals. We consider any object proposal method which computes segmentation masks. For

each object proposal we compute a context proposal which encompasses the object proposal and indicates the part of the image which describes its direct surrounding. To compute the saliency with respect to the context proposals, we use a similar approach as [80]. For an object to be salient, it should be so with respect to the region described by the context proposal. In addition, because typically an object is occluding a background, it is expected that the features in the context proposal do not vary significantly. As a consequence, the saliency of the object proposal is increased if the corresponding context-proposal is homogeneous in itself, and different with respect to the object segment. In [80] these observations on context-based saliency led to an iterative multi-scale accumulation procedure to compute the saliency maps. Here, however, we circumvent this iterative process by directly computing context proposals derived from object proposals, and subsequently computing the context saliency between the proposal and its context proposal.

Our main contribution is that we propose several context based features for saliency estimation. These are computed from context proposals which are computed from object proposals. To validate our approach we perform experiments on a number of benchmark datasets. We show that our method matches state-of-the-art on the FT dataset and improves state-of-the-art results on three benchmark (PASCAL-S, MSRA-B and ECSSD datasets). In addition, we evaluate several off-the-shelf deep features and object proposal methods for saliency detection and find that VGG-19 features and multiscale combinatorial grouping (MCG) obtain the best performance.

This chapter is organized as follows. In Section 2.2 we discuss the related work. In Section 2.3 we provide an overview of our approach to saliency detection. In Section 2.4 the computation of context proposals is outlined. Next we provide details on the experimental setup in Section 2.5 and give results in Section 2.6. Conclusions are provided in Section 2.7.

2.2 Related work

In this section we provide an overview of salient object detection methods and their connection with object proposal methods. More complete reviews on saliency can be found in [12, 133, 137].

2.2.1 Saliency detection

One of the first methods for computational saliency was proposed by [49]. Their model based on the feature integration theory of [113] and the work of [57] decomposes the input image into low level feature maps including color, intensity and orientation. These maps are subsequently merged together using linear filtering and center surround structures to form a final saliency map. Their seminal work initiated much research in biologically inspired saliency models [35, 87, 105] as well as more mathematical models for computational saliency [2, 41, 45, 68]. The central surround allows to measure contrast with the context, however it is confined to predefined shapes; normally the circle shape of the Gaussian filters [49] or rectangle shapes in the work of [73]. Here we will propose a method for arbitrary shaped contexts. Local and global approaches for visual saliency can be classified in the category of bottom-up approaches. Local approaches compute local center-surround contrast and rarity of a region over its neighborhoods. [49] derive a bottom-up visual saliency based on center surround difference through multiscale image features. [73] propose a binary saliency estimation method by training a CRF to combine a set of local, regional, and global features. [41] propose the GBVS method which is a bottom-up saliency approach that consists of two steps: the generation of feature channels as in Itti's approach, and their normalization using a graph based approach. A saliency model that computes local descriptors from a given image in order to measure the similarity of a pixel to its neighborhoods was proposed by [103]. [36] propose the adaptive whitening salience (AWS) method which is based on the decorrelation and the distinctiveness of local responses.

Another class of features for saliency are based on global context or rarity, the saliency of a feature is based on its rarity with respect to the whole image. [39] consider the difference of patches with all other patches in the image to compute global saliency. [121] compute saliency by considering the reconstruction error which is left after reconstructing a patch from other patches (other patches can be from the same image or from the whole dataset). [52] compute the rarity of a feature by comparing the contrast between a 15 pixel border around the image and the object proposal histogram. Other than these methods we propose a method to compute the

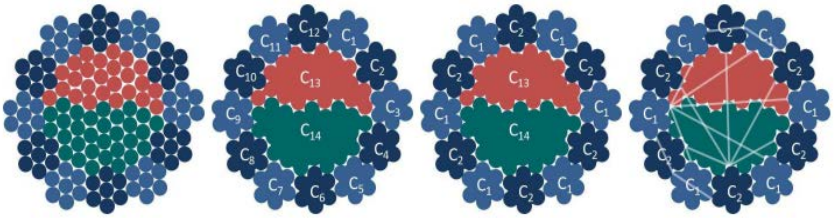


Figure 2.1 – The steps of Mairon's [80] approach. (From left to right) Initial coxels with their color-coded appearance content. Coxels with small contextual gaps, initially those which are very similar are merged to larger. The fusing of image regions based on their color distance into larger and larger context segments. The accumulation of saliency votes by the context segments. Figure taken from [80].

saliency with respect to the direct context of the object. Finally, to compute saliency [47] combined local and global objectness cues with a set of candidates location.

Our work has been inspired by a recent paper [80] which demonstrates the importance of visual context for saliency computation. The work is based on the observation that an object is salient with respect to its context. And since context is an integral part of saliency of an object, it should therefore be assigned a prominent role in its computation. The final saliency map is computed by alternating between two steps: 1. the fusing of image regions based on their color distance into larger and larger context segments, and 2. the accumulation of saliency votes by the context segments (votes are casted to the region which is enclosed by the context segments). The steps are alternated until the whole image is clustered together into a single context segment (see Fig 2.1). The procedure is elegant in its simplicity and was shown to obtain excellent results. However, the iterative nature of the computation renders it computationally very demanding.

Deep convolutional neural networks have revolutionized computer vision over the last few years. This has recently led to several papers on deep learning for saliency detection [23, 66, 91, 119, 138]. Both [66] and [138]

consider parallel networks which evaluate the image at various scales. [119] use two networks to describe local and global saliency. [112] combine a local and global model to compute saliency. The main challenge for saliency detection with deep networks is the amount of training data which is not always available. This is solved in [66, 119, 138] by training on the largest available saliency dataset, namely MSRA-B [73], and testing on the other datasets (both [64, 66] also use pretrained network weights trained on the 1M Imagenet dataset). Like these method, we will use a pretrained network for the extraction of features for saliency detection.

2.2.2 Object proposal methods

Object detection based on object proposals methods has won in popularity in recent years [114]. The main advantages of these methods is that they are not restricted to fixed aspect ratios as most sliding window methods are, and more importantly, they allow to evaluate a limited number of windows. As a consequence more complicated features and classifiers can be applied, resulting in state-of-the-art object detection results. The generation of object hypotheses can be divided into methods whose output is an image window and those that generate object or segment proposals. The latter are of importance for salient object detection since we aim to segment the salient objects from the background.

Among the first object proposal methods the work of [19], named the Constrained Parametric Min-Cuts (CPMC) method, uses graph cuts with different random seeds to obtain multiple binary foreground and background segments. [4] proposes to measure the objectness of an image window, where they rank randomly sampled image windows based on their likelihood of containing the object by using multiple cues among which edges density, multiscale saliency, superpixels straddling and color contrast. [27] proposed an object proposal method similar to the CPMC method by generating multiple foreground and background segmentations. A very fast method for object proposals was proposed by [21], which generates box proposals at 300 images per second.

An extensive comparison of object proposal methods was performed by [43]. Among the best evaluated object proposal methods (which generate

object segmentation) are the selective search [114], the geodesic object proposals [58] and the multiscale combinatorial grouping method [5]. Selective search proposes a set of segments based on hierarchical segmentations of the image where the underlying distance measures and color spaces are varied to yield a large variety of segmentations. [58], propose the geodesic object proposals method, which applies a geodesic distance transfer to compute object proposals. Finally, Multiscale Combinatorial Grouping [5] is based on a bottom-up hierarchical image segmentation. Object candidates are generated by a grouping procedure which is based on edge strength.

Several methods have applied object proposals to saliency detection [33, 70, 119]. The main advantage of saliency detection methods based on object proposals over methods based on superpixels [130] is that they do not require an additional grouping phase, since the object proposals are expected to encompass the whole object. Other than general object detection, salient object segmentation aims at detecting objects which are salient in the scene. Direct surrounding of objects is of importance to determine the object's saliency. Therefore, we extend the usage of object proposals for saliency detection with context proposals, which allow us to directly assess the saliency of the object with respect to its context.

2.3 Method Overview

The main novelty of our method is the computation of context features from context proposals. To illustrate the advantage of this idea consider Fig. 2.2. In this figure several implementation of the center surround idea for saliency detection are shown. The circular surround was used in the original work by [49]. This concept was later generalized to arbitrary rectangles [73]. Both these approaches have the drawback that they only are a rough approximation of the real object shape and the contrast between the (circle or rectangular) object and its surround does not very well represent the real saliency of the object. This is caused by the fact that when we approximate the object by either a square or circle, part of the object is in the surround, and part of the surround is in the object.

In principle the center surround idea could be extended to superpixels which are often used in saliency detection [130], see Fig. 2.2. However, super-

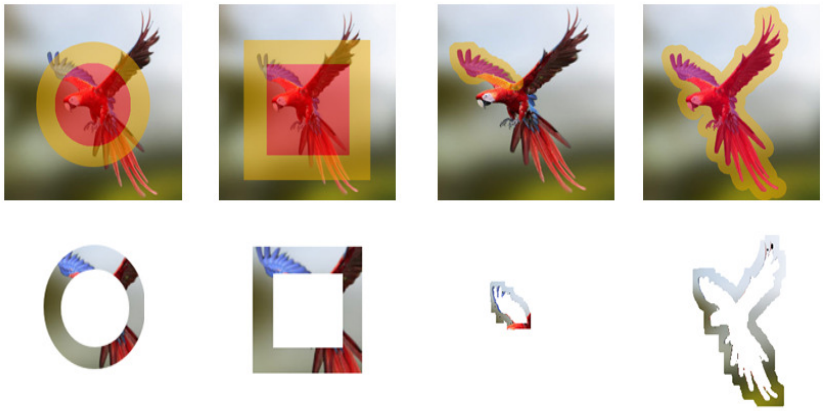


Figure 2.2 – Top row: examples of different center surround approaches (a) circular surround (b) rectangular surround (c) superpixels surround, and (d) the context proposals. Bottom row: the surround for each of the methods. It can be seen that only the object proposal based surround correctly separates object from background.

pixels generally only cover part of the object, and therefore their surround is often not homogeneous, complicating the analysis of the saliency of the center. Finally, [80] show that a surround which can adapt to the shape of the object (center) is an excellent saliency predictor. For its computation they propose an iterative procedure. Here, we propose to use object proposals methods [5], which are designed to directly provide a segmentation of the object, for the computation of context-based saliency. Since object proposals have the potential to correctly separate object from surround (see final column on the right in Fig. 2.2), we hypothesize that considering their contrast can lead to a better saliency assessment than with the other methods.

An overview of the saliency detection algorithm is provided in Fig. 2.3. Next, any object proposal algorithm can be used here that provides pixel-precise object contours, such as [5, 46, 58, 95, 114]. We extend each object

proposal with a context proposal which is its immediate surround (see Section 2.4.1). We then proceed by computing deep features for both the object proposal and its context proposal from which we derive several context features (see Section 2.4.2).

Given the feature vector of the object and context for each of the proposals in the training set we train a random forest classifier. As the saliency score for each object proposal we use the average saliency of the pixels in the proposal: pixels have a saliency of one if they are on the ground truth salient object or zero elsewhere (this procedure is further explained in Section 2.4.5). At testing time we infer the saliency for all the object proposals by applying the random forest regressor. The final saliency map is computed by taking for each pixel the average of the saliency of all the proposals that contain that pixel.

The overall method is similar to several previous papers on saliency. A similar approach was proposed by [52] and later used by [70]. In [52] they use a random forest classifier to score each region in the image instead of every object proposal in our method. [70] use the CPMC method for object proposals [19] and similar as [52] they apply a random forest to predict region saliency based on regional features. In contrast to these methods we investigate the usage of context proposal for object saliency detection.

2.4 Context Proposals for Saliency Computation

The main novelty of our approach is the introduction of context proposals for saliency computation. Here, we describe our approach to context proposal generation and how, from these, we compute context features.

2.4.1 Context Proposal Generation

Recently, several saliency methods have applied object proposal algorithms to generate proposals for salient objects [52, 70]. Consider an object proposal, represented by the mask M which is equal to one for all pixels within the object proposal and zero otherwise. Then we define the context of the

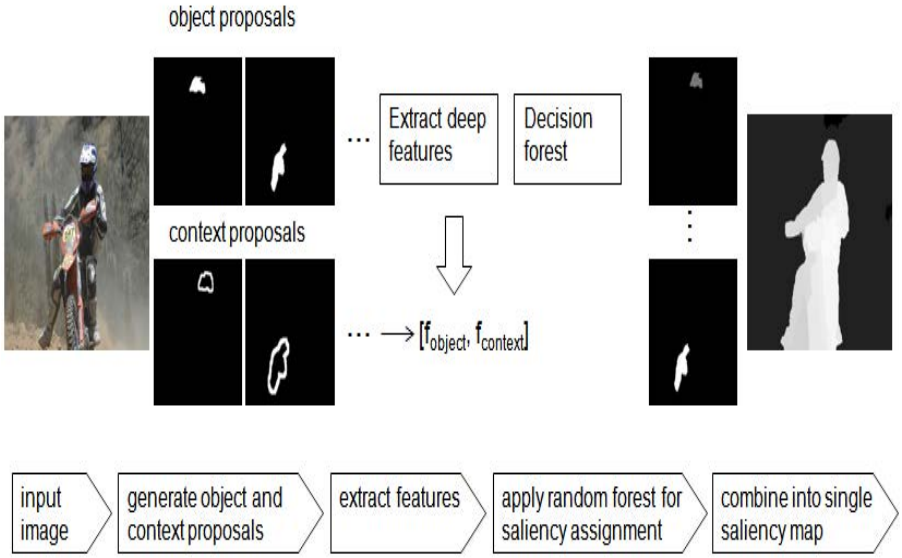


Figure 2.3 – Overview of our method at test time. A set of object proposals is computed. From these a set of accompanying context proposals is derived. We extract deep convolutional features from both object and context ($\mathbf{f}_{object}$ and $\mathbf{f}_{context}$). At training for each object proposal its saliency is computed based on the ground truth, and a random forest is trained to regress to the saliency. At testing this random forest is applied to predict the saliency of all proposals, which are combined in the final saliency map

proposal to be

$$C = \left(M \oplus B^{(n)}\right) \setminus M \quad \text{smallest } n \text{ for which } |C| \geq |M| \tag{2.1}$$

where B is a structural element and $\oplus$ is the dilation operator. We used the notation

$$B^{(n)} = \overbrace{B^{(1)} \oplus B^{(1)} \oplus B^{(1)} \oplus B^{(1)}}^{n\ times} \tag{2.2}$$

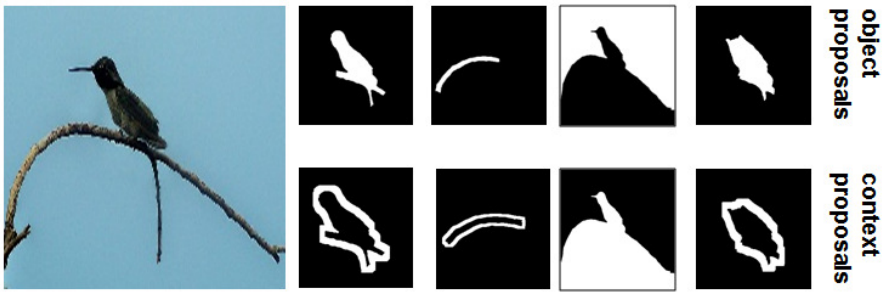


Figure 2.4 – Input image and (top row) examples of object proposals and (bottom row) examples of context proposals.

to indicate multiple dilations. In our work we choose $B = N_8$ which is the eight connected set (a 3x3 structural element with all ones). We use $|C|$ to indicate the number of non-zero values in C. If we would consider arbitrary n in the first part of this equation, this equation could be interpreted as generating a border for the object proposal M which thickness is equal to n. We define the context to be the smallest border which has equal or more pixels than M. In practice, the context is computed by iteratively dilating with B until we reach a border which contains more pixels than the object proposal M. In Fig. 2.4 we provide examples of context borders for several object proposals. Note that the context border is wider for larger object proposals. The idea is to verify if the object proposal is salient with respect to its context.

2.4.2 Context Feature Computation

Next we outline the computation of the context features. We consider two properties which define a good context proposal. *Context contrast* which measures the contrast between the features which make up the salient object and the features which describe its context. Secondly *Context continuity* which is based on the observation that the salient object is often occluding a background which continues behind it. As a consequence, we expect the features which describe the context on opposite sides of the salient object to be similar. In human vision research it was verified that salient objects (targets)

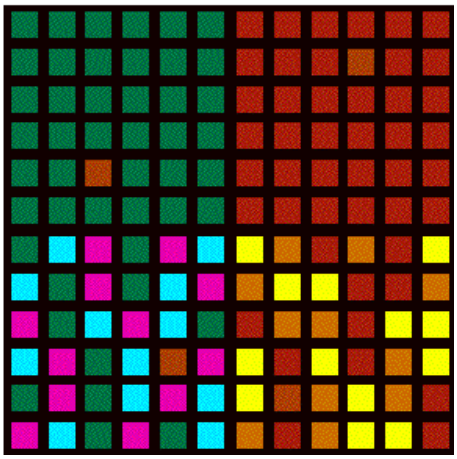


Figure 2.5 – We consider the dark orange target in each quadrant. The figure contains four quadrants. From left to right the context contrast decreases. From top to bottom context continuity decreases. If the distractors around a target object are homogeneous (e.g. all green), the target is found faster and perceived as more salient then in the case where the distractors are non-homogeneous (e.g. pink, cyan and green). Figure taken from [24].

are faster found on a homogeneous background than when surrounded by a heterogeneous background (distractor) [24]. An example is shown in Fig. 2.5. Context continuity is an indicator of background homogeneity, since homogeneous backgrounds lead to higher context continuity, and heterogeneous ones would lead to low context continuity.

The first context saliency feature which we consider combines both described properties, context contrast and context continuity, into a single measure. Consider a pixel m_i in the object proposal M. Then we define two related coordinates d_i^φ and u_i^φ which are coordinates of the points on the context when considering a line with orientation φ through point m_i (see Fig. 2.6). The saliency of a point m_i is larger when the feature representation at m_i is more different from the feature representation on its context at d_i and u_i. In addition, we would like the distance between the points on the context (d_i and u_i) to be similar. Combining these two factors in one

saliency measures yields:

$$c_1^{\varphi}(m_i) = \arctan\left(\frac{\min\left(s_i^{d,\varphi}, s_i^{u,\varphi}\right)}{s_i^{du,\varphi} + \lambda}\right). \tag{2.3}$$

where the numerator contains the context contrast and the denominator the context continuity. The $arctan$ and the constant λ are used to prevent large fluctuations in saliency for small values of $s_i^{du,\varphi}$. The distances are defined with

$$s_i^{u,\varphi} = \left\|\mathbf{f}\left(u_i^{\varphi}\right) - \mathbf{f}(m_i)\right\|, \tag{2.4}$$

$$s_i^{d,\varphi} = \left\|\mathbf{f}\left(d_i^{\varphi}\right) - \mathbf{f}(m_i)\right\|, \tag{2.5}$$

$$s_i^{du,\varphi} = \left\|\mathbf{f}\left(d_i^{\varphi}\right) - \mathbf{f}\left(u_i^{\varphi}\right)\right\|. \tag{2.6}$$

Here $\mathbf{f}(m_i)$ denotes a feature representation of the image at spatial location m_i, and $\|.\|$ is the L2 norm. This feature representation could for example be the RGB value at that spatial location, but also any other feature representation such as for example a deep convolutional feature representation as we will use in this chapter. We have also performed experiments with both tanh and L1 norm. We found arctan and L2 to yield superior results and therefore use these in our experiments.

Now that we have defined the saliency for a single point considering its context points along a line with orientation φ, we define the overall saliency for a context proposal as the summation over all pixels m_i in the object proposal considering all possible lines:

$$C^1 = \frac{1}{|M|}\sum_{m_i \in M}\int_0^{\pi} c_1^{\varphi}(m_i)\, d\varphi. \tag{2.7}$$

It should be noted that we exclude lines which do not have context on both sides of the object. This happens for example for objects on the border of the image.

Considering all orientations is computationally unrealistic and in prac-

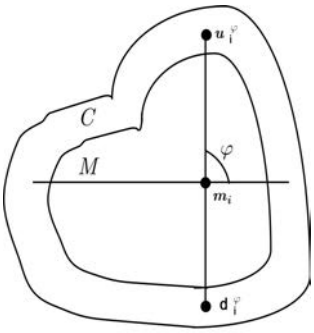


Figure 2.6 – Graphical representation of variables involved in context feature computation.

tice we approximate this equation with

$$C^1 = \frac{1}{|M|} \sum_{m_i \in M} \sum_{\varphi \in \Phi} c_1^{\varphi}(m_i), \tag{2.8}$$

where Φ is a set of chosen orientations between $[0, \pi)$. So we have considered four orientations

$$\Phi = \left\{ 0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4} \right\}. \tag{2.9}$$

The saliency of one point in the object proposal is hence computed by considering its context along four orientations. To be less sensitive to noise on the context both $\mathbf{f}(d_i^{\varphi})$ and $\mathbf{f}(u_i^{\varphi})$ are extracted from a Gaussian smoothed context proposal.

As a second context feature we ignore the object proposal and only consider the values on the context proposal to compute the saliency. This feature solely focuses on context continuity. In this case we would like the saliency to be larger when the values on the context have a smaller distance. We

Figure 2.7 – Context continuity: features on opposites sides of the object proposal are expected to be similar. Examples of (left) omni-directional context continuity and (right) horizontal context continuity.

propose to use the following measure:

$$c_2^{\varphi}(m_i) = \arctan\left(\frac{1}{s_i^{du,\varphi} + \lambda}\right) \tag{2.10}$$

again λ prevents large fluctuations for low values of s_i.

Similarly we compute the $C^2(m_i)$ for the object proposal with

$$C^2 = \frac{1}{|M|} \sum_{m_i \in M} \int_0^{\pi} c_2^{\varphi}(m_i)\, d\varphi. \tag{2.11}$$

and its approximation

$$C^2 = \frac{1}{|M|} \sum_{m_i \in M} \sum_{\varphi \in \Phi} c_2^{\varphi}(m_i). \tag{2.12}$$

In addition to C^1 and C^2 which measure context saliency based on com-

paring features on all sides of the object proposal, we introduce also a measure for horizontal context continuity C^3 where we use $\Phi^H = \{0\}$, and we compute

$$C^3 = \frac{1}{|M|} \sum_{m_i \in M} \sum_{\varphi \in \Phi^H} c_1^{\varphi}(m_i). \tag{2.13}$$

The motivation for a special measure for horizontal context continuity is provided in Fig. 2.7. Natural scenes contain more horizontal elongation than other orientations; the C^3 measure is designed to detect horizontal clutter.

The context measures proposed here are motivated by the work of [80]. They propose an iterative procedure to compute context based saliency. We prevent the iterative procedure by directly computing the context from the object proposals. In addition, we propose a measure of horizontal context which is not present in [80]. Also instead of RGB features we use deep features to compute the context saliency.

2.4.3 Off-the-Shelf Deep Features

The deep features, we use as the feature **f** in Eq. 2.4-2.6 to compute the three context features Eq. 2.8, Eq. 2.12- 2.13. These are combined into one context feature

$$\mathbf{f}_{context} = \{C^1, C^2, C^3\} \tag{2.14}$$

for each context proposal. The deep feature is also used directly as a descriptor for the object proposal by pooling the deep feature over all pixels in the object proposal with

$$\mathbf{f}_{object} = \frac{1}{|M|} \sum_{m_i \in M} \mathbf{f}(m_i) \tag{2.15}$$

Deep convolutional features have shown excellent results in recent papers on saliency [65, 66, 119, 138]. A straight-forward way to use deep features is by using a pre-trained network, for example trained for the task of image classification on ImageNet [59], to extract features. These so called off-the-shelf features can then be used as local features. A good overview

net. / bl.	AlexNet	VGG-19	ResNet-152
1.	$[11\times11,96]$	$[3\times3,64]\times2$	$[7\times7,64]$
2.	$[5\times5,256]$	$[3\times3,128]\times2$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
3.	$[3\times3,384]$	$[3\times3,256]\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times8$
4.	$[3\times3,384]$	$[3\times3,512]\times4$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times36$
5.	$[3\times3,256]$	$[3\times3,512]\times4$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$

Table 2.1 – Overview of the *convolutional* layers of different networks. The convolutional part can be divided in 5 blocks (bl.) for all three networks. For each block we show the convolutional size, the number of features, and how many times this layer pattern is repeated. The non-linear activation layers are omitted. In our evaluation we will use the last layer of each block to extract convolutional features.

of this approach is given by [99], who successfully apply this technique to a variety of tasks including object image classification, scene recognition, fine grained recognition, attribute detection and image retrieval.

A convolutional network alternates linear and non-linear layers. The convolutional layers are typically used in the first layers of the network. These consist of a set of filters which describe local structure in images. In most networks the output of the convolutional layer is processed by a nonlinear Rectified Linear Unit (ReLu) which removes all non-positive activations, and a max-pooling layer. The max-pooling layer reduces the total capacity (number of parameters) of the network by reducing the resolution and it provides robustness to small translations of the features. The final layers often consist of fully connected layers, which are alternated with ReLu layers.

Extensive studies on understanding the image representations of the

different layers have revealed that the feature complexity increases with layers. Starting with simple low-level features such as corners, blobs and edges in the first layers, the network learns to represent increasingly more complex structures in the higher layers, such as faces, instruments, tools, objects. Of interest in this chapter is to establish which layer is optimal for salient object detection.

To choose the best deep features for saliency detection we evaluate three popular networks, namely AlexNet [59], VGG-19 [107] and ResNet [42]. The configuration of the convolutional layers of the networks is given in Table 2.1. We evaluate the performance of the different blocks for saliency estimation. The results using both object features $\mathbf{f}_{object}$ and context features $\mathbf{f}_{context}$ are summarized in Fig. 2.9. We found the best results, similar to the ResNet, were obtained with block 5 of VGG-19 (which layer name is *conv5_4*). Based on these results we choose to extract block 5 deep features with VGG-19 for all images. We spatially pool the features within each object to form a 512-dimensional $\mathbf{f}_{object}$ and the 3-dimensional $\mathbf{f}_{context}$ according to Eq. 2.14-2.15. In addition, we found that applying a standard whitening, where we set the variance over all features of $\mathbf{f}_{object}$ to 1, prior to applying the classifiers improved results.

2.4.4 Whitening

Whitening has since long been considered an operation which closely relates to the human saliency assignment, it is based on the idea that frequent features are suppressed and rare features are amplified. This idea is also confirmed by information theory where rare events are more informative than normal events.

We used the color boosting function g which is approximated from the distribution of image derivatives proposed by [116]. According to their study the distribution of image derivative has an ellipsoid shape, whose parameters can be estimated by the covariance matrix M.

Mathematically, whitening can be done using the singular value decomposition (SVD) of the covariance matrix M of the filters of the layer. The covariance matrix could be decomposed into an eigenvector matrix U, a diagonal matrix S and an eigenvalue matrix matrix V. After the rotation each

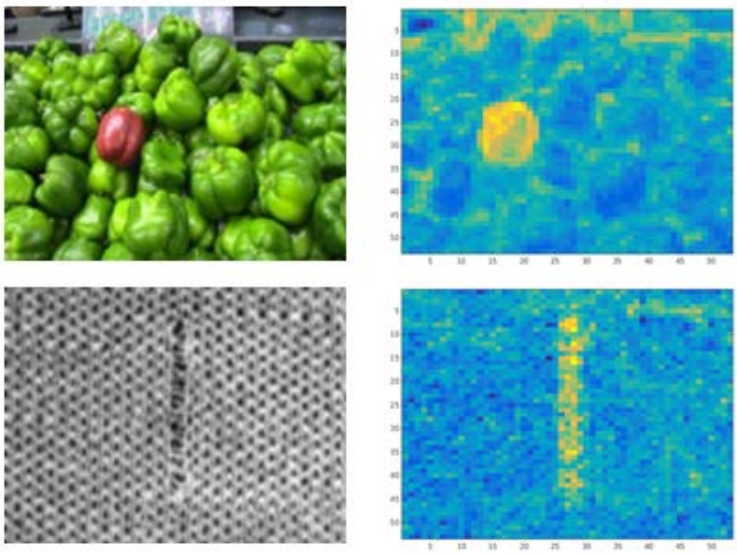


Figure 2.8 – Example of applying whitening on the $\mathbf{f}_{object}$ features.

component will have the variance given by its corresponding eigenvalue. Therefore, to make variances equal to 1, we divide by the square root of S.

Using the function g proposed by [116], we are able to apply the whitening in the layer, with which the ellipses are reshapes to spheres.

$$M = USV \tag{2.16}$$

$$g(L) = U.\left(diag\left(\frac{1}{\sqrt{(diag(S)+\lambda)}}\right).V^T\right).L \tag{2.17}$$

where L is the layer to be whitened. An example of whitening results is shown in Figure 2.8.

We compute *Deep whitened features* which are based on the value of the object proposal within the whitened different filter of the layer resulting in a 512 dimensional vector for each object proposal.

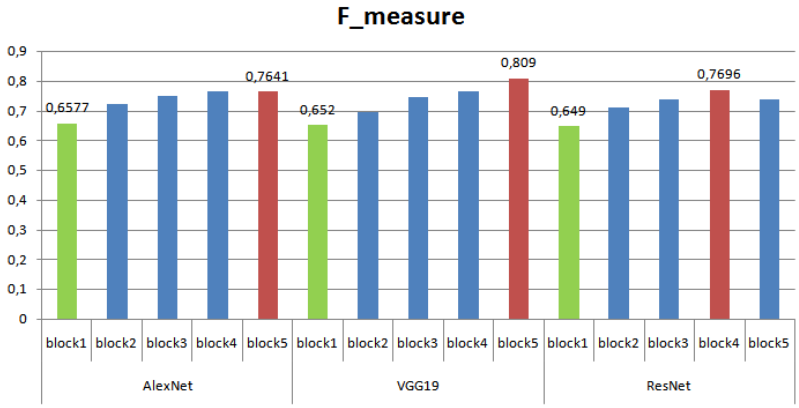


Figure 2.9 – Evaluation on 5 convolutional layers for the three architectures used in our framework.

2.4.5 Saliency Score of Object Proposals

Based on the features which are extracted from the object proposal and its context we train a random forest regressor to estimate the saliency of the object proposal. To compute the saliency score, sal^{object}, for object proposals we use the following equation:

$$sal^{object} = \frac{|M \cap S|}{|M|} \tag{2.18}$$

here M is the set of all pixels in the object proposal and S is the set of all pixels which are considered salient in the ground truth. A $sal = 0.8$ means that 80% of the pixels in the object proposal are considered salient.

We found that this score is not optimal when considering context proposals, and we propose to use the following equation

$$sal^{context} = \max\left(\frac{|M \cap S|}{|M|} - \frac{|C \cap S|}{|C|}, 0\right) \tag{2.19}$$

where C is the set of pixels in the context. The $sal^{context}$ measure lowers the score if salient pixels are in the context.

We train two separate random forest regressors, one based on the deep features of the object proposal regressing to sal^{object} and one based on the context features regressing to $sal^{context}$. The final saliency score at testing time is computed by adding results of the two regressors. The final saliency map is computed by averaging the saliency of all the object proposals which are considered in the image. We have also considered to assign to each pixel the maximum saliency of all object proposals which include the pixel, but found this to yield inferior results.

As an additional experiment we have also considered the following scenarios, where we consider a single regressor to obtain the saliency:

1. We regress from both $\mathbf{f}_{object}$ and $\mathbf{f}_{context}$ (e.g. concatenation) directly to sal^{object} (2.18)

2. We regress from both $\mathbf{f}_{object}$ and $\mathbf{f}_{context}$ directly to $sal^{context}$ (2.19)

Results obtained are 79.82 for method (1) and 79.94 for method (2) which are both inferior to our current method (which obtains 80.90) where we regress from $\mathbf{f}_{object}$ to sal^{object} and $\mathbf{f}_{context}$ to $sal^{context}$, and then average the outputs.

2.5 Experimental Setup

In this section we describe the features on which we base the saliency computation, the datasets on which the experiments are performed, and the evaluation protocol we use.

2.5.1 Datasets

We evaluate our proposed algorithm on several benchmark datasets that are widely used.

Pascal-S [70]: This dataset was built on the validation set of the Pascal VOC 2010 segmentation challenge. It contains 850 images with both saliency segmentation ground truth and eye fixation ground truth. Saliency ground truth masks were labeled by 12 subjects. Many of the images in this dataset contain multiple salient objects.

MSRA-B [73]: This dataset contains 5,000 images and is one of the most used datasets for visual saliency estimation. Most of the images contain only one salient object.

FT [2]: This dataset contains 1,000 images, most of the images contain one salient object. It provides only salient object ground truth which is derived from [126] and is obtained using user-drawn rectangles around salient objects.

ECSSD [129]: It contains 1,000 images acquired from the PASCAL VOC dataset and the internet and the ground truth masks were annotated by 5 subjects.

2.5.2 Evaluation

We evaluate the performance using PR (precision-recall) curve and F-measure. Precision measures the percentage of salient pixels correctly assigned, and recall the section of detected salient pixels which belongs to the salient object in the ground truth.

We compute precision and recall of saliency maps by segmenting the salient object with a threshold T and comparing the binary map with the ground truth. All saliency maps are also evaluated using the F-measure score which is defined as:

$$F_\beta = \frac{\left(1+\beta^2\right)\cdot \text{precision}\cdot \text{recall}}{\beta^2\cdot \text{precision} + \text{recall}} \tag{2.20}$$

where β^2 is set to 0.3 following [70, 111, 119, 138]. As a threshold we use the one which leads to the best F_β. This was proposed in [13, 85] as a good summary of the precision-recall curve. We compare our method against 8 recent CNN methods: Deeply supervised salient object (DSS)[44], Deep contrast learning (DCL) [64], Recurrent fully convolutional networks (RFCN) [120], Deep hierarchical saliency (DHS) [72], Multi-task deep saliency (MTDS) [69], Multiscale deep features (MDF) [66], Local and global estimation (LEGS) [119], Multi context (MC) [138] and we compare also against 8 classical methods including Discriminative regional feature integration (DRFI) [52], Hierarchical saliency (HS) [129], Frequency tuned saliency (FT) [2], Regional principal color based saliency detection (RPC) [76], (CPMC-GBVS) [70],

Graph-based manifold ranking (GBMR) [130], Principal component analysis saliency (PCAS) [84], Textural distinctiveness (TD) [101] and a Context aware method [39] (GOF). For a fair comparison we did not include (CPMC-GBVS) method[70] because they use eye fixation label in training.

Based on crossvalidation experiments on PASCAL-S training set we set the number of trees in the random forest to 200, we set $\lambda = 40$ in Eq. 2.3 and Eq. 2.10 and we set the minimum area of object proposals to be considered at 4,500 pixels. We use these settings for all datasets.

2.6 Experimental Results

In this section we provide our experimental results. We provide an evaluation of five popular object proposal approaches. Next we evaluate the relative gain which is obtained by adding the features based on context proposals. We evaluate also our context features with different context shapes including the conventional circular or rectangular neighborhood. Finally, we compare to state-of-the-art methods on several benchmark datasets.

2.6.1 Object Proposal based Saliency Detection

In recent years several methods have proposed to use object proposals for salient object segmentation. However, to the best of our knowledge, there is no work which evaluates the different object proposal approaches to saliency detection. [43] have provided an extensive evaluation of object proposals for object detection. Based on their analysis we have selected the three best object proposal methods which output segments based on their criteria, namely repeatability, recall, and detection results. The object proposal methods we compare to are selective search (SS) [114], the geodesic object proposals (GOP) [58], and the multiscale combinatorial grouping (MCG) method [5]. We have added two recent object proposals to this list which are based on deep learning, namely FastMask [46] and SharpMask [95]. We do these experiments on the PASCAL-S dataset because it is considered one of the most challenging saliency datasets; also it is labeled by multiple subjects without restriction on the number of salient objects [70].

We present here a brief description of the object proposal methods which

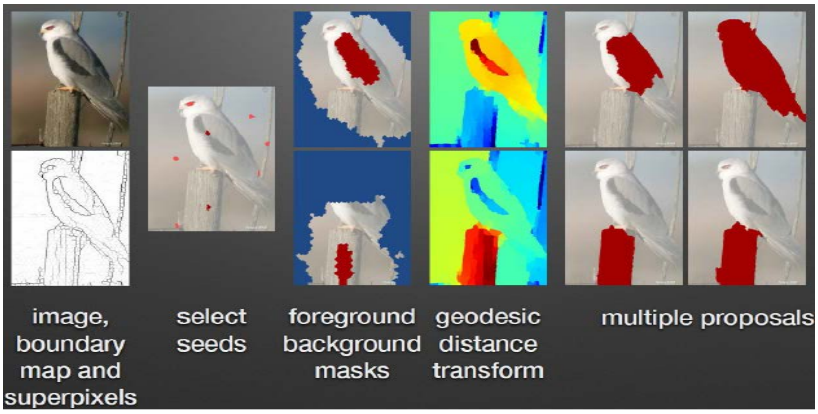


Figure 2.10 – Overview of GOP method. Figure taken from [58].

we consider:

Geodesic Object Proposals (GOP) Krähenbühl et al. [58] proposed the Geodesic Object Proposals method. The GOP method generates the least number of proposals when we use the default settings. The pipeline presented in their paper is shown in Fig 2.10, and can be divided into four steps. starting by decomposing the input image into super-pixels. The next step is seed selection, which aim is to select maximum objects using the least number of seeds. This leads to the least number of object proposals, which reduces the computational cost. Using these selected seeds they generate foreground and background masks. Finally, they compute the signed geodesic distance transform (SGDT) for the background and the foreground masks of the image. Every level set of the SGDT is an object proposal.

Multiscale Combinatorial grouping (MCG) Arbelaez et al. [5] merge regions from different scales into possible object proposals. They uses a bottom-up hierarchical image segmentation approach which segment the image separately into multiple resolutions forming a pyramid. They divide the image into a set of superpixels, then they align and merge all hierarchical boundaries in a multiscale hierarchy (see Fig 2.11). Finally, they generate object candidates using the information about location, size and shape.

Selective Search (SS) Uijlings et al [114] propose the selective search method.

Figure 2.11 – Overview of MCG method. Figure taken from [5].

This method is based on an exhaustive search to generate object candidates. They create small segment proposals at every possible scales. Then, combine them together in an iterative way using the similarity score. After merging two homogeneous regions, the score is updated. They continues until obtaining a single region. In Fig 2.12 the pipeline and hierarchy of the selective search method is provided.

SharpMask Contrary to traditional approaches [5, 58, 114], which use low level features, DeepMask method [94] is a deep learning approach which generate object segments from CNN features. It is using a body-head structure to generate object segments. The SharpMask method [95] is a refinement method of DeepMask which adds a backward branch to refine the masks, which helps to carefully segment the objects boundaries. They generate image segments at multiple scales using image pyramid structure during inference. Fig 2.13 shows the framework of the SharpMask.

FastMask Hu et al. [46] propose a one-shot model which enables training and inference to avoid image pyramid structure. They use the body-head structure of [95] and propose a new component, called neck. This neck could be used on the feature map and zoom it out into feature pyramid. Then, a new module (shared head) is used on the pyramid of features to generate segments at different scales. Using body-neck-head structure, they generate

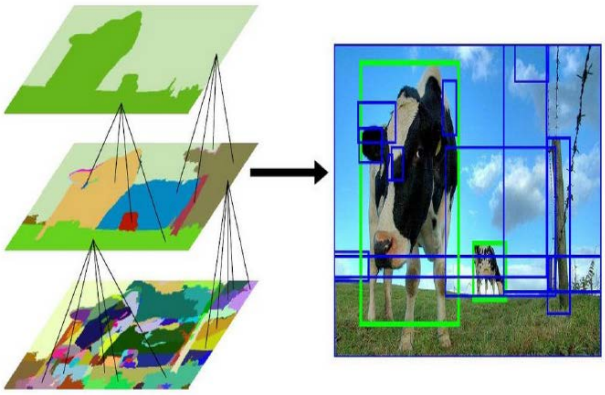


Figure 2.12 – Overview of SS method. Figure taken from [114].

segment proposal in one shot. In Fig 2.14 the one-shot segment proposal is shown.

In addition to comparing the three object proposal methods, We evaluate also the performance of object proposal methods as a function of proposals. Results are provided in Table. 2.2. Results of MCG are remarkable already for as few as 16 proposals per image, and they stay above the other methods when increasing the number of proposals. The results of SS can be explained by the fact that the ranking of their proposals is inferior to the other methods. The inferior ranking is not that relevant for object detection where typically thousands of proposals are considered per image[2]. The results of the two methods based on deep learning, namely FastMask and SharpMask, are somewhat surprising because they are known to obtain better results for object detection [46, 95]. In a closer analysis we found that MCG obtains higher overlap (as defined by IoU) with the salient object ground truth(Fig 2.15). In addition, deep learning approaches typically extract the salient object among the first 8-16 proposals, and therefore do not improve, and sometimes even deteriorate, when considering more proposals. Based on the results we select MCG to be applied on all further experiments, and we set

[2]Selective search applies a pseudo random sorting which combines random ranking with a ranking based on the hierarchical merging process.

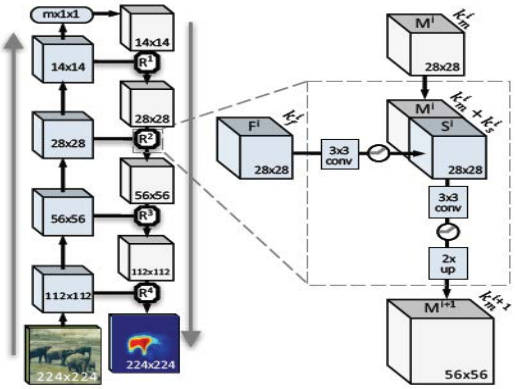


Figure 2.13 – SharpMask framework. Figure taken from [94].

the number of object proposals to 256.

2.6.2 Evaluation of the Context Features

The proposed context features are motivated by the work of [80]. Different from it, our approrach does not use an iterative procedure but is based on object proposals. We add a comparison in Table 2.3 of the performance of our context features against their method on the PASCAL-S dataset. Note that here we only consider our context feature for a fair comparison, and do not use the object feature. We have also included results when only using RGB features, which are the features used by [80]. Our context features clearly outperform the context features based on both RGB and deep features. We have also included timings of our algorithm. Since most of the time was spend by the MCG algorithm (35.3s) we have also included results with the FastMask object proposals (using 8 proposals). In this case the computation of the context features takes (5.4s). Note that this is based on an unoptimized matlab implementation. Also we add a visual comparison between our method and [80] in Fig. 2.16.

Next we compare our context proposals, which follow the object proposal boundary, with different context shapes. We consider rectangular and circular context, which are derived from the bounding boxes based on the object

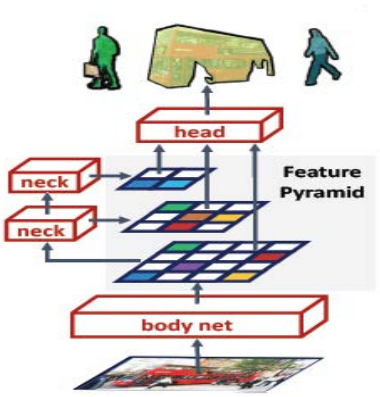


Figure 2.14 – FastMask one-shot structure. Figure taken from [46].

proposals. The context of the rectangular bounding box is computed by considering its difference with a rectangle which is $\sqrt{2}$ larger. In case of the circular context we consider the circle center to have a radius of ($r = \frac{w+h}{4}$) and its context is computed by considering the difference with a radius larger by a factor of $\sqrt{2}$. Like this the context for both the rectangle and the circle has again the same surface area as the object (center). For the three different context shapes we extract the same context features. The results are summarized in Table 2.4 and show that our approach clearly outperforms the rectangular and circular shaped contexts. Thereby showing that accurate context masks result in more precise saliency estimations.

In the following experiment we evaluate the additional performance gain of the saliency features based on context proposals. The results are presented in Table 2.5 for four datasets. We can see that a consistent performance gain is obtained by the usage of context proposals. The absolute gain varies from 0.7 on FT to 1.6 on PASCAL-S. This is good considering that the context feature only has a dimensionality of 3 compared to 512 for the object feature.

2.6.3 Comparison state-of-the-art

The results are presented in Figs. 2.17-2.18 and in Table 2.6. Note that we have only included the curves of the methods in Figs. 2.17-2.18 when this

Number of proposals	8	16	32	64	128	256
SS	59.00	64.60	70.20	74.20	77.50	78.40
GOP	66.20	71.50	73.30	76.30	77.70	79.60
MCG	77.20	77.50	78.60	79.30	80.20	**80.90**
SharpMask	73.79	74.07	73.34	73.15	73.70	74.01
FastMask	75.87	75.03	74.42	74.04	–	–

Table 2.2 – The F-measure performance as the number of proposals evaluated on the PASCAL-S dataset for selective search (SS), geodesic object proposals (GOP), multiscale combinatorial grouping (MCG), SharpMask and FastMask

	feature	proposals	PASCAL-S	Time(s)
Mairon	RGB	-	65.57	140
Our context	RGB	MCG	69.06	40.9
Our context	DF	MCG	74.90	49.0
Our context	DF	FastMask	73.65	6.7

Table 2.3 – Comparison between our context features and the context method proposed by [80] in terms of F-measure and computational speed in seconds. We provide results for our method based on RGB and deep features (DF), and with MCG or FastMask as an object proposal method.

.

data is made available by the authors.

Experiments have been conducted on the PASCAL-S, MSRA-B, FT and ECSSD datasets. Traditionally these datasets proposed an original train and testset split [70]. However, several of these datasets are too small to train deep neural networks. Therefore, methods based on deep learning generally train on the MSRA-B trainset which is the largest available dataset [52, 64, 66]. To be able to compare with all results reported in the literature, we report in Table 2.6 both results; the results trained on the original training set and those based on training on the MSRA-B training set (these results are indicated by an asterix). As an evaluation metric we use the F-measure. We report both qualitative and quantitative comparison of our methods

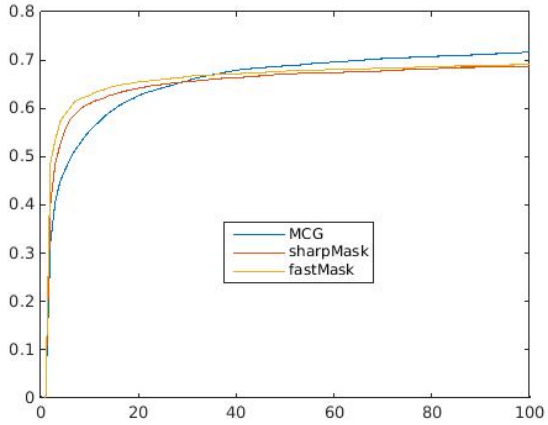


Figure 2.15 – Comparison based on the intersection over union (IoU) with the salient object groundtruth.

with state-of-the-art methods. We also report our results in Figs. 2.17-2.18. Note that these are based on training on the original training set of each dataset. Furthermore, we have only included the curves of the methods in Figs. 2.17-2.18 when this data is made available by the authors.

On the challenging PASCAL-S dataset our method trained on the original dataset obtains an F-measure of 82.3, and is the third method. On the MSRA-B dataset we are outperformed by several recent end-to-end trained saliency methods but still obtain competitive results of 90.9. On the FT dataset

Method	PASCAL-S
Our context features	74.90
Rectangular center surround	67.64
Circular center surround	63.71

Table 2.4 – Comparison between our context shape and the conventional circular or rectangular neighborhood in terms of F-measure.

.

	object	context	object & context
PASCAL-S	80.64	74.90	82.31
MSRA-B	89.90	89.24	90.90
FT	89.80	87.96	91.5
ECSSD	86	82.64	86.90

Table 2.5 – The results on four datasets in F-measure for saliency based only on object proposals, only context proposals and a combination of the two.

we obtain similar to state-of-the-art results when trained on the original dataset, and slightly better than state-of-the-art when trained on the MSRA-B dataset. Finally, on the ECSSD dataset we obtain the best results when considering only those which are trained on the ECSSD training dataset, but are outperformed by recent end-to-end trained networks trained on MSRA-B.

We added a qualitative comparison in Figs. 2.19-2.20. We tested our method in different challenging cases, multiple disconnected salient objects (first two rows), and low contrast between object and background (third and fourth row). Notice that our method correctly manages to assign saliency to most parts of the spider legs. Finally, results of objects touching the image boundary are shown where our method successfully includes the parts that touch the border (last two rows).

2.7 Conclusions

In this chapter, we have performed that the direct context of an object is believed to be important for the saliency humans attribute to it. To model this directly in object proposal based saliency detection, we pair each object proposal with a context proposal. We propose several features to compute the saliency of the object based on its context; including features based on omni-directional and horizontal context continuity. We propose context proposals for center-surround saliency from deep features.

We evaluate several object proposal methods for the task of saliency segmentation and find that multiscale combinatorial grouping outperforms

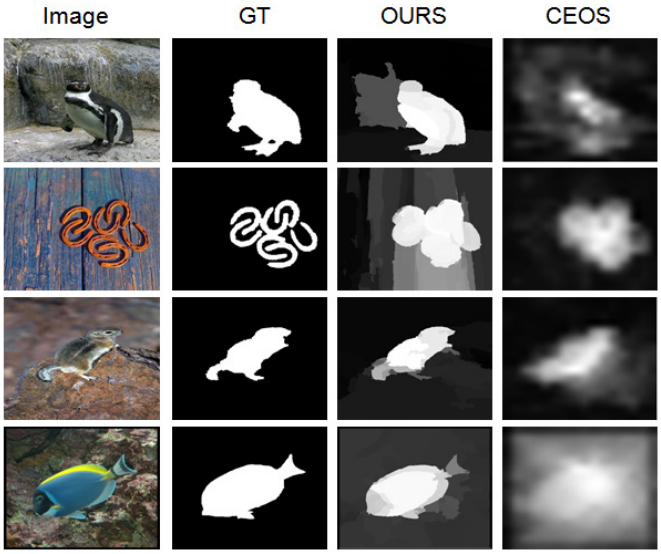


Figure 2.16 – Visual comparison between our method and the method of [80]. Our method results in clearer edges since saliency is assigned to whole object proposals.

selective search, geodesic object, SharpMask and Fastmask. We evaluate three off-the-shelf deep features networks and found that VGG-19 obtained the best results for saliency estimation. In the evaluation on four benchmark datasets we match results on the FT datasets and obtain competitive results on three datasets (PASCAL-S, MSRA-B and ECSSD). When only considering methods which are trained on the training set provided with the dataset, we obtain state-of-the-art on PASCAL-S and ECSSD.

For future research, we are interested in designing an end-to-end network which can predict both object and context proposals and extract their features. We are also interested in evaluating the usage of context proposals for other fields where object proposals are used, notably in semantic image segmentation. Finally, extending the theory to object proposals and saliency detection in video would be interesting [100].

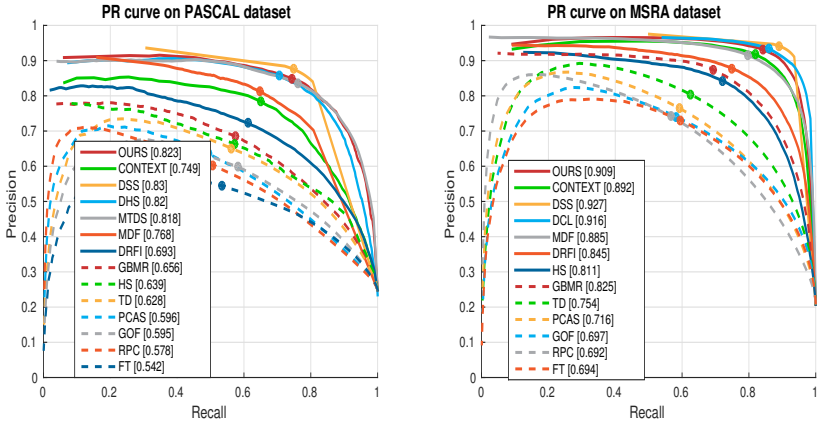


Figure 2.17 – Precision-Recall curves on (left) Pascal-S dataset and (right) on MSRA-B dataset

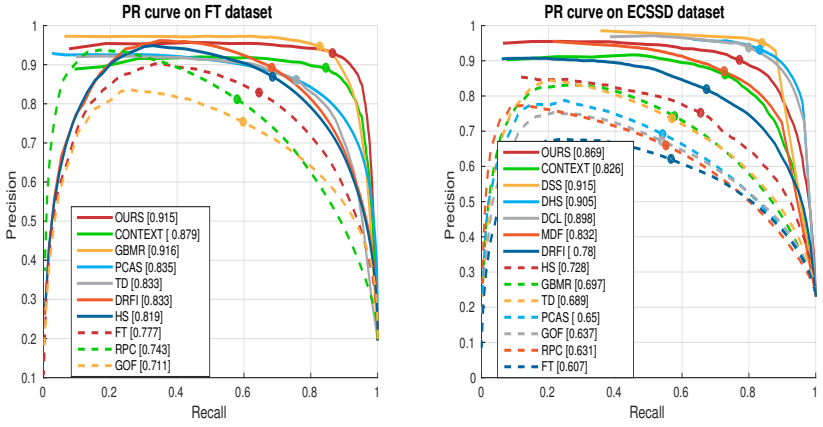


Figure 2.18 – Precision-Recall curves on (left) FT dataset and (right) ECSSD dataset.

	Pascal-S	MSRA-B	FT	ECSSD
FT[2]	54.2	69.4	77.7	60.7
PRC[76]	57.8	69.2	74.3	63.1
GOF[39]	59.5	69.7	71.1	63.7
PCAS [84]	59.6	71.6	83.5	65
TD [101]	62.8	75.4	83.3	68.9
HS[129]	63.9	81.1	81.9	72.8
GBMR [130]	65.6	82.5	91.6	69.7
DRFI[52]	69.3	84.5	83.3	78
LEGS[119]	75.2	87	–	82.5
MC[138]	79.3	–	–	73.2
MDF[66]	76.8*	88.5	–	83.2*
MTDS [69]	81.8*	–	–	80.9*
DHS[72]	82*	–	–	**90.5***
DCL [64]	82.2*	91.6	–	89.8*
RFCN[120]	**82.7***	**92.6**	–	89.8*
DSS[44]	**83***	**92.7**	–	**91.5***
Ours (trained on original trainset)	82.3	90.9	91.5	86.9
Ours (trained on MSRA-B)	78.1*	90.9	**91.8***	85.4*

Table 2.6 – Comparison of our method and context features against state-of-the-art methods. The results are based on training on the original trainset of each datasets. The methods which use the MSRA-B dataset to train are indicated with a $*$.

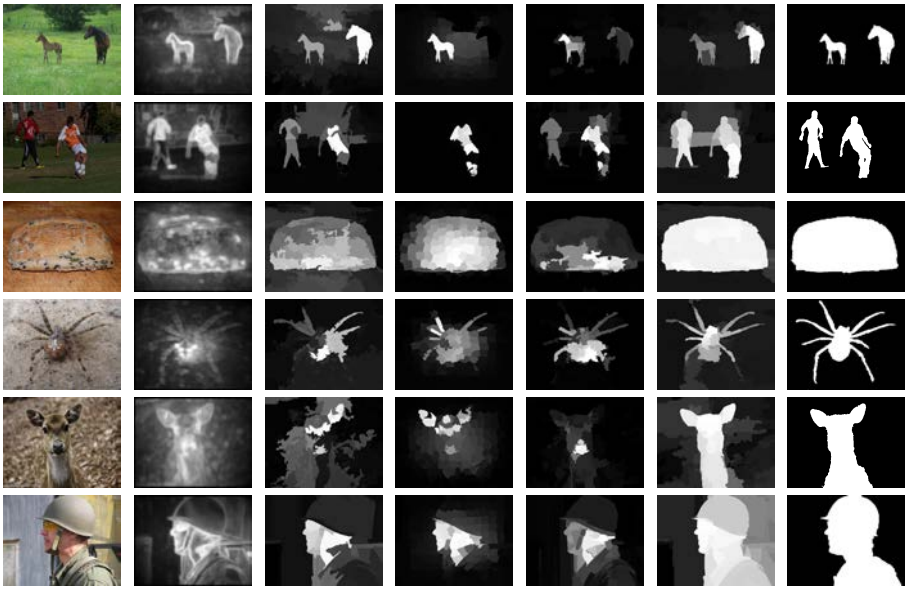


Figure 2.19 – Visual comparison of saliency maps generated from 4 different methods, including our method. Methods for comparison includes DRFI [52], GOF [39], HS [129], and GBMR [130].

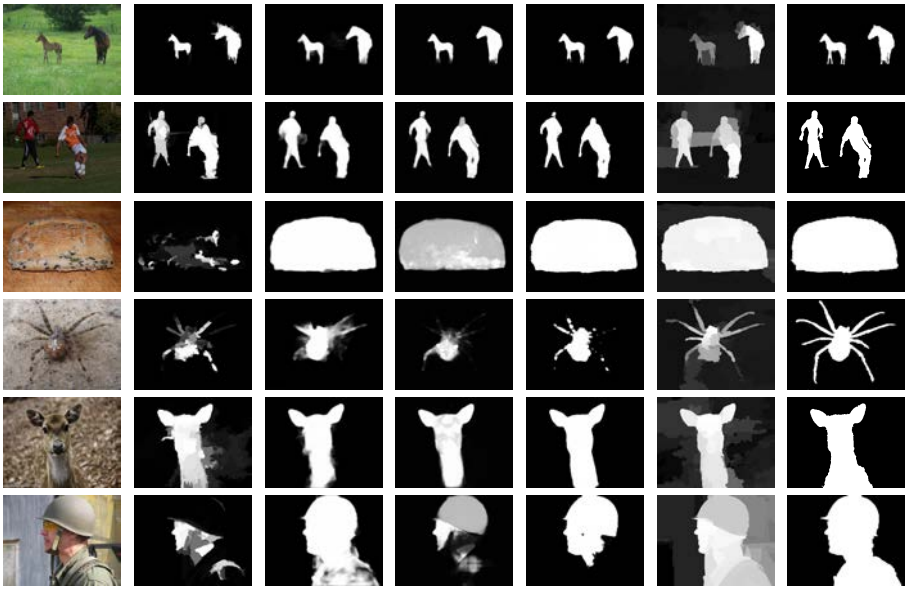


Figure 2.20 – Visual comparison of saliency maps generated from 4 different methods, including our method. Methods for comparison includes DSS [44], DCL [64], DHS [72] and MDF[66].

3 Saliency from High-Level Semantic Image Features[1]

Top-down semantic information is known to play an important role in assigning saliency. Recently, large strides have been made in improving state-of-the-art semantic image understanding in the fields of object detection and semantic segmentation. Therefore, since these methods have now reached a high level of maturity, evaluation of the impact of high-level image understanding on saliency estimation is now feasible. We propose several saliency features which are computed from object detection and semantic segmentation results combined with object proposals and these to a standard baseline method for saliency detection. Experiments demonstrate that the proposed features derived from object detection and semantic segmentation improve saliency estimation significantly. Moreover, they show that our method obtains state-of-the-art results on four benchmark data sets (FT, ImgSal, ECSSD and PASCAL-S data sets).

3.1 Introduction

Saliency is the quality of objects that make them pop-out with respect to others thereby grabbing the viewer's attention. Computational saliency detection can be divided in two approaches: research which aims to estimate the saliency maps obtained with eye-tracking devices on human subjects [22, 104, 135], and work which identifies the salient objects in scenes [10, 140]. The latter is also called saliency object detection and is the focus of this chapter. Computational salient object detection aims to detect the most attractive objects in the image in a manner which is coherent with the perception of the human visual system. Visual saliency has a wide range of applications such as image retargeting [29], image compression [110] and image retrieval [117].

[1] The materials in this chapter are used in a journal submission with the same name by Aymen Azaza, Joost van de Weijer, Ali Douik and Javad Zolfaghari.

Initially, most saliency models were bottom-up approaches which are based on low-level features which are merged using linear and non-linear filtering to get the final saliency map [9, 20]. Itti et al. [49] propose one of the first models for computational visual saliency which is based on the integration theory of Treisman [113] and uses several low-level bottom-up features including color, orientation and intensity. Even though this method has been surpassed on popular baselines by many approaches, a recent study which optimized all its parameters found that it could still obtain results comparable to state-of-the-art [33]. Yang et al. [130] improve low-level features by considering their contrast with respect to the boundary of the image. The boundary is used to model the background. The final saliency map is computed using graph-based manifold ranking. Perazzi et al. [93] apply a Gaussian filtering framework which is based on computing regional contrast, and element color uniqueness to rank the saliency of regions.

Top-down approaches consider that high-level semantic understanding of the image plays an important role in saliency assignment. These methods first identify a subset of high-level concepts, such as faces, text, and objectness, which are detected in the image, and in a subsequent phase are used to compute the saliency map. Judd et al. [53] propose a top-down algorithm to detect objects such as faces, people, text, body parts and animals. Yang et al. [131] propose a top-down approach based on learning a conditional random field and a dictionary. Borji et al. [9] combine low-level bottom-up features with top-down features computed such as face and text. Ehinger et al. [25] compute saliency by combining scene context features, target features and location. Cerf et al. [20] add a face detector to their saliency approach. All of these methods show that adding high-level semantic features to saliency computation improves results significantly.

Convolutional neural networks [59, 60] have significantly improved the state-of-the-art of high-level image understanding. Instead of separately designing hand-crafted features and optimal classifiers for computer vision problems, these networks propose to learn end-to-end, optimizing both the feature representation and the classifier at the same time. These techniques have led to impressive performance gains in semantic image understanding. For example the results for object detection on the popular PASCAL

VOC 2010 data set have improved from 29.6 [30] in 2010 to 68.8 with fast R-CNN in 2015 [37]. Impressive improvements can also be seen for semantic segmentation on PASCAL VOC 2011 from 47.6 in 2012 [18] to 62.7 with fully convolutional networks in 2015 [75]. Given this large improvement in performance we think it is timely to revisit top-down high-level features for saliency. Given the significant improvements of high-level object detection and semantic segmentation, we aim to evaluate the impact of these high-level methods on the task of saliency estimation. An example of the importance of high-level features for saliency is shown in Figure 3.1. As discussed above it is well known that high-level semantic information plays an important role when attributing saliency [20, 79]. However to the best of our knowledge an analysis of the current state of the art methods in object detection on saliency estimation is still missing in the literature. In addition, a recent article titled "Where should saliency models look next?" [17] concluded that models continue to miss semantically-meaningful elements in scenes, these missed elements are parts of people, faces, animals, and text. In this chapter we evaluate the impact of two methods for high-level image understanding, namely object detection and semantic segmentation. The knowledge of these high-level classes, which are from a variety of object groups including humans, vehicles, indoor and animals, is expected to lead to better saliency estimates. We will combine these algorithms with object proposal methods and we will propose several new saliency features based on them. We will perform an extensive analysis on several standard data sets and evaluate the gain which is obtained by having access to this high-level information.

The organization of the chapter is as follow. In Section 3.2, we present the related work. In Section 3.3, we give an overview of the proposed method. In Section 3.4, we describe the features computed from object detection, segmentation results and object proposals. Next we provide details on the experimental setup and results are presented in Section 3.5. Conclusions are provided in Section 3.6.

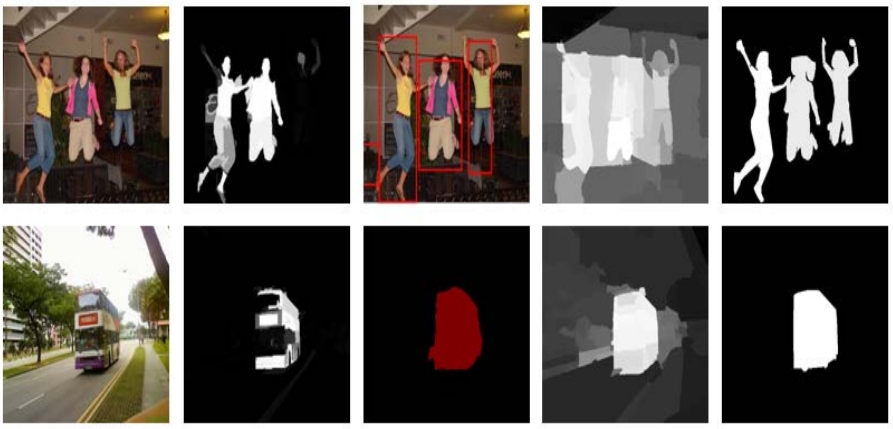


Figure 3.1 – From left to right input image, saliency map by MDF method [66], third column first row object detection results and second row semantic segmentation results, fourth column our saliency map and last column the ground truth. Examples show that high-level features is important for saliency detection.

3.2 Related Work

In this section, we provide an overview of salient object detection methods. After the seminal work of Itti et al. [49], who propose one of the first models for computational saliency, saliency estimation has led to both biologically inspired models [7, 35, 87], and many mathematical motivated methods [2, 41, 45]. Complete reviews on saliency can be found in [11] and [137].

Some models generate a list of bounding boxes with a saliency score [31, 109]. In [31], they propose to select the salient object by ranking bounding boxes with a saliency score. Siva et al. [109] propose an unsupervised method to sample the most important patches in the image via patch features and combine them into object proposals. Alexe et al. [3, 4] measure the objectness of a bounding box by combining local appearance contrast and boundary cues.

Due to its success in object detection, object proposal methods became

a hot topic of research [21, 58, 114, 141]. Recently, these object proposals methods have been applied in the field of saliency detection. The advantage of using object proposals approaches against the methods based on superpixels is that they do not require an additional regrouping step (often implemented with a conditional random field) [73]. Also the use of object proposals methods has another advantage which is avoiding the use of the costly sliding window approaches. The methods which use object proposals include [70, 119], where saliency features are extracted for all object proposals after which a classifier is used to assign saliency to the object proposals. Wang et al. [119] propose local and global deep network for saliency to predict the saliency of each object proposal generated from the Geodesic object proposal method [58]. We use a similar method to [70] as our baseline.

Only few methods have explicitly used high-level object detection as part of the saliency estimation. Xu et al. [127] introduce a visual saliency approach which includes the object and the semantic level information of the pixel level. The object level includes size, convexity, solidity, complexity and the eccentricity. The semantic level has four categories. It includes a category with information which is directly related to humans (e.g., face, emotion, touched, gazed). The second category is motion. The third category focuses on the senses of humans (e.g., sound, smell, taste, touch) and the last category is based on the interaction with humans (text). Nuthmann et al. [89] prove that object is important in leading attention. Einhauser et al. [26] demonstrate that objects predict fixations better than early saliency, so they propose a model based on detecting or segmenting objects to predict salient regions. Other than these methods we consider a wider group of twenty object classes and evaluate their impact on saliency estimation. In addition, we evaluate both the influence of object detection and semantic segmentation for saliency estimation.

The usage of convolutional networks has also quickly been applied to visual saliency research. Initially, several works used off-the-shelf deep features to replace previous hand-crafted features [66, 99, 119, 138]. Further progress was made when fully convolutional networks allowed for end-to-end estimation of saliency [50, 64], which led to convolutional features which were optimized for saliency detection. Li et al. [69] propose a multi-task deep model for semantic segmentation task and saliency prediction task,

they investigate the correlation between the semantic segmentation and saliency detection. They prove that using features collaboratively for two correlated tasks can improve overall performance. In this chapter, we study the influence of state-of-the-art semantic image understanding methods, such as object detection and semantic segmentation, on saliency detection. We use a simple method as a baseline which is not based on deep learning but the method could be extended to include bottom-up deep features.

3.3 Method Overview

The main novelty of our approach is the use of high level semantic information (object detection and semantic segmentation results) for the task of saliency prediction. To evaluate the impact of high-level semantic information on saliency we use a standard saliency pipeline. A similar approach was for example used by Li et al. [69] where they propose a multi-task deep model for semantic segmentation and saliency prediction task.

An overview of the baseline saliency approach at testing time is shown in Figure. 3.2. Given an image we compute a set of object proposals using the multiscale combinatorial grouping (MCG) method [5] (which was proven in chapter 2 to be the best object proposal method for the task of saliency detection). Based on the extracted feature vector for each of the object proposals, we train a random forest for regression to produce a saliency model which will be used for saliency estimation. As the saliency score for each object proposal we use the average saliency of the pixels in the proposal (pixels have a saliency of one if they are on the ground truth salient object or zero elsewhere). At testing time we assign saliency for all the object proposals using the random forest regressor. The final saliency map is computed by taking for each pixel the average of the saliency of all the proposals that contain that pixel.

To incorporate high-level semantic information into the saliency pipeline we only adapt the feature extraction phase of the baseline method. An overview is provided in Figure. 3.3. We will consider two types of high-level semantic information, namely object detection and semantic segmentation results. We will use both systems which are trained on the PASCAL VOC dataset which contains of twenty classes, including humans, animals, vehi-

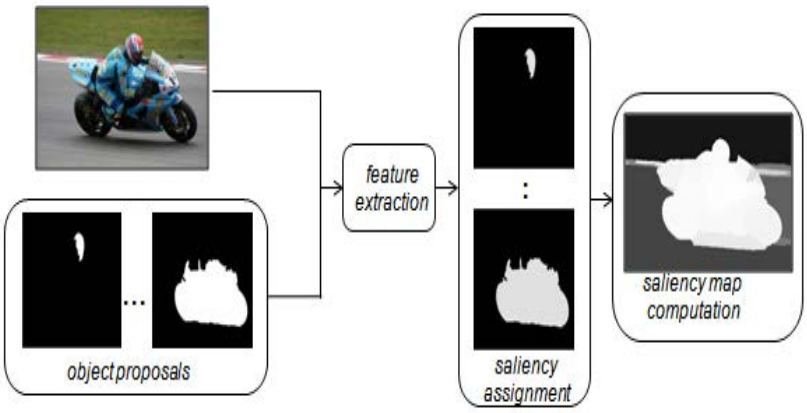


Figure 3.2 – Overview of our proposed method, from the input image we compute a set of object proposals using MCG method, from these objects we compute shape features, object detection and segmentation features, we train a random forest, when testing we assign a saliency score to each object proposal

cles, and indoor objects (the classes of objects are presented in table 3.1). We propose several object detection features which are derived from the detection bounding boxes and the object proposals. Similarly, we derive semantic segmentation features by comparing the semantic segmentation results with the object proposals (see Section 3.4).

Before introducing the high-level features we derive from object detection and semantic segmentation results we shortly describe the standard shape features which are directly computed from the object proposals. We will apply these features in our baseline method and in combination with the semantic features.

Shape features: We extract 17 object proposal features, namely *shape features* which are based on the shape of the binary mask and its position in the image resulting 17 features. The features described here for shape are existing features from saliency literature. For every object proposal we compute a set of shape features similar to the ones proposed in [52, 70]. The shape

Table 3.1 – PASCAL VOC dataset (20 classes)

Vehicles	Animals	Indoor objects	Humans
aeroplane	bird	bottle	person
bicycle	cat	chair	
boat	cow	table	
bus	dog	plant	
car	horse	sofa	
motorbike	sheep	tv	
train			

features we consider are centroid (2 dimensions), area, perimeter, convex area, Euler Number, major axis Length, minor axis Length, eccentricity, orientation, equivalent diameter, solidity, extent, width and height of the object proposal. As an additional shape feature, we add the border-clutter feature [96] which is a binary feature indicating if the object proposal touches the boundaries of the image, and is therefore cluttered by the field of view of the image. We also model the fact that salient objects are more frequent near the center of the image [53, 136]. This feature is modeled by placing a Gaussian in the center of the image (for standard deviation $\sigma_x = width/4$ along the horizontal coordinates and $\sigma_y = height/4$ along the vertical coordinates was chosen). The centrality of object proposals is equal to the average value of the Gaussian over all pixels within the object proposal. It should be noted that for datasets where such a bias does not exist this can be learned by the classifier, and this feature would subsequently be ignored.

3.4 High-level semantic features

The human visual system gives more attention to specific semantic objects classes such as persons, cars etc. In this section, we present high-level semantic features that we extract to compute saliency. These high-level features contain semantic knowledge of the object class. Therefore, the amount of saliency can depend on the semantic class and can be learned in a training phase.

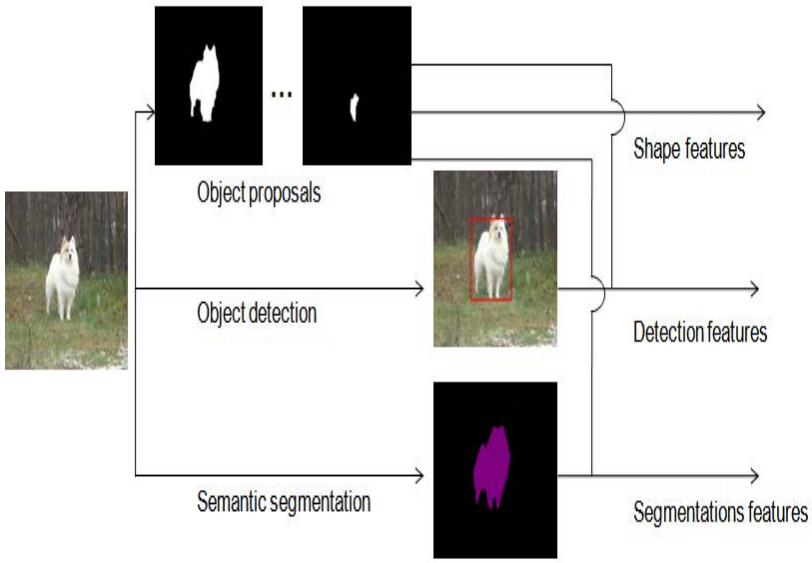


Figure 3.3 – Overview of feature extraction, input image and a set of object proposals are used to compute shape features and combined with object detection and semantic segmentation based saliency features.

Based on the human perception high-level features, such as people, faces and text have been proposed to capture visual attention [9, 20]. As for example [53] which assigns saliency to regions of faces, or the work of [9] which combines low level bottom-up features with top-down features such as text. Other than these works we consider a wider class of objects in the chapter: the twenty classes of the PASCAL VOC which includes persons, animals, vehicles, and objects. Since recently with deep learning the semantic understanding of images has improved significantly and now is of high quality [3, 37], we think it is timely to evaluate the influence of a wider class of objects on saliency.

Object detection features: Here we propose several saliency features derived from object detection results. Object detectors in general detect a

number of bounding boxes in the image. The detection provides a score related to an object class which indicates the confidence of the detector. Often a threshold on the score is defined. Bounding boxes above this threshold are then considered detected objects. The idea here comes from the importance of the semantic information in the object detection results which will help us to detect the salient part in the image. In the pipeline which we described in Section 3.3 the aim is to assign saliency to object proposals. Therefore to exploit high level object detection we have to combine the object detection bounding boxes with the object proposals. To do so, we consider three different features which are all based on the intersection between detection bounding box and object proposals. They differ in the way they are normalized.

As a first measure we consider the popular intersection over union, which is equal to the intersection of the object proposal O_i and the detection bounding box B_i divided by the union between the object proposal O_i and the detection bounding box B_i:

$$ODF_1 = \frac{|O_i \cap B_i|}{|O_i \cup B_i|} \tag{3.1}$$

where $|O_i \cap B_i|$ is equal to the number of pixels in set $O_i \cap B_i$. This measure is typically used in the evaluation of semantic segmentation [28].

The second measure computes the intersection over the minimum of the detection bounding box B_i and the object proposal O_i:

$$ODF_2 = \frac{|O_i \cap B_i|}{\min(O_i,\, B_i)} \tag{3.2}$$

and is sometimes considered as an alternative for intersection over union [102].

A drawback of the first measure is that in case the object proposal is part of the bounding box, but a significant part of the bounding box is outside the object proposal, this measure will assign a low saliency. The second measure addresses this problem, however when the bounding box is included in the object proposal, this measure will assign a high saliency to the whole object proposal, even though the bounding box might only be a small part of the object proposal. Both these problems are addressed by the third measure

which computes the percentage of pixels in object proposal O_i which are in the detection bounding box B_i:

$$ODF_3 = \frac{|O_i \cap B_i|}{|O_i|} \tag{3.3}$$

An example of the object detection features computation is shown in Figure 3.4. Comparison of object detection features computed on three example images (top row) from an example object proposal and object detection bounding box (bottom row). Superposed on the images in the bottom row are the object detection features. In these three examples the saliency which should be assigned to the object proposal is high for the first two images and low for the last example. Only the third object detection feature correctly correlates with this.

It should be noted that we compute the equations Eq.3.1–3.3 with the object proposal mask and with the bounding box representation for the detection. One could also decide to represent the object proposal with a bounding box, by drawing the smallest enclosing bounding box around the object proposal. Again the same three features could be computed but now based on the bounding box for O_i. We compared both approaches on the PASCAL-S dataset and report the F-score in Table 3.2. One can observe that using the original object proposal obtains better results than using bounding boxes. In addition, we see that the best results are obtained with object detection feature ODF_3. In all our experiments we combine the three measures based on segmentation masks into the final ODF feature.

Table 3.2 – Comparison of detection features on the segmentation mask and the bounding box representation in terms of F-score.

Features	ODF_1	ODF_2	ODF_3
segmentation mask	45.40	58.90	64.30
bounding box	45.10	30.10	53.30

Segmentation Features: As second feature for high-level information, we use semantic segmentation results. Semantic segmentation algorithms yield

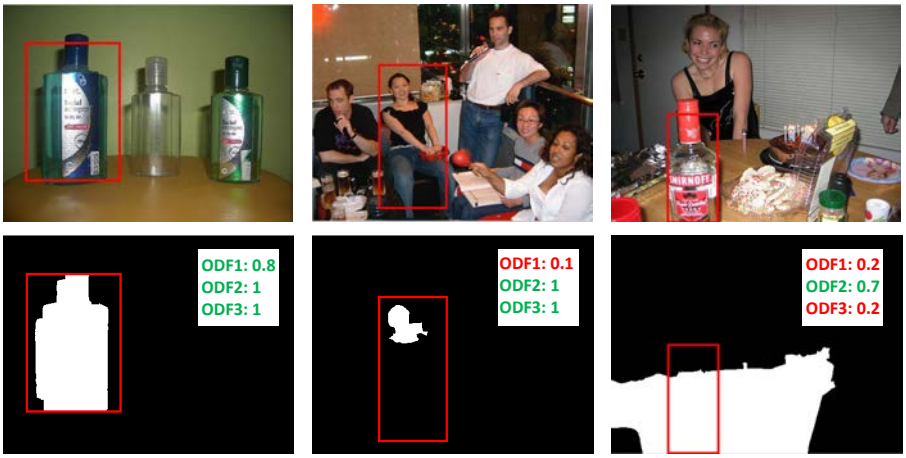


Figure 3.4 – Example of object feature computation for three example images. See text for details.

a probability map of the same size as the input image. For each pixel it provides the probability that it belongs to one of the semantic classes. Typically a background class is introduced for all pixels which do not belong to any of the semantic classes. Semantic segmentation can be considered a more difficult task than object detection because for good results the exact borders of objects need to be correctly detected.

We use the semantic segmentation results to propose a semantic segmentation feature (SSF) for saliency. For every semantic class c and object proposal O_i we compute the SSF according to:

$$SSF(c) = p(c|O_i) = \frac{\sum_{x_j \in O_i} p(c|x_j)}{|O_i|} \tag{3.4}$$

where $p(c|x_j)$ is the output of the semantic segmentation algorithm and provides the probability of a semantic class conditioned on the pixel location. In this chapter we will evaluate a semantic segmentation features derived from algorithms trained on VOC PASCAL, which has 21 classes, and therefore the SSF feature of each object proposal will also have a dimensionality of

21).

3.5 Experiments and Results

In this section, we provide the implementation details, the experimental setup that we use in our approach, the benchmark datasets and the evaluation metrics.

3.5.1 Implementation details

The overall pipeline of our method is provided in Section 3.3. Here, we report the implementation details.

Object proposals generation: From the input images we compute a set of object proposals using the multiscale combinatorial grouping (MCG) method [5]. This method is based on a bottom-up hierarchical image segmentation. It was found to obtain improved results compared to other object detection methods [58, 114]. It is proven that MCG method is the best object proposal method for the task of saliency detection (see section 2.6.1 in chapter 2). We use the algorithm with the default settings. With these settings the method generates an average of 5153 object proposals per image.

Object detection: To generate the object detection bounding boxes we used the fast R-CNN of Girshick [37]. Fast R-CNN is an improved version of the R-CNN [38]; it obtains a significant speed-up by sharing the computation of the deep features between the bounding boxes. We use the fast R-CNN detector [37] and which is trained on PASCAL VOC 2007. The architecture of R-CNN method is provided in Figure 3.5.

Segmentation results: For the semantic segmentation we use the algorithm proposed by Long et al. [75]. They compute the segmentation maps with a fully convolutional neural network (FCN) using end-to-end training. They improve the accuracy of their approach by using features extracted at multiple scales and adding skip connections between layers. We used the code provided by [75] and trained on the 20 classes of the PASCAL VOC 2007.

Random Forest: To assign saliency to every object we used random forest and we set the number of trees to 200. As a protocol we train on 40 % of the images, and test on the 60% of images which were not included in the

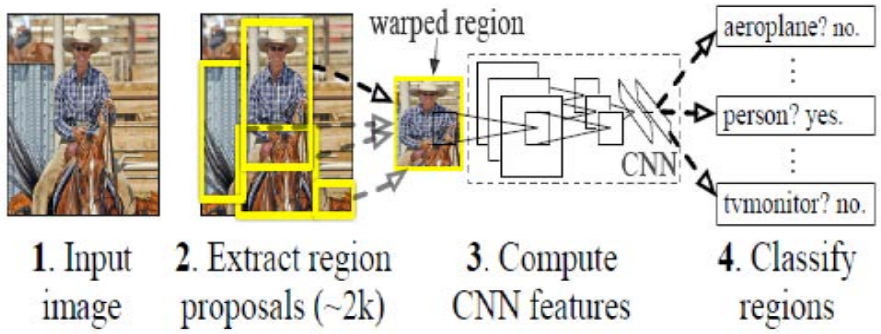


Figure 3.5 – Architecture of Fast R-CNN. Figure taken from [37].

training set.

Saliency Features: We will compare results of several different saliency features. As a baseline we will only use the shape features (SF) explained in Section 3.3. With ODF we indicate the method which is only based on the the object detection features and with SSF the method which only uses the semantic segmentation features. Combinations of features are indicating as e.g. SF&ODF for joining shape feature and object detection features.

3.5.2 Experimental setup

Datasets

To evaluate the performance of the proposed method, we provide both qualitative and quantitative results on four benchmark datasets: FT [2], ImgSal [67], ECSSD [129] and PASCAL-S [70]. The FT dataset contains 1,000 images, most of which have one salient object. It provides the salient object ground truth which is provided by [126]. The ground truth in [126] is obtained using user-drawn rectangles around salient objects. The ImgSal dataset contains 235 images collected from the internet. It provides both fixations as well as salient object masks. The ECSSD dataset contains 1,000 images. It is obtained by collecting images from the internet and PASCAL VOC, the ground truth masks are labeled by 5 subjects. The PASCAL-S dataset contains 850 images and was built on the validation set of PASCAL VOC which has 20

classes of objects. In contrast to the other datasets it often contains more than one salient object. All the datasets contain manually labeled ground truth.

Evaluation metric

We evaluate the performance of our method using two metrics which are F-measure and Precision Recall curve (PR). The PR curves are computed by binarizing the saliency map at different thresholds and comparing it to the ground truth mask. The F-measure is defined in Eq.2.20.

We conduct a qualitative and quantitative comparison of our method against the following methods: a context aware method [39] (GOF), Multi task deep saliency [69](MTDS), discriminative regional feature integration (DRFI) [52], frequency tuned saliency (FT) [2], graph-based manifold ranking (GBMR) [130], local and global estimation (LEGS) [119] hierarchical saliency (HS) [129], multiscale deep features (MDF) [66], regional principal color based saliency detection (RPC) [76], Principal component analysis saliency (PCAS) [84] and textural distinctiveness (TD) [101].

3.5.3 Results

We start by evaluating the additional gain obtained when adding object detection features (ODF) and semantic segmentation features (SSF). The results for the four datasets are provided in Table 3.3 and Table 3.4. When we look at the performance of ODF and SSF alone, we observe that semantic segmentation provides much better features for saliency detection than object detection. We think that this is caused by the fact that segmentation algorithms provide pixelwise results rather than bounding boxes, and therefore the saliency feature computation for each object proposal is more accurate.

Next we consider the absolute gain which is obtained by adding ODF and SSF features to our baseline method (indicated by SF). For both features, and on all four datasets, the features provide a significant improvement. This clearly shows the importance of high-level semantic features for saliency assignment. Again the improvement is largest when adding features derived from semantic segmentation. The best results are obtained on PASCAL-S

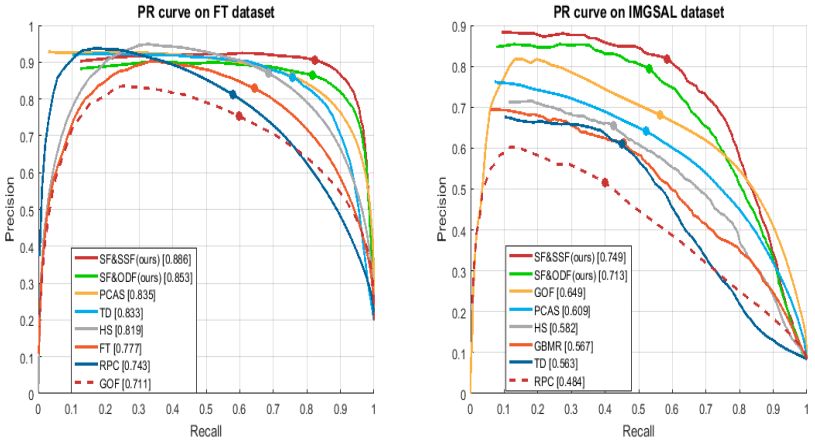
dataset where a gain of over 11% is reported. This is partially caused by the fact that the object detector and the semantic segmentation algorithm have been trained on the PASCAL VOC dataset [2]. The images in the PASCAL-S dataset contain therefore always classes which are detected (or segmented) by these algorithms. However, on the other datasets, especially on ImgSal and ECSSD also large improvements of around 7% are obtained.

To get a better insight in which classes contribute to the improvement in saliency detection, we have performed an additional experiment for the method based on SSF. We add an analysis to investigate which semantic classes are important. We evaluate the drop of saliency if we remove one class. The results show that removing both bird and person significantly deteriorates saliency estimates on all four datasets. Some other classes contribute only on a part of datasets, such as aeroplane, bicycle, potted plant, sofa and tv-monitor also lead to a drop of over 0.6 when removed for some dataset. Removing some classes actually leads in some cases to a small increase in performance. Possibly caused to overfitting or noise in the semantic segmentation algorithm.

Next we compare our method to state-of-the-art saliency detection methods. The proposed method matches or outperforms clearly all the state-of-the-art methods used in comparison in the four datasets in terms of PR curves and F-measure. In Figure. 3.6 the PR curves on FT and ImgSal datasets are provided. On the FT dataset we clearly outperform all the other salient object detection methods with both object detection and segmentation features. Also we obtain the best F-measure compared to other state-of-the art methods. On the ImgSal dataset we have also the best F-measure. The performance is better over a wide range of recalls only to be slightly outperformed for the highest recalls by GOF.

In Figure. 3.7 the PR curves on the ECSSD and PASCAL-S datasets are reported. On the ECSSD dataset we are the first using the segmentation based saliency features. The saliency method derived from object detection features is fourth after MDF, LEGS and MTDS. Similar results are obtained on the PASCAL-S dataset. Here we significantly outperform all the methods using the segmentation and object detection results. Only the MDF and

[2] None of the images used for training are included in the PASCAL-S dataset.



a b

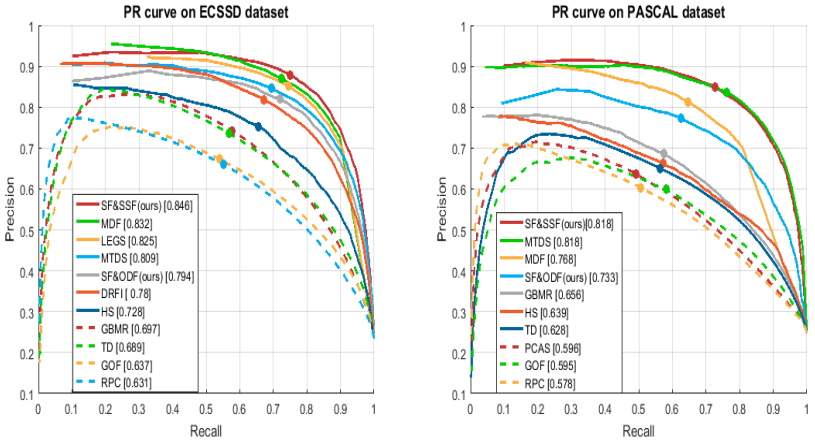
Figure 3.6 – PR curves for a variety of methods on (a) the FT dataset and (b) the ImgSal dataset

MTDS methods outperforms our object detection based method.

We provide a qualitative comparison in Figures. 3.9-3.10. We tested our method in several challenging cases, low contrast between object and background (first two rows), results of objects touching the image boundary are shown where our method successfully includes the regions that touch the border (third and fourth row). Finally, the case when multiple disconnected objects is investigated (last two rows).

3.6 Conclusions

The importance of high-level semantic image understanding on saliency estimation is known [20, 79]. However, most computational methods are bottom-up or only include few semantic classes such as faces, and text [9, 53]. Therefore, we have evaluated the impact of recent advances in high-level semantic image understanding on saliency estimation. To do this, we have derived saliency features from two popular algorithms, namely fast-RCNN



a b

Figure 3.7 – PR curves for a variety of methods on (a) the ECSSD dataset and (b) the Pascal-S dataset

and a fully convolutional approach to semantic segmentation. We found that the features based on semantic segmentation obtained superior results, most probably due to the fact that they provide pixel wise labels, which lead to more accurate saliency estimation maps.

To evaluate the derived features from object detection and semantic segmentation, we perform experiments on several standard benchmark datasets. We show that a considerable gain is obtained from the proposed features and we examine which semantic class boost more the task of saliency. We found that the classes of person and bird are among the most important. In the evaluation on four benchmark datasets we outperform state-of-the-art on four standard benchmark datasets (FT, ImgSal, ECSSD and PASCAL-S).

For future work, we are interested in extending current end-to-end networks for saliency with explicit modules for object detection, and evaluate if such architectures could further improve state-of-the-art. It would also be interesting to evaluate the impact of a larger set of object classes on saliency detection (currently we evaluate the 20 classes from the PASCAL VOC chal-

Table 3.3 – F-measure of baseline (SF) and object detection feature (ODF), their combination and the absolute gain obtained by adding semantic object detection features.

	SF	ODF	SF & ODF	gain ODF
FT	84.23	57.26	85.30	1.07
ImgSal	67.19	54.76	71.30	4.11
ECSSD	77.47	64.57	79.40	1.93
Pascal-S	70.40	66.84	73.32	2.92

Table 3.4 – F-measure of baseline (SF) and semantic segmentation feature (SSF), their combination and the absolute gain obtained by adding semantic segmentation features.

	SF	SSF	SF & SSF	gain SSF
FT	84.23	85.84	88.60	4.37
ImgSal	67.19	73.47	74.90	7.71
ECSSD	77.47	82.21	84.60	7.13
Pascal-S	70.40	81.16	81.80	11.40

lenge). Finally, we evaluated the impact of high-level information on salient object detection, but it would also be interesting to perform a similar study for saliency maps derived from eye-tracking experiments.

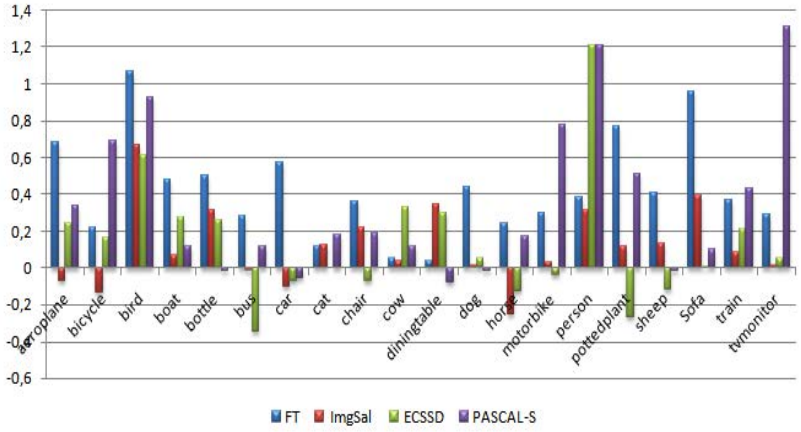


Figure 3.8 – Saliency drop as a consequence of removing a single semantic class on the four datasets from left to right FT, ImgSal, ECSSD and Pascal-S dataset

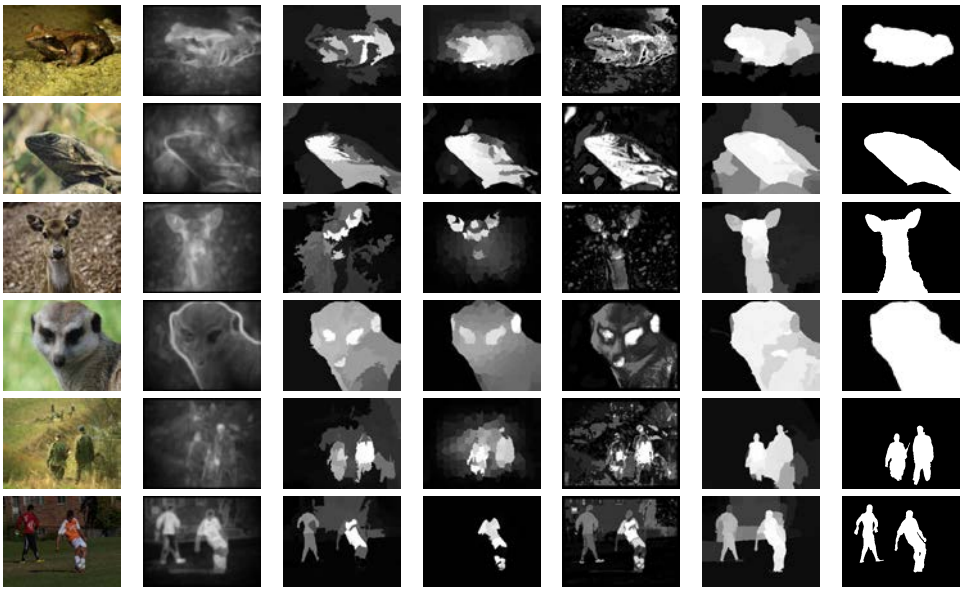


Figure 3.9 – *Qualitative comparison of saliency maps generated from 4 state-of-the-art methods, including our method. Methods for comparison includes GOF [39], HS [129], GBMR [130] and TD [101].*

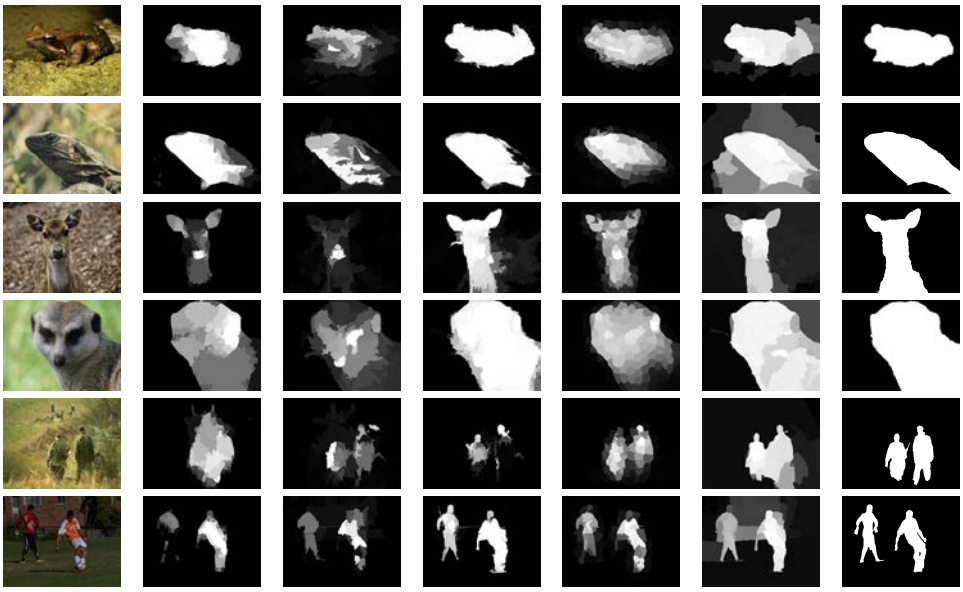


Figure 3.10 – *Qualitative comparison of saliency maps generated from 4 state-of-the-art methods, including our method. Methods for comparison includes LEGS [119], DRFI [52], MDF [66], and MTDS [69].*

4 Context Proposals for Salient Object Segmentation in Videos[1]

Salient object segmentation in videos is generally broken up in a video segmentation part and a saliency assignment part. Recently, object proposals, which are used to segment the image, have had significant impact on many computer vision applications, including image segmentation, object detection, and recently saliency detection in still images. However, their usage has not yet been evaluated for salient object segmentation in videos. Also the importance of context has not yet been evaluated for saliency in video.

Therefore, in this chapter, we investigate the application of object proposals to salient object segmentation in videos. We evaluate several motion features for saliency. In addition, we extend the proposed context approach in still images developed in chapter 2 to detect saliency in video. Experiments on two standard benchmark datasets for video saliency show that object proposals are an efficient method for salient object segmentation. Moreover, we prove that the proposed features computed from the context proposals are the single most important feature for saliency in video, outperforming features based on motion and shape. Results on the challenging SegTrack v2 and Fukuchi benchmark data sets show that we significantly outperform the state-of-the-art.

[1] The materials in this chapter are partially based on a joint journal submission (currently under review): Rahma Kalboussi, Aymen Azaza, Joost van de Weijer, Mehrez Abdellaoui and Ali Douik 'Object Proposals for Salient Object Segmentation in Videos'. The theory of that joint research has been further extended in this chapter with the context proposals of chapter 2.

4.1 Introduction

One of the aspects of human vision, which is extensively studied, is visual saliency. Research on saliency addresses the question: how does our brain assign saliency to the objects (regions) in the scene? This research tries to discover what features – like edges, borders, or colors – are important for the assignment of saliency. Computational saliency can be divided into two categories. The first category focuses on the prediction of eye-fixations [16, 53]. The second category focuses on the segmentation of salient objects in a scene [33, 70, 119, 124]. It is important to distinguish this from image segmentation [19, 58], where we want to segment all objects in the scene, whereas in salient object segmentation only the salient objects need to be segmented. In this chapter, we focus on the latter category of salient object segmentation.

Saliency detection methods for still images are numerous [10, 11]. The field of video saliency is still in its early stages. Similar as for still images, computational video saliency is divided into methods which predict eye-fixations on videos [82, 118], and methods which segment the salient objects in videos [86, 124, 139]. Here we focus on *salient object segmentation* in videos, where we aim to segment the salient object from its background. Such segmentation can be used in divers applications such as object detection, activity recognition, human-computer interaction, video compression and summarization, content based retrieval, image quality assessment, photo collage, media cropping, thumb-nailing, re-targeting and visual tracking.

The problem of salient object segmentation is generally broken up into two parts: segmentation and saliency assignment [124]. In the first part, the image or video is segmented in larger regions which are expected to belong to the same object, and therefore have similar saliency. In the second part, features are extracted and then a classifier is applied to assign saliency to all segments. For videos several motion segmentation methods have been proposed. Some analyze trajectories of a given point in order to extract its motion information [15, 32, 62]. Brox et al. [15] propose a method that consists in building an affinity matrix between each pair of trajectories. Lezama et al. [62] compute motion vectors from the previous

and next frames. Then they group pixels that have coherent motion together. Fragkiadaki et al. [32] detect embedding density discontinuities between trajectories that have a common spatial neighboring. Several works [61, 134] have shown remarkable weaknesses of these techniques, related to occlusion, initialization, complexity of computational models and lack of prior information to elaborate a robust object segmentation.

Lately, object proposals [19] have been applied to correct the aforementioned flaws related to traditional segmentation methods. Most semantic segmentation methods were predominantly based on superpixels. Superpixel approaches provide an over-segmentation of the image into smaller non-overlapping regions. These methods were also very popular in image saliency [56, 130]. However, because these superpixels are rather small, an additional step is required to impose spatial coherence, e.g. by means of a conditional random field [108]. The advantage of object proposal methods over these superpixel methods is that they directly provide possible object proposals, and therefore do not require an additional step to impose spatial coherence. Because of this advantage, object proposal methods have been popular in recent years for saliency detection [70, 119]. However, video saliency methods are still based on superpixels [74, 108, 124].

As shown in Chapter 2 context provides important features for saliency assignment. To the best of our knowledge the importance of context features in video has not yet been evaluated. Similarly as for still images, objects in video often occlude a background; the object is moving in front of a still (in case of fixed camera) or moving (in case of moving camera) background. Again context features could be used to see if the features on different sides of the objects look similar, which would indicate that object is occluding a continuous background. Therefore, we are interested to investigate the importance of context features for video saliency detection.

We will start by evaluating the effectiveness of object proposals for video saliency detection. The first step consists in the extraction of object proposals from each video frame using a spatio-temporal object proposals method. Then, to compute the importance of context for video saliency, we compute a set of context proposals which is the direct context of each object proposals. After that, for each object proposal, a set of features, namely shape, motion and context features are extracted. These features are subsequently

used to train a random forest regressor to predict the saliency. Our main contribution is the evaluation of context proposals for video saliency. For the evaluation we use two standard benchmark data sets for video saliency, namely the SegTrack v2 dataset [63] and Fukuchi dataset [34]. We report state-of-the-art results on both these datasets.

This chapter is organized as follows. In section 4.2 we discuss related works. In section 4.3 we will explain the main steps for saliency feature generation and saliency computation. Then, we will introduce the usage of our context features for video saliency detection in section 4.4. We will discuss experimental results in section 4.5. Finally, conclusions are provided in section 4.6

4.2 Related Work

In this section, we review the main methods on video saliency, for excellent reviews of saliency methods in still images we refer to [10] and [11].

While there are numerous image saliency methods, video saliency detection are still in their early stages. Itti et al. [48] defined a salient event as an important event that can stimulate attention, such events are called surprise. They developed a model which computes immediate low-level events at each location in a video sequence. Rahtu et al. [98] proposed a saliency model for both natural images and videos which combines local features and a statistical framework, with a conditional random field model. Jiang et al. [51] proposed a salient object segmentation algorithm that includes bottom-up salient stimuli and an object-level shape prior, which assumes that each salient object has a precise closed boundary. Mancas et al. [81] introduced a motion selection in crowd model where optical flow is used to determine the motion region. A spatio-temporal visual saliency model was proposed where the final saliency map is the fusion of static and dynamic saliency maps [97]. Goferman et al. [39] proposed a model that considers local and global surroundings of an object to compute its saliency level, so that the whole context of the input image is considered. Zhong et al. [139] proposed a novel video saliency model based on the fusion of a spatial saliency map which is inherited from a classical bottom-up spatial saliency model and a temporal saliency map issued from a new optical flow model. Wang

et al. [124] propose a video saliency object segmentation model based on geodesic distance where both spatial edges and temporal motion boundaries are used as foreground indicators. Also, Wang et al. [125] use a geodesic Bayesian model to produce a saliency map. Mauthner et al. [86] proposed an approach based on the Gestalt principle of figure-ground segregation for appearance and motion cues. Singh et al. [108] presented a method that extracts salient objects in video by integrating color dissimilarity, motion difference, objectness measure, and boundary score feature. They use temporal superpixels to simulate attention to a set of moving pixels.

There is a vast literature on object proposals in still images (e.g. [19], and [114]), however there is relatively little work on object proposals in video. In their paper, Lee et al. [61] discover persistent groups of appearance and motion by computing series of key-segments. Then, for each discovered hypothesis, through all frames, a pixel-level object labeling is estimated. Ma et al. [78] propose a method based on finding a maximum weight clique in a weighted region graph where a region with maximum weight clique has a high objectness score. Zhong et al. [134] proposed a segmentation method based on object proposals which consists on selecting primary object segments in each video frame. Then use these object regions to build object models. Oneata et al. [90] present a spatio-temporal object proposal method based on a hierarchical clustering of superpixels in a graph which considers both spatial and temporal connections. They obtain state-of-the-art results for 3D segmentation, and we therefore selected that method for our saliency detection method.

A similar pipeline as the one we use here for video saliency detection has been introduced in several saliency approaches for still images, but it has never been combined with object proposals for video saliency. Wang et al. [122] used global features extracted from divers saliency indicators to train a random forest model which is used for saliency prediction in web images. In their paper, Nah et al. [88] used features extracted from sampled image patches to derive a random forest model which will be used to predict whether a patch is salient or not. Jiang et al. [52] assigns a saliency score to every image region using a feature vector which is called discriminative regional feature integration by using a random forest regressor. None of these methods is applied to video, and therefore they are based on another

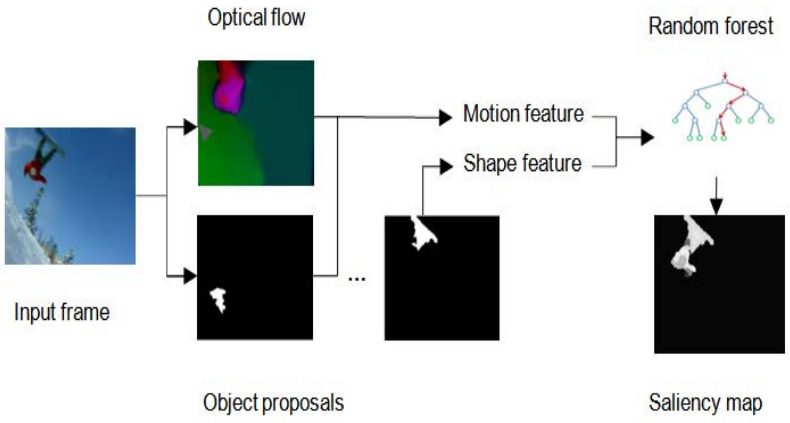


Figure 4.1 – Overview of object proposal based saliency detection in video. A set of object proposals for the video is computed. Based on these proposals we directly compute shape features. By combining the object proposals with the optical flow estimation we derive motion features. A random forest is applied to assign saliency to each object proposal. These are combined in the final saliency map.

set of features.

4.3 Object Proposal based video saliency

In this section we provide a baseline method for object proposal based saliency detection in video. We will extend this method with context proposals in the next section.

Object proposals in videos consist of a set of supervoxels grouped together and hypothesized to contain a single object [90]. The advantage of object proposals is that no spatial reasoning is required as a post-processing step, something which is required for superpixels, and would typically be done by a conditional random field (CRF) which promotes label smoothness [71].

Given an input video frame, we use the spatial-temporal object proposal

method of [90] to get a set of object proposals (see Figure 4.1). For each object proposal we extract a set of features (discussed in Section 4.3) including shape, and a motion feature. The motion feature is computed by combining the optical flow measurement with the object proposals. The motion features addresses the fact that one should not only consider the motion of the region, which could be caused by camera motion, but one should focus on the presence of multiple motions within local regions. In the following sections we will detail the computation of the static and motion features.

4.3.1 Static saliency features

We will consider **shape features** which are directly computed from the binary object proposals. We combine several shape feature which were successfully applied to saliency in still images [19, 53, 70]. The shape feature which we compute include perimeter, area, Euler Number, major Axis Length, convex Area, minor Axis Length, eccentricity, orientation, centroid (2 dimensions), equivalent diameter, solidity, extent, width and height of the object proposal mask. It should be noted that the shape features are the same which we used in our saliency approach in chapter 3.

4.3.2 Dynamic Saliency Features

It is well known that the movement of objects is one of the important features to which saliency is attributed by humans [40]. Therefore, in video saliency, it is important to identify the dynamic objects in the scene. In this section we describe a saliency feature derived from the optical flow estimation in the scene.

To identify the dynamic object in the scene, we compute the optical flow. Consider that the optical flow is given by $u(x,y)$ and $v(x,y)$, respectively the movement in the x and y direction. For its computation, we use the Lucas-Kanade optical flow estimation method [77]. We define the optical flow magnitude to be equal to

$$M\left(x,y\right)=\sqrt{u^2\left(x,y\right)+v^2\left(x,y\right)}. \tag{4.1}$$

It is important to note that the optical flow magnitude in itself is not a

good saliency measure, since camera movement also results in optical flow for all pixels in the image. Therefore, prior work [92, 108] has focused on detecting **motion boundaries**, which can be found by assessing the change in optical flow:

$$B(x,y) = \sqrt{u_x^2(x,y) + u_y^2(x,y) + v_x^2(x,y) + v_y^2(x,y)} \tag{4.2}$$

where $u_x(x,y)$ is the derivative in the x-direction of the optical flow in the x-direction, etc.

The quantity $B(x,y)$ is averaged for all pixels in a superpixel and used as the motion feature in the work of [108]. Papazoglou and Ferrari [92] consider that this measure is subject to noise and combine it with a measure which computes the variation in angular direction of the optical flow. This has the problem however that the derivative of the angular direction of the optical flow can also be noisy or unstable, e.g. when M is small. We therefore consider an alternative approach to assessing the presence of a moving object. Consider the optical flow structure tensor given by:

$$\mathbf{G}(x,y) = \begin{bmatrix} \overline{u(x,y) \cdot u(x,y)} & \overline{u(x,y) \cdot v(x,y)} \\ \overline{v(x,y) \cdot u(x,y)} & \overline{v(x,y) \cdot v(x,y)} \end{bmatrix} \tag{4.3}$$

here $\bar{\cdot}$ refers to a local averaging operation for which we use a standard Gaussian kernel. The structure tensor was originally proposed by [8] and also is derived by [106] within the context of optical flow, and by [115] for the computation of color features. Here we use it to describe the local optical flow gradient field.

We can now compute the two eigenvalues of the structure tensor with:

$$\begin{aligned} \lambda_1 &= \tfrac{1}{2}\left(\overline{u^2} + \overline{v^2} + \sqrt{\left(\overline{v^2} - \overline{u^2}\right)^2 - 4\overline{u \cdot v}^2}\right) \\ \lambda_2 &= \tfrac{1}{2}\left(\overline{u^2} + \overline{v^2} - \sqrt{\left(\overline{v^2} - \overline{u^2}\right)^2 - 4\overline{u \cdot v}^2}\right) \end{aligned} \tag{4.4}$$

where we omitted the spatial coordinates x and y for brevity. Here λ_1 is the energy of the optical flow in the main orientation of the local region and λ_2 is the optical flow energy perpendicular to the main orientation. We use this

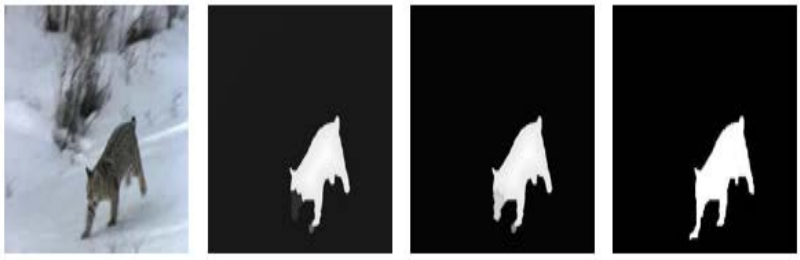


Figure 4.2 – Motion map, from left to right: input frame, motion map, estimated saliency map, ground truth.

feature to detect moving objects. Camera motion is expected to be locally in a single direction, and therefore a high λ_2 reveals the presence of a second motion direction within the local region, which could indicate the presence of a moving object[2]. We found that scaling them to the range of $[0,1]$ slightly improved results, and hence we consider:

$$\hat{\lambda}_2 = 1 - \exp\left(\sqrt{\lambda_2}\Big/\sigma\right) \tag{4.5}$$

In addition we added the motion boundary strength of Eq. 4.2 according to

$$\hat{B} = 1 - \exp\left(B/\sigma_B\right) = 1 - \exp\left(\sqrt{\lambda_1 + \lambda_2}\Big/\sigma_B\right) \tag{4.6}$$

We found that $\sigma = \sigma_B = 6$ worked fine for all datasets. To compute the two dimensional motion feature for an object proposal we average the value of $\hat{B}$ and $\hat{\lambda}_2$ for all pixels in the object proposal. Fig. 4.2 shows an example of the importance of the motion feature for saliency assignment.

In conclusion, other than prior work [92, 108] we derive the motion feature directly from the optical flow by using the optical flow structure tensor and do not have to revert to the derivative of the optical flow. This has the advantage that it is not based on the more noisy second order derivative.

[2]Note that you cannot know with certainty which of the λ's is related to the object and which to the background, but you do know that a high λ_2 indicates the presence of two local motion vectors.

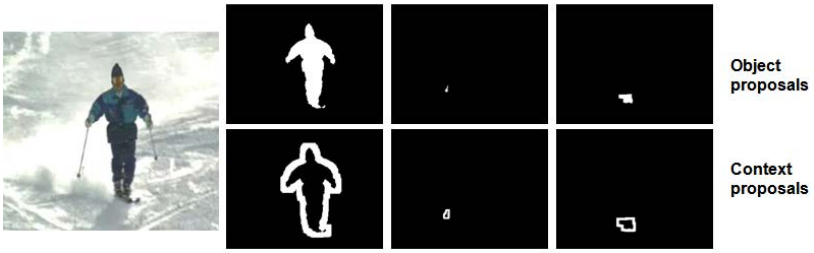


Figure 4.3 – Example of context proposals in video. (top) object proposals; (bottom) context proposals.

In the experiments we will compare this motion feature with these previous methods.

4.4 Context Proposals for Video Saliency

In the previous section we described our baseline method for saliency detection based on object proposals in video. We will here extend the method with context proposals.

An overview of the proposed method for context proposals in video is proposed in Section 4.4. From the object proposals, we compute a set of context proposals which we use to compute the context features. Then for each object proposal, the shape, motion and context features are combined in a single feature vector. Then, a regression method is trained over the training dataset, which maps the feature vector to a saliency score. The ground-truth saliency score of object proposals is given by the ratio of pixels in the object proposal which are considered salient by the ground-truth divided by the total number of pixels in the object proposal. We apply a random forest algorithm [14] for regression. At testing time, we extract the features for all object proposals and apply the random forest to assign a saliency estimate to each proposal. The final saliency map is computed by taking for each pixel the mean of the saliency of all the proposals that contain that pixel.

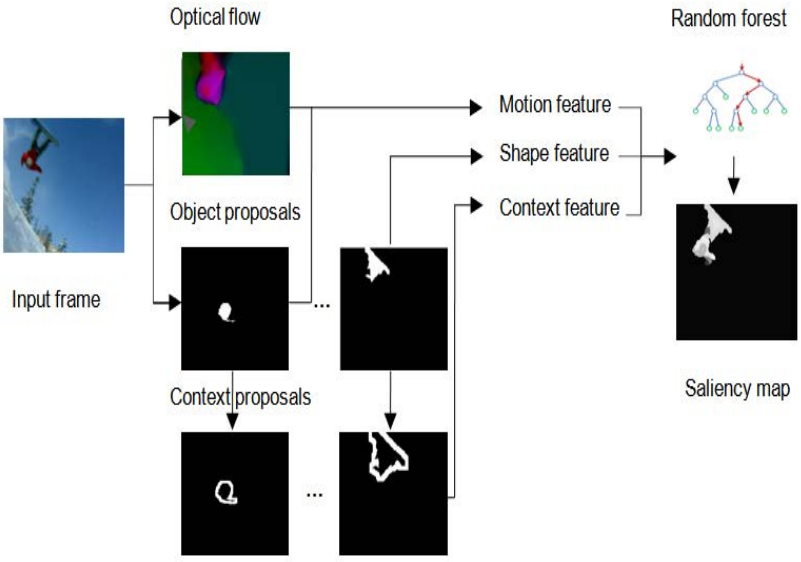


Figure 4.4 – Overview of method. We extend the method shown in Fig. 4.1 with context proposals. Based on the computed proposals we compute the context features from the context proposals. These features are added to the shape and motion features. Again a random forest is applied to assign saliency to each object proposal. These are combined in the final saliency map.

4.4.1 Context Proposal Generation

The context proposal generation is done by finding the immediate surround of each object proposal with a similar surface as the object proposal. The method is similar as the one we proposed for context proposals in still images (see Section2.4.1.)

Consider a mask M of the object proposal. The mask is one for all pixels which belong to the object proposal and zero otherwise. The context is

computed with:

$$\begin{aligned} &C = \left(M \oplus B^{(n)}\right) \setminus M \\ &\text{smallest } n \text{ for which } |C| \geq |M| \end{aligned} \tag{4.7}$$

where B is a structural element, $|C|$ is the number of non-zero values in C, and $\oplus$ is the dilation operator. We used the notation

$$B^{(n)} = \overbrace{B^{(1)} \oplus B^{(1)} \oplus B^{(1)} \oplus B^{(1)}}^{n\ times} \tag{4.8}$$

to indicate multiple dilations. We found that the same settings as used for still images worked fine for video. We used therefore $B = N_8$ which is the eight connected set (a 3x3 structural element with all ones). Examples of context proposals are provided in Fig. 4.3.

4.4.2 Context features

In addition to the shape features and the proposed motion features, we evaluate the usage of the context features in video. For that, we extend our approach for saliency detection in still images developed in Chapter 2 to video. Motivated by the gain obtained using the proposed context features, we will investigate the usage of these features to compute saliency in video data. We combine shape, motion and context features.

As context features we consider the same features as in Chapter 2. We will consider both *Context contrast* and *Context continuity*. Context contrast C_{cc} measures the contrast between the features which make up the salient object and the features which describe its context. For its computation we use Equation Eq. 2.8.

Context continuity measures the continuity of the context around the object proposal. Typically the object is occluding a background region and therefore the pixels on both sides of an object proposal are expected to by similar. Other than the context contrast this feature is purely based on the context and does not use any features from the object. For the of omni-directional context continuity C_{dc} we use the equation Eq. 2.12 and for horizontal context continuity C_h we use equation Eq. 2.13.

We combine the context features together in the feature **f**.

$$\mathbf{f}_{context} = \{C_{cc}, C_{dc}, C_{hc}\} \tag{4.9}$$

These are computed for each context proposal.

4.4.3 Saliency map computation

We learn a regressor from the extracted shape and motion features to the saliency of the object proposal. For the combination of the object proposals we consider the saliency of a pixel to be the average saliency of all the object proposals which include the pixel:

$$S(x,y) = \frac{\sum_i S_i O_i(x,y)}{\sum_i O_i(x,y)} \tag{4.10}$$

here $S(x,y)$ is the saliency at pixel (x,y) and $O_i(x,y)$ is the mask of object proposal i which is one for all pixels which are included in the object proposal and zero everywhere else. Finally, S_i is the saliency estimate for object proposal i which is computed with the random forest based on the shape and motion features. This equation attributes equal saliency to all pixels within an object proposal. We found it to be better to attribute slightly higher saliency to pixels in the center of the proposal, according to:

$$S(x,y) = \frac{\sum_i S_i \left((1-\gamma) O_i(x,y) + \gamma\, dist_{Eu}(O_i(x,y))\right)}{\sum_i O_i(x,y)} \tag{4.11}$$

where we found a low value of $\gamma = 0.1$ to be sufficient. Here $dist_{Eu}(O_i(x,y))$ is the Euclidean distance transform, which is equal to the minimum Euclidean distance for each pixel in the set to a pixel which is not in the set. We normalize the Euclidean distance transform for a single proposal to have a maximum of 1, by dividing by the value of the pixel with the maximum Euclidean distance.

4.5 Experiments

We evaluate the performance of our approach on video saliency detection on two standard benchmark datasets. We first present the experimental setup then we introduce the benchmark datasets and the evaluation metrics that are used. Finally, we compare our results to state of the arts approaches.

4.5.1 Experimental setup

All settings are equal for both datasets. For the estimation of the optical flow we use Lucas-Kanade method [77] which provides magnitude and orientation between each pair of frames. For the object proposals in videos we use the method proposed in [90] which outputs different levels of supervoxel segmentations. For each frame, these supervoxels are grouped together to give spatio-temporal object proposals.

As a classifier we use random forest for regression where we set the number of trees equal to 200. To train the random forest classifier we use a leave-one-out protocol: we train on all videos except one, and test on the video not included in the training set. This is applied separately to the two considered datasets. We report comparison of our methods with six state-of-the-art methods, namely GB [41], ITTI [48], CBS [51], RR [81], RT [98] and GVS [124] .

4.5.2 Datasets

We evaluate our approach on two benchmark datasets that are used by most of the state-of-the-art video saliency methods.
Fukuchi [34], which is a video saliency dataset and contains 10 video sequences with a total of 768 frames with one dynamic object per video. The groundtruth consists of the segmented images. The dynamic objects are from different classes including, horse, flower, skyman, snow cat, snow fox, bird, etc.
SegTrack v2 [63] which is a video segmentation dataset, is used also for video saliency detection methods like [125] and [55]. It contains 14 sequences with a total of 1066 frames. Videos can contain more than one dynamic object. Salient objects include objects with challenging deformable shapes includ-

ing birds, a frog, cars, a soldier, etc.

4.5.3 Evaluation

We evaluate the performance using the precision-recall (PR) curve, ROC curve and F-measure. **Precision** is the proportion of predicted positive pixels that are real positives

$$\text{precision} = \frac{\sum_{x,y} S(x,y)G(x,y)}{\sum S(x,y)} \tag{4.12}$$

where S is the binarized estimated saliency map and G is the binary ground truth. **Recall** is the proportion of real positive pixels that are correctly predicted positives, and is given by

$$\text{recall} = \frac{\sum_{x,y} S(x,y)G(x,y)}{\sum G(x,y)} \tag{4.13}$$

A PR curve can be computed by varying the threshold which is used to binarize $S(x,y)$. The F-score is defined in Eq 2.20.
The ROC curve plots the true positive rate against the false positive rate. A perfect approach has a 0 value for the false positive rate and 100 per cent value for the true positive rate which indicates that predictions are identical to the ground truth.

4.5.4 Baseline Method

In this experiment we evaluate the important choices of our baseline method (the method without context features); we evaluate the motion features, and we evaluate the effectiveness of object proposals for salient object segmentation.

One of the important features of video saliency is the motion feature. We first evaluate the usage of the second eigenvalue λ_2 for saliency assignment and compare it to using λ_1. To do so we run the system described in Fig. 4.1 without including shape features. The results are provided in Table. 4.1 and show that λ_2 obtains significantly better results. This shows that the presence of two local motion directions (as measured by λ_2) is more relevant

Table 4.1 – Comparison in F-score of the first and the second eigenvalues.

Dataset	First eigenvalue λ_1	Second eigenvalue λ_2
Segtrack v2	0.454	**0.684**
Fukuchi	0.604	**0.656**

Table 4.2 – Comparison in F-score of motion features for saliency estimation.

Method	Singh (2015)	Papazoglou (2013)	Ours
Segtrack v2	0.320	0.347	**0.684**
Fukuchi	0.433	0.364	**0.656**

for saliency assignment than just the magnitude of the principle motion direction (as measured by λ_1).

Next, we evaluate the effectiveness of the motion feature by comparing it against two alternatives [92, 108]. The results of this experiment are provided in Table 4.2. The approach based on the optical flow structure tensor obtains significantly better results. This is due to the fact that the second eigenvalue λ_2 is a more stable estimation of the variation of local motion than looking directly at the derivative of the local orientation of the motion vectors, as was done by [92].

As a second experiment we evaluate the gain obtained with the motion feature when also using the shape features. The results are provided in Table. 4.3. The results show that using only shape features leads to 0.73 on Fukuchi dataset. When combining shape and motion we get 0.771 with an absolute gain of **4.1%**. On segtrack v2 dataset the gain is lower at **3.4%**. The gain shows the importance of motion features for saliency detection in

Table 4.3 – Different values of F-score using different combinations of features

Dataset	S	S+M	gain (%)	Proposals
Segtrack v2	0.741	0.775	3.4	0.074
Fukuchi	0.730	0.771	4.1	0.173

Table 4.4 – Evaluation of the impact of adding context features C

Dataset	S	M	C	S+M	S+M+C	gain (%)
Segtrack v2	0.741	0.684	0.779	0.775	0.791	1.6
Fukuchi	0.730	0.656	0.751	0.771	0.776	0.5

Table 4.5 – Comparison with state-of-the-art in F-score

Method	Segtrack v2	Fukuchi
GVS	0.671	0.724
RR	0.570	0.551
RT	0.367	0.663
GB	0.480	0.539
CBS	0.590	0.670
ITTI	0.443	0.567
Ours (C only)	**0.779**	**0.751**
Ours (S+M+C)	**0.791**	**0.776**

videos.

To make sure that the obtained gain is due to our framework and not to the object proposals method we performed an additional test where we evaluated the saliency estimation of only the object proposals without any feature (neither shape nor motion) by directly assigning a saliency of **1** to all object proposals in the image. This was found to obtain a very low F-scores of **0.074** on Segtrack V2 and **0.173** on Fukuchi dataset.

4.5.5 Context Proposals for Video Saliency

Here we evaluate if context features are also efficient for the detection of saliency in video. To the best of our knowledge this has not been evaluated yet.

We add in Table. 4.4 the results when combining with the context features. First, if we consider only the context features, we notice that these features alone obtain considerably better results when we compare it to the shape features and to the motion features alone. The results obtained on

Segtrack improve from Shape features 0.741 to Context features 0.779. Also on Fukuchi a great jump is seen from 0.730 for Shape to 0.751 for Context features. This shows that also in videos context features are very important features for the assigning of saliency. Furthermore, the results show that the combination of shape, motion and context features leads to an absolute gain of **1.6%** on Segtrack v2 dataset and to **0.5%** on Fukuchi.

Finally, we compare our video saliency approach to several state-of-the-art methods in Table. 4.5. On both datasets Segtrack v2 and Fukuchi datasets we clearly outperform the other methods in term of F-score. Interestingly, our method which only uses context features also outperforms state-of-the-art. The precision-recall curves in Fig. 4.6 provide similar conclusions where our method obtains state-of-the-art results for most recall values. Recall values of **RR** [81] and **GVS** [124] are very small when we vary the threshold to 255 and even decrease to 0 in case of **ITTI** [48], **RT** [98] and **GB** [41] since the output saliency maps do not respond to the salient object detection. For the Segtrack v2 and Fukuchi datasets, the minimum value of recall is different from zero, which means that our proposed method is able to highlight the salient object even under challenging situations and complex backgrounds. ROC curves are presented in Fig. 4.5. For low false positive rate our method obtains much higher true positive rates.

We added qualitative comparison on different challenging cases in Figs. 4.7-4.8. When the object touches the image boundaries (first row) a higher saliency probability is assigned to the frog legs. In case of a moving object and a static camera (last two rows), our method detects the dynamic object perfectly. A result of a moving object with higher speed and a static camera is shown in the third row, and produces a good saliency map. In case of an object with high speed and a moving camera, (forth rows) our proposed motion feature highlights only the moving object.

The saliency maps provided by **GB** [41] and **ITTI** [48] do not show the exact location of the salient object because of lack of motion information with complex backgrounds. Saliency maps provided by **RT** [98] detect correctly the salient object but provide blurred saliency maps (first and sixth rows). The performance of the saliency method **RR** [81] which is based on the optical flow, failed to highlight the whole salient object because, optical flow draws the change of position when moving from one frame to another.

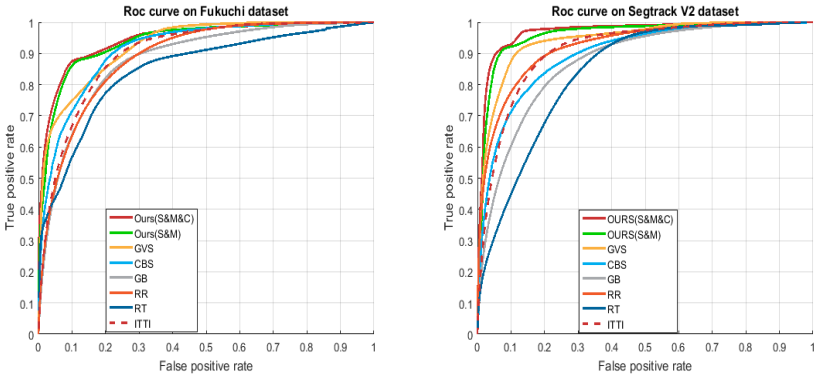


Figure 4.5 – Roc curves on Fukuchi dataset (left) and on Segtrack v2 dataset (right)

In case of slow motion, this method will not be able to produce good saliency maps (see third and sixth rows).

In most cases, **CBS** [51] and **GVS** [124] are able to locate the salient object even in complex situations. For example if the foreground-background colors are similar (see first and last rows) temporal information will serve to highlight only pixels with higher motion information which are considered salient. Based on the aforementioned analysis, two main conclusions can be drawn. First, to detect salient objects in videos, it is essential to examine motion information. Second, developing a method that depends only on motion information is not enough and other features, like shape features, should be included. Combining spatial and temporal information into a video saliency framework produces good results.

4.6 Conclusion

In this chapter, we set out to evaluate the usage of context proposals for video saliency. We start by using object proposal methods which have been previously used successfully in various computer vision applications, such as semantic segmentation and object detection. Recently, they have shown excellent results for saliency estimation in still images. We use the method

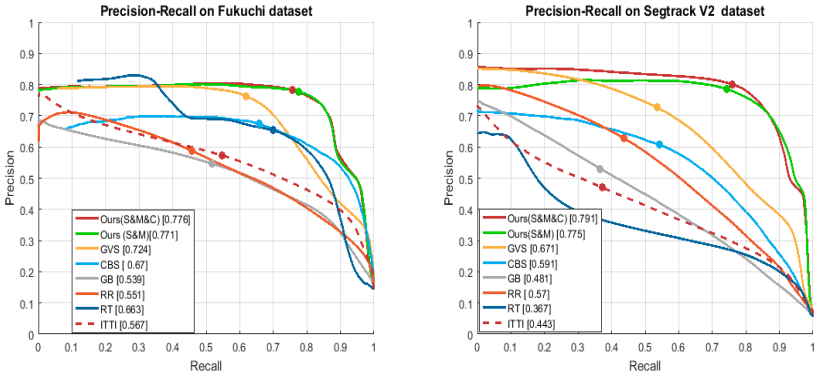


Figure 4.6 – Precision-Recall curves on Fukuchi dataset (left) and on Segtrack v2 dataset (right)

of [90] as an object proposal method in videos. From this method we then compute our context proposals, and their features. In addition, we evaluate several motion features. We prove that the proposed context features lead to better saliency estimation. Actually, as a single feature they obtain better results than motion or shape features alone. Actually, context features alone obtain state-of-the-art results on the challenging Segtrack v2 and Fukuchi datasets. These results are futher improved when we combine shape, motion and context features together.

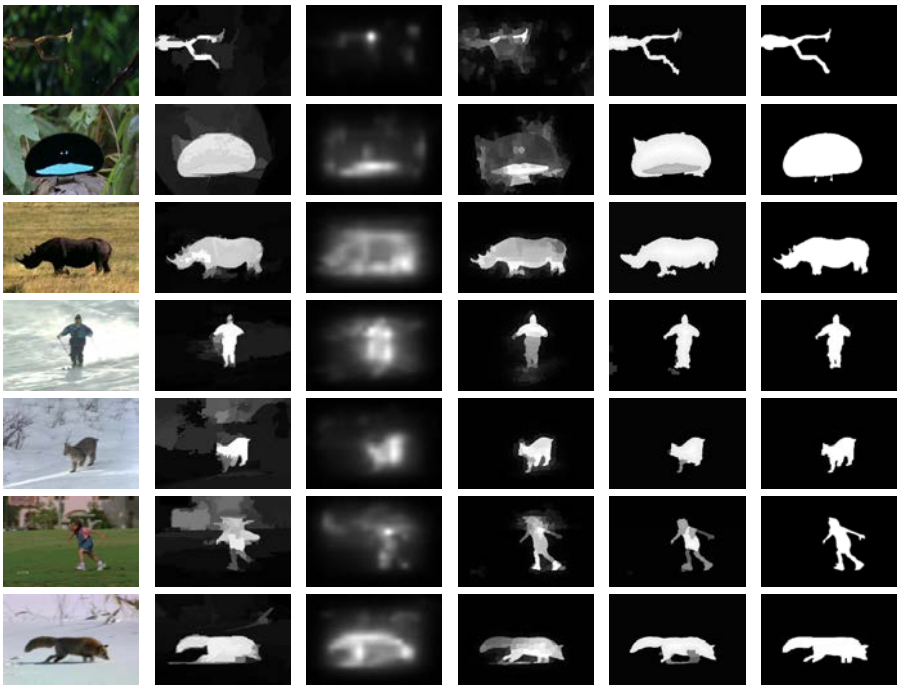


Figure 4.7 – Visual comparison of saliency maps generated from 3 different methods, including our method, CBS [51], GB [41], GVS [124].

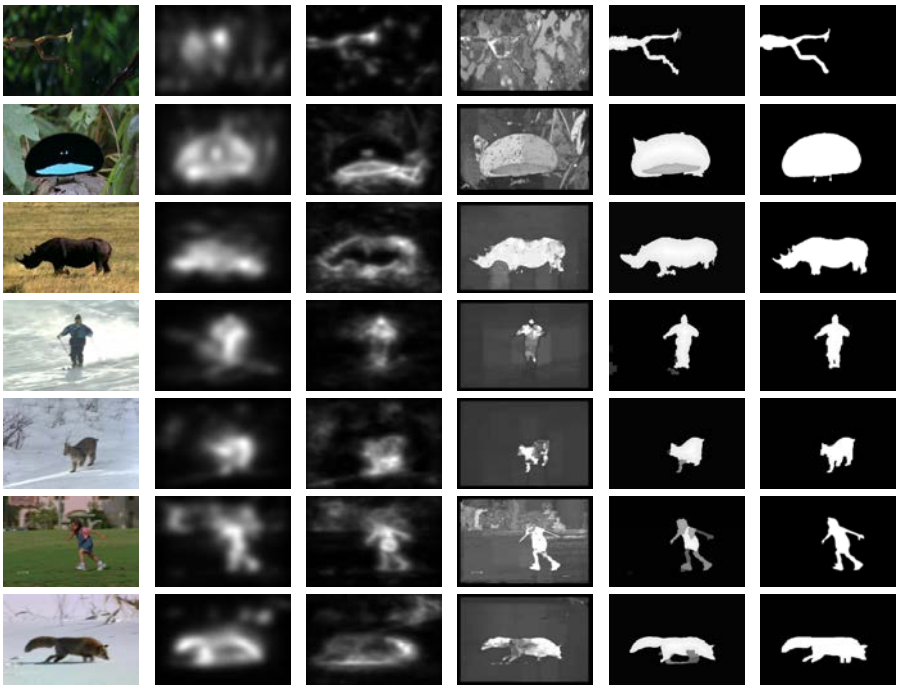


Figure 4.8 – Visual comparison of saliency maps generated from 3 different methods, including our method, ITTI [48], RR [81] and RT [98].

5 Conclusions and Future Directions

In this thesis, we set out to detect the most salient region in images and video. For that purpose we have proposed three different approaches. In this chapter we summarize the approaches proposed to highlight the salient object. We finish this chapter with future work.

5.1 Conclusions

Computational saliency focuses on designing algorithms which, similarly to human vision, predict which regions in a scene are salient. In this chapter, we will discuss our three main contribution of this thesis. We will discuss saliency from context, we will present high-level semantic approach for saliency prediction. Finally, we will provide our approach for saliency in video.

In this thesis, we have evaluated the importance of the direct *context* in the task of saliency prediction. We proposed a saliency approach based on saliency features computed from the direct surround (context) of every object proposal. The direct context provides an important information on the saliency of the object, because the object is typically occluding a background. Here we use deep features from context proposals for center-surround contrast computation to compute saliency. Our method shows how to extend object proposals with context proposals which allow for a precise description of the objects context. It is shown that these can significantly improve the performance of saliency detection.

In our approach, we investigate which network and which layer are optimal for the task of saliency detection. We found that high-level features are more suitable for saliency. Also, we propose to use of the MCG object proposals method [5] for saliency prediction. We prove in chapter 2 that this method is the best object proposal method for the task of saliency prediction among five methods including two recent methods based on deep

learning [46, 95]. These object proposals are designed to directly provide a segmentation. We hypothesize that considering the contrast and the homogeneity of the context can lead to a better saliency assessment.

We demonstrate the effectiveness of the irregular-shaped contextual region *context*. We compare our context proposals, which follow the object proposal boundary, with different context shapes. We consider them to the conventional circular or rectangular neighborhood. For the three different context shapes we extract the same context features. The results show that our approach clearly outperforms the rectangular and circular shaped contexts. Thereby showing that accurate context masks result in more precise saliency estimations.

In the second part of this thesis, we investigated the usage of *high-level semantic information* in the task of saliency prediction. With semantic information we refer to object detection and semantic segmentation results. Due to the success of convolutional neural networks which improved significantly the state-of-the-art of high-level image understanding in recent years, we think it is relevant to re-evaluate the importance of high-level semantic information for saliency detection. Especially, since it is known from human vision research that semantic information plays an important role in the saliency estimation task.

We derived several saliency features which are computed from object detection and semantic segmentation results combined with features based on object proposals [5]. Most computational methods are bottom-up or only include few semantic classes such as faces, and text. Other than these methods we considered a wider group of twenty object classes and evaluated their impact on saliency estimation. We carried on with our study of which semantic classes boost more the task of saliency. We found that the classes of person and bird are among the most important classes.

Finally, in the last part of this thesis, we propose a method to detect saliency in video, by computing saliency from supervoxels [90] and optical flow by combining shape features (used in chapter 3) with motion features. Also, we evaluated the impact of the proposed context proposals proposed in chapter 2 for video saliency using the context features combined with motion features. Our approach is tested on standard object recognition data sets. The results obtained clearly demonstrate the effectiveness of our

approach.

5.2 Future Directions

There are plenty of research opportunities which can be pursued following this report.

In chapter 2 of this thesis, we show the importance of explicit context modelling for saliency estimation (it confirms the results of Mairon and Ben-Shahar [80]). Explicit context estimation is not modelled by current end-to-end networks for saliency. Incorporating context in end-to-end trainable networks for saliency is an interesting research direction for saliency estimation. Also, it would be interesting to apply the proposed context proposals in the field of object detection and semantic segmentation.

The approaches presented in chapter 3 are based on merging shape features and high-level semantic features. However, other saliency features such as Appearances features, deep features etc. can also be combined using the proposed approaches. Moreover, our high-level saliency approach can strongly benefit if we evaluate a wider set of object classes on saliency detection (we currently evaluate the 20 classes from the PASCAL VOC challenge). Evaluating the impact of high-level semantic information on saliency maps derived from eye-tracker devices is also an interesting research direction. It would also be interesting to extend current end-to-end networks for saliency with explicit modules for object detection, and evaluate if such architectures could further improve state-of-the-art.

Furthermore, we acknowledge the necessity of a new dataset which is based on the-odd-one-out to include more variety in images. This dataset should aim to test our saliency approaches when there is multiple objects and only one attracts the attention. We expect such a dataset to be better positioned to evaluate the difference between foreground-background segmentation and salient object detection.

Finally, we propose as our future work for chapter 4. Currently we combine shape features, motion features and context features, but it would be interesting to do wider comparison of saliency features. Video saliency is also clearly in need of larger datasets, and future research in this interesting research direction will be much served by new larger datasets.

Bibliography

[1] Richard A Abrams and Shawn E Christ. Motion onset captures attention. *Psychological Science*, 14(5):427–432, 2003.

[2] Ravi Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604. IEEE, 2009.

[3] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object? In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 73–80. IEEE, 2010.

[4] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2189–2202. IEEE, 2012.

[5] Pablo Arbelaez, Jordi Pont-Tuset, Jonathan Barron, Ferran Marques, and Jagannath Malik. Multiscale combinatorial grouping. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 328–335. IEEE, 2014.

[6] Aymen Azaza, Joost van de Weijer, Ali Douik, and Marc Masana. Context proposals for saliency detection. *Computer Vision and Image understanding*, 2018.

[7] Peng Bian and Liming Zhang. Biological plausibility of spectral domain approach for spatiotemporal visual saliency. In *International conference on neural information processing*, pages 251–258. Springer, 2008.

[8] Josef Bigun, Goesta H. Granlund, and Johan Wiklund. Multidimensional orientation estimation with applications to texture analysis

and optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):775–790, 1991.

[9] Ali Borji. Boosting bottom-up and top-down visual features for saliency estimation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 438–445. IEEE, 2012.

[10] Ali Borji. What is a salient object? a dataset and a baseline model for salient object detection. *IEEE Transactions on Image Processing*, 24(2):742–756, 2015.

[11] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *arXiv preprint arXiv:1411.5878*, 2014.

[12] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013.

[13] Ali Borji, Dicky N Sihite, and Laurent Itti. Salient object detection: A benchmark. In *European Conference on Computer Vision*, pages 414–429. Springer, 2012.

[14] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[15] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *European conference on computer vision*, pages 282–295. Springer, 2010.

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

[17] Zoya Bylinskii, Adrià Recasens, Ali Borji, Aude Oliva, Antonio Torralba, and Frédo Durand. Where should saliency models look next? In *European Conference on Computer Vision*, pages 809–824. Springer, 2016.

[18] Joao Carreira, Rui Caseiro, Jorge Batista, and Cristian Sminchisescu. Semantic segmentation with second-order pooling. *Computer Vision–ECCV 2012*, pages 430–443, 2012.

[19] Joao Carreira and Cristian Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3241–3248. IEEE, 2010.

[20] Moran Cerf, E Paxon Frady, and Christof Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of vision*, 9(12):10–10, 2009.

[21] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3286–3293. IEEE, 2014.

[22] Antoine Coutrot and Nathalie Guyader. Learning a time-dependent master saliency map from eye-tracking data in videos. *arXiv preprint arXiv:1702.00714*, 2017.

[23] Philippe Pérez de San Roman, Jenny Benois-Pineau, Jean-Philippe Domenger, Aymar De Rugy, Florent Paclet, and Daniel Cataert. Saliency driven object recognition in egocentric videos with deep cnn: toward application in assistance to neuroprostheses. *Computer Vision and Image Understanding*, 2017.

[24] John Duncan and Glyn W Humphreys. Visual search and stimulus similarity. *Psychological review*, 96(3):433, 1989.

[25] Krista A Ehinger, Barbara Hidalgo-Sotelo, Antonio Torralba, and Aude Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition*, 17(6-7):945–978, 2009.

[26] Wolfgang Einhäuser, Merrielle Spain, and Pietro Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):18–18, 2008.

[27] Ian Endres and Derek Hoiem. Category-independent object proposals with diverse ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):222–234, 2014.

[28] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[29] Yuming Fang, Zhenzhong Chen, Weisi Lin, and Chia-Wen Lin. Saliency detection in the compressed domain for adaptive image retargeting. *IEEE Transactions on Image Processing*, 21(9):3888–3901, 2012.

[30] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[31] Jie Feng, Yichen Wei, Litian Tao, Chao Zhang, and Jian Sun. Salient object detection by composition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1028–1035. IEEE, 2011.

[32] Katerina Fragkiadaki, Geng Zhang, and Jianbo Shi. Video segmentation by tracing discontinuities in a trajectory embedding. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1846–1853. IEEE, 2012.

[33] Simone Frintrop, Thomas Werner, and Germán Martín García. Traditional saliency reloaded: A good old model in new shape. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 82–90, 2015.

[34] Ken Fukuchi, Kouji Miyazato, Akisato Kimura, Shigeru Takagi, and Junji Yamato. Saliency-based video segmentation with graph cuts and sequentially updated priors. In *2009 IEEE International Conference on Multimedia and Expo*, pages 638–641. IEEE, 2009.

[35] R Gaborski, Vishal S Vaingankar, and RL Canosa. Goal directed visual search based on color cues: Cooperative effects of top-down & bottom-up visual attention. *Proceedings of the Artificial Neural Networks in Engineering, Rolla, Missouri*, 13:613–618, 2003.

[36] Antón Garcia-Diaz, Xosé R Fdez-Vidal, Xosé M Pardo, and Raquel Dosil. Decorrelation and distinctiveness provide with human-like saliency. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 343–354. Springer, 2009.

[37] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[38] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[39] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):1915–1926, 2012.

[40] Bing Han, Xuelong Li, Xinbo Gao, and Dacheng Tao. A biological inspired features based saliency map. In *International Conference on Computing, Networking and Communications*, pages 371–375. IEEE, 2012.

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

[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[43] Jan Hosang, Rodrigo Benenson, Piotr Dollár, and Bernt Schiele. What makes for effective detection proposals? *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2015.

[44] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5300–5309. IEEE, 2017.

[45] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[46] Hexiang Hu, Shiyi Lan, Yuning Jiang, Zhimin Cao, and Fei Sha. Fastmask: Segment multi-scale object candidates in one shot. *arXiv preprint arXiv:1612.08843*, 2016.

[47] Lina Huo, Licheng Jiao, Shuang Wang, and Shuyuan Yang. Object-level saliency detection with color attributes. *Pattern Recognition*, 49:162–173, 2016.

[48] Laurent Itti and Pierre Baldi. A principled approach to detecting surprising events in video. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 631–637. IEEE, 2005.

[49] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[50] Saumya Jetley, Naila Murray, and Eleonora Vig. End-to-end saliency mapping via probability distribution prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5753–5761, 2016.

[51] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Tie Liu, Nanning Zheng, and Shipeng Li. Automatic salient object segmentation based on context and shape prior. In *BMVC*, volume 6, page 9, 2011.

[52] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2083–2090, 2013.

[53] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2106–2113. IEEE, 2009.

[54] Timor Kadir and Michael Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[55] Hansang Kim, Youngbae Kim, Jae-Young Sim, and Chang-Su Kim. Spatiotemporal saliency detection for video sequences based on random walk with restart. *IEEE Transactions on Image Processing*, 24(8):2552–2564, 2015.

[56] Jiwhan Kim, Dongyoon Han, Yu-Wing Tai, and Junmo Kim. Salient region detection via high-dimensional color transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 883–890, 2014.

[57] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987.

[58] Philipp Krähenbühl and Vladlen Koltun. Geodesic object proposals. In *European Conference on Computer Vision*, pages 725–739. Springer, 2014.

[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[60] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[61] Yong Jae Lee, Jaechul Kim, and Kristen Grauman. Key-segments for video object segmentation. In *2011 International Conference on Computer Vision*, pages 1995–2002. IEEE, 2011.

[62] José Lezama, Karteek Alahari, Josef Sivic, and Ivan Laptev. Track to the future: Spatio-temporal video segmentation with long-range motion cues. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011.

[63] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2192–2199, 2013.

[64] G. Li and Y. Yu. Deep contrast learning for salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 478–487, June 2016.

[65] G. Li and Y. Yu. Deep contrast learning for salient object detection. *arXiv preprint arXiv:1603.01976*, 2016.

[66] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5455–5463, 2015.

[67] Jian Li, Martin D Levine, Xiangjing An, Xin Xu, and Hangen He. Visual saliency based on scale-space analysis in the frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):996–1010, 2013.

[68] Shuang Li, Huchuan Lu, Zhe Lin, Xiaohui Shen, and Brian Price. Adaptive metric learning for saliency detection. *IEEE Transactions on Image Processing*, 24(11):3321–3331, 2015.

[69] Xi Li, Liming Zhao, Lina Wei, Ming-Hsuan Yang, Fei Wu, Yueting Zhuang, Haibin Ling, and Jingdong Wang. Deepsaliency: Multi-task deep neural network model for salient object detection. *IEEE Transactions on Image Processing*, 25(8):3919–3930, 2016.

[70] Yin Li, Xiaodi Hou, Christian Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–287. IEEE, 2014.

[71] Xiao Lin, Josep R Casas, and Montse Pardas. 3d point cloud segmentation using a fully connected conditional random field. In *Signal Processing Conference (EUSIPCO), 2017 25th European*, pages 66–70. IEEE, 2017.

[72] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 678–686, 2016.

[73] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Transactions on Pattern analysis and machine intelligence*, 33(2):353–367, 2011.

[74] Zhi Liu, Xiang Zhang, Shuhua Luo, and Olivier Le Meur. Superpixel-based spatiotemporal saliency detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(9):1522–1540, 2014.

[75] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[76] Jing Lou, Mingwu Ren, and Huan Wang. Regional principal color based saliency detection. *PLoS ONE*, 9(11):e112475, 2014.

[77] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. In *IJCAI*, volume 81, pages 674–679, 1981.

[78] Tianyang Ma and Longin Jan Latecki. Maximum weight cliques with mutex constraints for video object segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 670–677. IEEE, 2012.

[79] Norman H Mackworth and Anthony J Morandi. The gaze selects informative details within pictures. *Attention, Perception, & Psychophysics*, 2(11):547–552, 1967.

[80] Rotem Mairon and Ohad Ben-Shahar. A closer look at context: From coxels to the contextual emergence of object saliency. In *The European Conference on Computer Vision*, pages 708–724. Springer, 2014.

[81] Matei Mancas, Nicolas Riche, Julien Leroy, and Bernard Gosselin. Abnormal motion selection in crowds using bottom-up saliency. In *2011 18th IEEE International Conference on Image Processing*, pages 229–232. IEEE, 2011.

[82] Sophie Marat, Tien Ho Phuoc, Lionel Granjon, Nathalie Guyader, Denis Pellerin, and Anne Guérin-Dugué. Modelling spatio-temporal saliency to predict gaze direction for short videos. *International journal of computer vision*, 82(3):231, 2009.

[83] Luca Marchesotti, Claudio Cifarelli, and Gabriela Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *12th International Conference on Computer Vision*, pages 2232–2239. IEEE, 2009.

[84] Ran Margolin, Avishay Tal, and Lihi Zelnik-Manor. What makes a patch distinct? In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1139–1146. IEEE, 2013.

[85] David R Martin, Charless C Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):530–549, 2004.

[86] Thomas Mauthner, Horst Possegger, Georg Waltner, and Horst Bischof. Encoding based saliency detection for videos and images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2494–2502, 2015.

[87] Naila Murray, Maria Vanrell, Xavier Otazu, and C Alejandro Parraga. Low-level spatiochromatic grouping for saliency estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2810–2816, 2013.

[88] Seungjun Nah and Kyoung Mu Lee. Random forest with data ensemble for saliency detection. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 604–607. IEEE, 2015.

[89] Antje Nuthmann and John M Henderson. Object-based attentional selection in scene viewing. *Journal of vision*, 10(8):20–20, 2010.

[90] Dan Oneata, Jerome Revaud, Jakob Verbeek, and Cordelia Schmid. Spatio-temporal object detection proposals. In *European conference on computer vision*, pages 737–752. Springer, 2014.

[91] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Shallow and deep convolutional networks for saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–606, 2016.

[92] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1777–1784, 2013.

[93] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 733–740. IEEE, 2012.

[94] Pedro O Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. In *Advances in Neural Information Processing Systems*, pages 1990–1998, 2015.

[95] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *European Conference on Computer Vision*, pages 75–91. Springer, 2016.

[96] Yao Qin, Huchuan Lu, Yiqun Xu, and He Wang. Saliency detection via cellular automata. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 110–119, 2015.

[97] Anis Rahman, Dominique Houzet, Denis Pellerin, Sophie Marat, and Nathalie Guyader. Parallel implementation of a spatio-temporal visual saliency model. *Journal of Real-Time Image Processing*, 6(1):3–14, 2011.

[98] Esa Rahtu, Juho Kannala, Mikko Salo, and Janne Heikkilä. Segmenting salient objects from images and videos. In *European Conference on Computer Vision*, pages 366–379. Springer, 2010.

[99] S. Razavian, A. Azizpour, J. H. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014.

[100] Dmitry Rudoy, Dan B Goldman, Eli Shechtman, and Lihi Zelnik-Manor. Learning video saliency from human gaze using candidate selection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1147–1154. IEEE, 2013.

[101] Christian Scharfenberger, Alexander Wong, Khalil Fergani, John S Zelek, and David A Clausi. Statistical textural distinctiveness for salient region detection in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 979–986, 2013.

[102] Andrew Senior and Andrew W Senior. *Protecting privacy in video surveillance*, volume 1. Springer, 2009.

[103] Hae Jong Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 9(12):15–15, 2009.

[104] Puneet Sharma. Perceptual image difference metrics. saliency maps & eye tracking. Master's thesis, 2008.

[105] Christian Siagian and Laurent Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):300–312, 2007.

[106] Eero P Simoncelli, Edward H Adelson, and David J Heeger. Probability distributions of optical flow. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 310–315. IEEE, 1991.

[107] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[108] Anurag Singh, Chee-Hung Henry Chu, and M Pratt. Learning to predict video saliency using temporal superpixels. In *Pattern Recognition Applications and Methods, 4th International Conference on*, pages 201–209, 2015.

[109] Parthipan Siva, Chris Russell, Tao Xiang, and Lourdes Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3238–3245, 2013.

[110] X Yu Stella and Dimitri A Lisin. Image compression based on visual saliency at individual scales. In *Advances in Visual Computing*, pages 157–166. Springer, 2009.

[111] Jingang Sun, Huchuan Lu, and Xiuping Liu. Saliency region detection based on markov absorption probabilities. *IEEE Transactions on Image Processing*, 24(5):1639–1649, 2015.

[112] Na Tong, Huchuan Lu, Ying Zhang, and Xiang Ruan. Salient object detection via global and local cues. *Pattern Recognition*, 48(10):3258–3267, 2015.

[113] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[114] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[115] Joost Van De Weijer, Theo Gevers, and Arnold WM Smeulders. Robust photometric invariant features from the color tensor. *IEEE Transactions on Image Processing*, 15(1):118–127, 2006.

[116] Gevers Th. Van De Weijer J. and Bagdanov A.D. Boosting color saliency in image feature detection. *IEEE TRANS. PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 28:150–156, 2005.

[117] Shouhong Wan, Peiquan Jin, and Lihua Yue. An approach for image retrieval based on visual saliency. In *International Conference on Image Analysis and Signal Processing*, pages 172–175. IEEE, 2009.

[118] Julius Wang, Hamed R Tavakoli, and Jorma Laaksonen. Fixation prediction in videos using unsupervised hierarchical features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 50–57, 2017.

[119] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Deep networks for saliency detection via local estimation and global search. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3183–3192, 2015.

[120] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Saliency detection with recurrent fully convolutional networks. In *European Conference on Computer Vision*, pages 825–841. Springer, 2016.

[121] Meng Wang, Janusz Konrad, Prakash Ishwar, Kevin Jing, and Henry Rowley. Image saliency: From intrinsic to extrinsic context. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 417–424. IEEE, 2011.

[122] Peng Wang, Jingdong Wang, Gang Zeng, Jie Feng, Hongbin Zha, and Shipeng Li. Salient object detection for searched web images via global saliency. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3194–3201. IEEE, 2012.

[123] Shuo Wang, Ming Jiang, Xavier Morin Duchesne, Elizabeth A Laugeson, Daniel P Kennedy, Ralph Adolphs, and Qi Zhao. Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron*, 88(3):604–616, 2015.

[124] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3395–3402, 2015.

[125] Xiang Wang, Huimin Ma, and Xiaozhi Chen. Geodesic weighted bayesian model for saliency optimization. *Pattern Recognition Letters*, 75:1–8, 2016.

[126] Zheshen Wang and Baoxin Li. A two-stage approach to saliency detection in images. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 965–968. IEEE, 2008.

[127] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of vision*, 14(1):28–28, 2014.

[128] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

[129] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1155–1162. IEEE, 2013.

[130] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[131] Yang M. H. Yang, J. Top-down visual saliency via joint crf and dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(3):576–588, 2016.

[132] Steven Yantis and Howard E Egeth. On the distinction between visual salience and stimulus-driven attentional capture. *Journal of Experimental Psychology: Human Perception and Performance*, 25(3):661, 1999.

[133] Dingwen Zhang, Huazhu Fu, Junwei Han, and Feng Wu. A review of co-saliency detection technique: Fundamentals, applications, and challenges. *CoRR*, abs/1604.07090, 2016.

[134] Dong Zhang, Omar Javed, and Mubarak Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 628–635, 2013.

[135] Jianming Zhang and Stan Sclaroff. Saliency detection: A boolean map approach. In *Proceedings of the IEEE international conference on computer vision*, pages 153–160, 2013.

[136] Lin Zhang, Zhongyi Gu, and Hongyu Li. Sdsp: A novel saliency detection method by combining simple priors. In *20th International Conference on Image Processing*, pages 171–175. IEEE, 2013.

[137] Qi Zhao and Christof Koch. Learning saliency-based visual attention: A review. *Signal Processing*, 93(6):1401–1407, 2013.

[138] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1265–1274, 2015.

[139] Sheng-hua Zhong, Yan Liu, Feifei Ren, Jinghuan Zhang, and Tongwei Ren. Video saliency detection via dynamic consistent spatio-temporal attention modelling. In *AAAI*, pages 201–209, 2013.

[140] Chunbiao Zhu, Ge Li, Wenmin Wang, and Ronggang Wang. Salient object detection with complex scene based on cognitive neuroscience. In *Multimedia Big Data (BigMM), 2017 IEEE Third International Conference on*, pages 33–37. IEEE, 2017.

[141] Dollar P. Zitnick, C. L. Edge boxes: Locating object proposals from edges. In *In European Conference on Computer Vision*, pages 391–405. IEEE, 2014.