# Attentive Systems: A Survey

**Tam V. Nguyen**[1] · **Qi Zhao**[2] · **Shuicheng Yan**[3]

Received: 21 July 2015 / Accepted: 1 September 2017 / Published online: 15 September 2017

**Abstract** Visual saliency analysis detects salient regions/objects that attract human attention in natural scenes. It has attracted intensive research in different fields such as computer vision, computer graphics, and multimedia. While many such computational models exist, the focused study of what and how applications can be beneficial is still lacking. In this article, our ultimate goal is thus to provide a comprehensive review of the applications using saliency cues, the so-called attentive systems. We would like to provide a broad vision about saliency applications and what visual saliency can do. We categorize the vast amount of applications into different areas such as computer vision, computer graphics, and multimedia. Intensively covering 200+ publications we survey (1) key application trends, (2) the role of visual saliency, and (3) the usability of saliency into different tasks.

**Keywords** Attentive systems · Visual attention · Salient region/object detection · Interestingness · Scene understanding · Computer graphics · Image retargeting · Feature pooling · Multimedia · Compression

Communicated by Yoichi Sato.

✉ Tam V. Nguyen
tamnguyen@udayton.edu

Qi Zhao
qzhao@umn.edu

Shuicheng Yan
eleyans@nus.edu.sg

[1] University of Dayton, Dayton, OH, USA

[2] University of Minnesota, Minneapolis, MN, USA

[3] National University of Singapore, Singapore, Singapore

## 1 Introduction

Human vision system is able to rapidly detect salient regions in the scene (Schneider and Shiffrin 1977; Shiffrin and Schneider 1977). These salient regions are later processed to extract the high-level information. This complex biological system is naturally built to effortlessly detect potential prey, predators, or mates in the real world. Visual saliency—particularly stimulus-driven, saliency-based attention—has been an active research field over the past decades. This research has been first studied by neuroscientists and cognitive scientists, and has recently attracted a lot of interest in other research communities such as computer vision, computer graphics, and multimedia applications. Even though visual saliency has been applied for different research and practical problems, there is not yet an extensive survey on its applications. Thus, in this paper, we aim to thoroughly review attentive systems that are built on top of visual saliency outputs, clarify less understood challenges, and offer learned lessons from existing works.

The remainder of this article is organized as follows. In Sect. 2, we provide an overview, including the application taxonomy and short survey of saliency models. Next, we introduce and discuss the applications in different domains, i.e., computer vision, computer graphics, multimedia, and miscellaneous applications in Sects. 3, 4, 5, and 6, respectively. The best practices of the applications are discussed and followed by conclusions in Sect. 7.

## 2 Overview

There exist hundreds of applications based on visual attention from three different sources, namely, generic saliency models, task-driven attention, and human gaze.

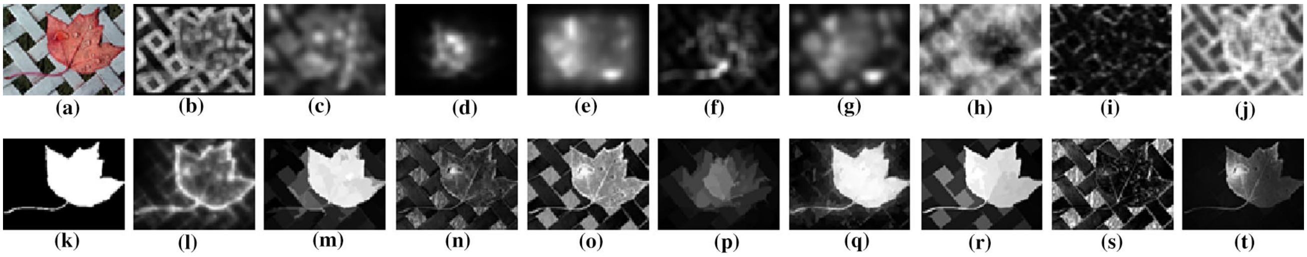

**Fig. 1** Original image (**a**), human binary map (**k**), and maps from 18 state-of-the-art saliency models (**b**–**j**, fixation prediction methods; **l**–**t**, salient object detection methods). **b** Attention based on information maximization (AIM, Bruce et al. 2005), **c** boolean map based saliency (BMS, Zhang and Sclaroff 2013), **d** saliency based on region covariance (COV, Erdem and Erdem 2013), **e** graph based saliency (GB, Harel et al. 2006), **f** incremental coding length (ICL, Hou and Zhang 2008), **g** visual attention measurement (IT, Itti et al. 1998), **h** induction model (SIM, Murray et al. 2011), **i** spectral residual (SR, Hou and Zhang 2007), **j** saliency using natural statistics (SUN, Zhang et al. 2008), **l** context-aware (CA, Goferman et al. 2010), **m** discriminative regional feature integration (DRFI, Jiang et al. 2013), **n** frequency tuned saliency (FT, Achanta et al. 2009), **o**, **p** global contrast saliency (HC and RC, Cheng et al. 2015), **q** high-dimensional color transform (HDCT, Kim et al. 2014), **r** hierarchical saliency (HS, Yan et al. 2013), **s** spatial temporal cues (LC, Zhai and Shah 2006), and **t** saliency filters (SF, Perazzi et al. 2012)

*Generic saliency models* yield a saliency map that later is utilized by some attentive systems. We need to emphasize that this survey sorely focuses on applications of generic saliency models. In literature, hundreds of computational saliency models (Itti et al. 1998; Bruce et al. 2005; Harel et al. 2006; Goferman et al. 2010; Perazzi et al. 2012; Achanta et al. 2009; Koch and Ullman 1985; Hou and Zhang 2007, 2008; Zhang and Sclaroff 2013; Erdem and Erdem 2013; Nguyen and Liu 2017) are available that predict important regions or objects in a scene. They are useful in a variety of tasks for the lofty goal of scene understanding. There exist two types of groundtruth data for visual saliency prediction, namely, the human fixation map (i.e., fixation points smoothened by a Gaussian kernel) for fixation prediction, and the binary object mask for salient object/region detection.

Figure 1 shows saliency maps of different 18 computational models recommended in Perazzi et al. (2012), Jiang et al. (2013), Cheng et al. (2015), and Borji et al. (2012). Most fixation prediction maps are of low resolution and highlight edges. Meanwhile, the salient object maps focus on the entire objects. Note that reviewing all computational models is beyond the scope of this paper. Besides predicting important regions on images, there are also approaches in the video domain, namely, video or dynamic saliency (Zhai and Shah 2006; Nguyen et al. 2013). In addition, other modalities that impact visual saliency are explored, i.e., depth (Lang et al. 2012; Desingh et al. 2013), touch (Ni et al. 2014), computer mouse movement (Jiang et al. 2015a), and audio factors (Chen et al. 2014).

Meanwhile, *task-driven attention* requires more than a generic saliency model. There are many works based on the task-driven saliency to facilitate some tasks. For example, attention-based recurrent networks have been successfully applied to a wide variety of tasks including handwriting synthesis (Graves 2013), machine translation (Bahdanau et al. 2014), image caption generation (Xu et al. 2015) and visual object classification (Mnih et al. 2014). *Human gaze* records eye fixations of a user as in two ways. First, the commercial eye tracking devices provide accurate fixation points. However, commercial eye trackers are usually expensive and requires specialized installation. Therefore, there are a few approaches (Zhang et al. 2015; Sugano et al. 2010; Choi etal. 2016) that utilize an off-the-shelf webcam for gaze estimation (i.e., appearance-based gaze estimation) which can be placed in front of a monitor. There are many works that leverage human gaze to facilitate tasks of interest (Lee et al. 2012; Mishra et al. 2012; Xu et al. 2015; Yun et al. 2013). In this paper, attentive systems using human gaze and task-driven saliency appear only as supplemental information.

In this paper, we consider the *task-driven saliency* and the *human gaze* as the *guided attention*. Other than human gaze, task-driven and generic saliency estimation, there also exist some dedicated computer vision methods, i.e., object hypotheses methods for object detection/recognition. As discussed in Elazary and Itti (2008), interesting objects are visually salient. In fact, object hypotheses generation and salient object detection approaches are closely related. On the one hand, object hypotheses generation approaches consider saliency as a useful cue for measuring objectness of a region (Alexe et al. 2012; Cheng et al. 2014; Krähenbühl and Koltun 2014; Zitnick and Dollár 2014). On the other hand, object hypothesis methods can be considered as support methods to locate salient objects. For example, object hypotheses generation models or objectness measures attempt to generate a small set (e.g., a few hundreds or thousands) of object regions, so that these regions cover every object in the input image, regardless of the specific categories of those objects. Estimating object hypotheses in a pre-processing stage greatly speeds up the computation by reducing the search locations, and also improves the detection accuracy. Nguyen (2015), Nguyen and Sepulveda (2015), and Srivatsa and Babu (2015) show that objectness hypothe-

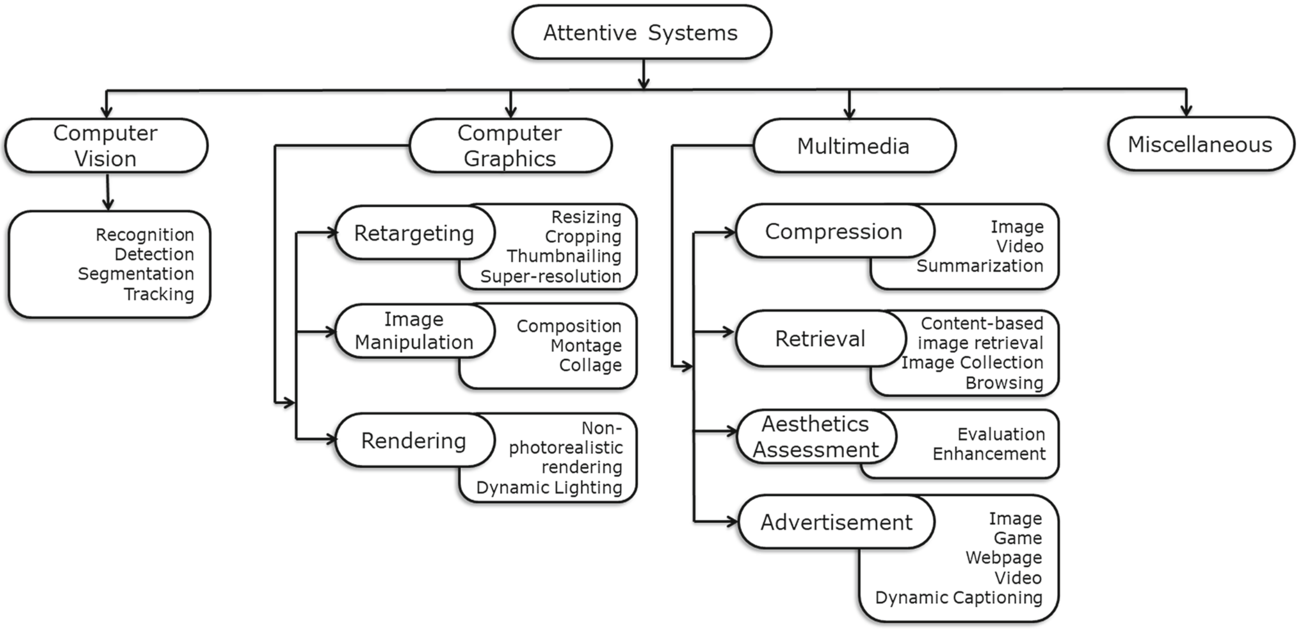

**Fig. 2** Taxonomy of popular attentive systems based on visual saliency research

ses can provide some important cues to locate salient objects. To do so, they incorporate other constraints, namely, distinctiveness and compactness.

Some of saliency prediction models and objectness hypothesis generators have been used in both academic and commercial products (Cheng et al. 2014; Harel et al. 2006; iLab 2010). Again, reviewing all these aforementioned models is out of the scope of this paper. The recent progresses of state-of-the-art works are extensively reviewed in Borji and Itti (2013), Borji et al. (2012, 2015), and Li et al. (2014). Visual attention plays a significant role in the human visual system to focus limited perceptual and cognitive resources to the most important regions in the scene. The question of how to apply this mechanism into the real artificial systems is very interesting. Even though saliency maps that mimic the attentional mechanism of the biological systems have been used for different research and practical problems, there is not yet an extensive survey on its applications. Thus, in this review paper, our main goal is to thoroughly review attentive systems that are built on top of visual saliency outputs, clarify less understood challenges, and offer learned lessons from existing works. Since different applications have different viewpoints about visual saliency, we categorize the applications of visual saliency into four categories, namely, computer vision, computer graphics, multimedia and miscellaneous applications. It is worth noting that we follow the sub-categories of Google Scholar in Engineering and Computer Science area (Martín-Martín et al. 2014).[1] It also notes that the categories in this paper are also recommended by the recent review paper in visual saliency computational models (Borji et al. 2015). Figure 2 summarizes the taxonomy of popular attentive systems reviewed in this paper.

## 3 Computer Vision Applications

In this section, we review computer vision applications that allow computers to perceive the similar way as humans do. We group the applications to different subcategories, namely, recognition, detection, segmentation and tracking.

### 3.1 Recognition

*Scene classification* is one of the most fundamental problems in computer vision. Visual saliency is used as a *criterion* for selecting local regions from where local features, e.g., HOG (Dalal and Triggs 2005), SIFT (Lowe 1999), are extracted. Kadir and Brady (2001) show that saliency, scale selection and content description are intrinsically related. In addtion to the local scale selection, the method considers saliency across scale as well as spatial dimensions. Siagian and Itti (2007) state that the gist feature (Oliva and Torralba 2001) can be useful in outdoor localization for a walking human, with straightforward application to autonomous mobile robotics. This capability reduces the need for detailed calibration in which a robot has to rely on the ad-hoc knowledge of designers for reliable landmarks. Frintrop and Jensfelt (2008) present a complete visual SLAM system, which includes feature detection, tracking, loop closing and

[1] https://scholar.google.com.sg/citations?view_op=top_venues&hl=en&vq=eng.

active camera control. Landmarks are selected based on biological mechanisms that favor salient regions. They discover that the repeatability of salient regions is considerably higher than the regions from standard detectors. Borji and Itti (2011) propose an approach for scene classification by extracting and matching visual features only at the focuses of visual attention instead of the entire scene. They calculate the overall similarity between two images by matching the salient regions. The $k$ nearest neighbors to the test image are retrieved and the class label of this image is assigned as the label of the most frequent class.

*Object recognition* aims to find the existence of a certain object in an image. The idea of using saliency is that not all parts of an image provide useful information. If we attend only to the relevant parts, we can recognize the image more quickly with less resources. Salah et al. (2002) develop a serial model for visual pattern recognition based on the primate selective attention mechanism. It simulates the primitive, bottom-up attentive level of the human visual system with a saliency scheme and the more complex, top-down, temporally sequential associative level with observable Markov models. Gao and Vasconcelos (2004), Gao et al. (2009) propose an alternative definition of saliency, denoted by discriminant saliency that is intrinsically grounded on the recognition problem. This work is based on the intuition that, for recognition, the salient features of a visual class are those that best distinguish it from all other visual classes of recognition interest. Rutishauser et al. (2004) use the object recognition algorithm by SIFT matching. Recognition is performed by matching keypoints found in the test image with stored object models. In their model, finding salient patches is done for learning and recognition before keypoints are extracted. The use of contrast modulation as a means of deploying object-based attention is motivated by neurophysiological experiments that show a tight link between luminance contrast and bottom-up attention as well as by its usefulness with respect to SIFT matching process.

Meanwhile, Walther and Koch (2006) model the object recognition process as the networks of linear threshold units. Once a proto-object region is selected, the object recognition system will be able to form hypotheses about the identity of the attended objects. This will then in turn instruct the attentional system to focus on features or regions that would provide information for the verification or falsification of those hypotheses. Moosmann et al. (2006) combine bottom-up and top-down processes in a way that classification errors are much lower than using the bottom-up process alone. They propose a novel classifier that combines saliency maps with an object part classifier: prior knowledge stored in the classifier is used to simultaneously build the saliency map online as well as to provide information about the object class. Kanan and Cottrell (2010) propose an approach based upon two facets of the visual system: sparse visual features that capture the statistical regularities in natural scenes and sequential fixation-based visual attention. In particular, saliency maps are used as interest point operators. Their approach works well since it employs a non-parametric exemplar-based classifier. This yields several immediate benefits: it does not degrade the discriminability of the features and it employs a simple representation of spatial relationships. By replacing the first layer of the hierarchical architecture in Riesenhuber and Poggio (1999) with saliency networks, Han and Vasconcelos (2010) report that saliency has a significant positive impact on recognition. Additionally, max-based pooling does not appear to have an advantage over averaging, indicating that selecting discriminant features is more important than locating them exactly.

Saliency is sometimes referred to as a criterion for *feature pooling*. Chen et al. (2012) introduce a hierarchical matching framework for image classification based on bag-of-words representation. Each image is expressed as a bag of orderless pairs, each of which includes a local feature vector encoded over a visual dictionary and its corresponding side information from priors or contexts. They use two types of side information: object confidence map and visual saliency map, from object detection priors and within-image contexts respectively. The side information is used for hierarchical clustering of the encoded local features. In particular, the saliency-guided pooling is described as followings. Denote $A$ as the number of saliency-guided spatial layers, the total number of attention-aware spatial channels is $2^A - 1$. For the $a$-th layer, image descriptors are grouped to $2^{a-1}$ channels according to threshold values $\theta_a = \{\frac{1}{2^{l-1}}, \frac{2}{2^{a-1}}, \ldots, \frac{2^{a-1}}{2^{a-1}}\}$. Based on their saliency values in the saliency map $\boldsymbol{S}$, the local descriptors are assigned to the corresponding channel. The saliency-guided channels are demonstrated in Fig. 3a–c.

In another work, Ren et al. (2014) apply saliency maps to better encode image features for object recognition. Since the objects usually correspond to salient regions, and these regions usually play more important roles for object recognition than the background, they incorporate a saliency map into sparse coding-based image representation.

Algorithms using “bag of features” for video representations achieve state-of-the-art performance (Laptev et al. 2008; Kläser et al. 2008; Wang et al. 2011, 2009; Wang and Schmid 2013) on action recognition tasks, such as on the challenging Hollywood2 benchmark. Many works (Mathe and Sminchisescu 2012, 2013; Nguyen et al. 2015; Vig et al. 2012) investigate the benefit of space-variant processing of inputs, inspired by attentional mechanisms in the human visual system. Saliency is considered as a cue to separate foreground actors and background environment. The visual content in the foreground relates to the actors performing the action whereas the visual content in the background provides the context information. Recently, Nguyen et al. (2015) pro-

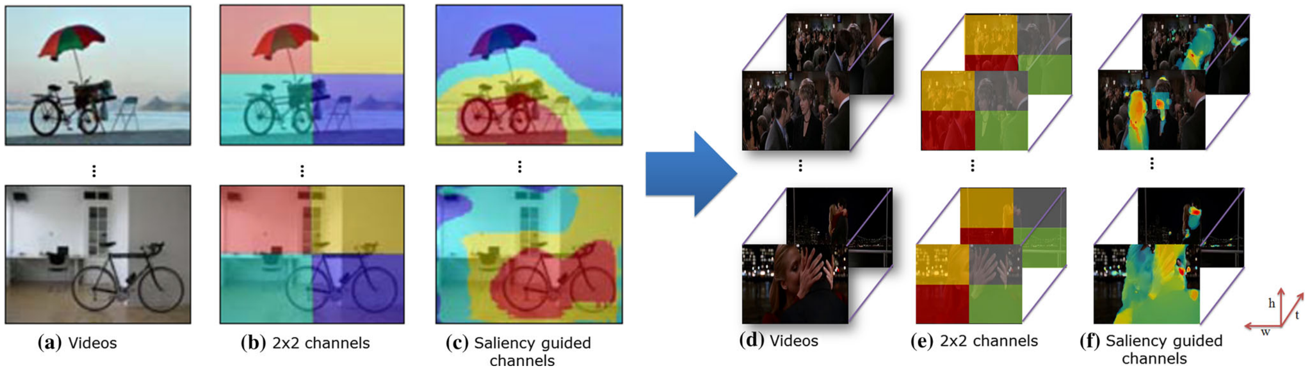

**Fig. 3** Illustration of the saliency guided matching for images (**a**–**c**) (Chen et al. 2012) and video (**d**–**f**) (Nguyen et al. 2015). The local features are pooled according to partition of **b**, **e** traditional SPM and **c**, **f** the saliency guided pooling illustration in the form of the heatmaps which superimpose the saliency maps onto the original color images/video frames. The figure shows the saliency-based framework is superior than SPM in object matching across different images. Images courtesy of Chen et al. (2012), Nguyen et al. (2015). For better viewing of all figures in this paper, please see original color pdf file

pose Spatial-Temporal Attention-aware Pooling procedure aims to pool video local descriptors into channels guided by the predicted video saliency maps. In addition to spatial pooling mentioned in Chen et al. (2012), the video frames are divided into $T$ temporal layers and the temporal channel of each descriptor $x$ is denoted as:

$$G_t(x) \subset \{1, 2, \ldots, 2^T - 1\}.$$

Then the visual descriptors belonging to the $a$-th attention-aware channel and the $t$-th temporal channel are pooled to produce the descriptor as illustrated in Fig. 3d–f. Similarly, Mathe and Sminchisescu (2012), Mathe and Sminchisescu (2013) explore the relationship between human visual attention and computer vision, with emphasis on action recognition in videos. They introduce saliency as a criterion to select features for action recognition. Likewise, Vig et al. (2012) employ saliency-mapping algorithms to find informative regions and descriptors corresponding to these regions are either used exclusively, or are given greater representational weights with additional codebook vectors.

In *robotics*, the problem of localization is central to endowing mobile machines with object recognition algorithms. As studied in Tatler et al. (2011), there is a consistent set of principles underlying search guidance involving behavioral relevance, reward, or uncertainty about the state of the environment, as well as the learned models of the environment, or priors. Range sensors such as sonar and ladar are particularly effective in indoor environments due to many structural regularities such as flat walls and narrow corridors. In outdoor environments, however, these sensors become less robust given all the protrusions and surface irregularities. Therefore, Ouerhani et al. (2005) propose a landmark-based localization method based on visual attention. In the learning phase, a multicue, multi-scale saliency-based model of visual attention is computed and used to automatically acquire robust visual landmarks that are integrated into a topological map of the navigation environment. During navigation, the same visual attention model detects the most salient visual features that are then matched to the learned landmarks. The matching result yields a probabilistic measure of the current location of the robot. Siagian and Itti (2009) use complementary (Oliva and Torralba 2001) and saliency features, implemented in parallel using shared raw feature channels (color, intensity, orientation), as study of human visual cortex suggests. With the saliency model, the system automatically selects consistently salient regions as localization cues. Since the system performs matching within a much smaller region rather than the entire scene, the process is more efficient in the number of SIFT keypoints compared. Further, the gist features along with saliency at almost no computational cost, approximate the image layout and provide segment estimation. Mertsching et al. (1998) introduce a system to recognize complex objects. The system is coupled with two different experimental platforms: a stereo camera head and a mobile robot with a smaller monocular camera head. The stereo camera head measures depth information from the scene while the mobile robot is able to navigate through the scene. The saliency map is computed based on the depth map and several static/dynamic features. The scene segments are further extracted from the saliency map for object recognition.

### 3.2 Detection

*Object detectors* conventionally use a sliding window across the image and apply a binary classifier at each window to detect the presence or absence of the target object. While this approach is successfully applied to detecting rigid and non-rigid objects such as faces, cars and pedestrians, it is slow and computationally expensive as each classifier (corresponding to every object category) is run independently at every window within the image. The speed bottleneck of the sliding window approach can be overcome by using saliency to quickly select a few interest regions in the image. This area receives much interest recently, with several systems using attention as a front-end to accelerate detection speed and reduce complexity of automated multi-target detection.

One of the most well-known works in object detection is the face detector proposed by Viola and Jones (2004). They combine successively more complex classifiers in a cascade structure that dramatically increases the speed of the detector by focusing attention on salient regions of the image. Later, Mitri et al. (2005) introduce VOCUS: Visual Object detection with a CompUtational attention System with a robust object detection method with an application to ball recognition. VOCUS finds regions of interest generating a hypothesis for possible locations of the ball. The classifier verifies the hypothesis by detecting balls at regions of interest. Fritz et al. (2004) introduce a saliency-based approach for object detection. Its key contribution to visual attention is to investigate information theoretic saliency measures with respect to object search and recognition. Early features are tuned to selectively respond to task related visual features, i.e., locally discriminative information that is useful in object recognition. The discriminative regions are determined from the information content in the local appearance patterns. A rapid mapping from appearances to discriminative regions is estimated using decision trees. The focus of attention on discriminative patterns enables the efficient detection of a searched object, but also the definition of sparse object representations to respond only to task relevant information. The performance in object recognition from single images dramatically increased considering only discriminative patterns.

Navalpakkam and Itti (2006) propose a model that combines both bottom-up as well as top-down attentional influences. Their proposed model first computes the naive, bottom-up saliency of every scene location for different local visual features (i.e., different colors, orientations and intensities) at multiple spatial scales. Next, the top-down component uses learnt statistical knowledge of local features of the target and distracting clutter, to optimize the relative weights of the bottom-up maps such that the overall saliency of the target is maximized relative to the surrounding clutter. Such optimization renders the targets more salient than the distractors, thereby maximizing target detection speed. Frintrop (2006) introduces a weighting function based on a measure of object uniqueness is applied to each map before summing up the maps for locating an object. Ehinger et al. (2009) present a model of search guidance that combines saliency, target features, and scene context, and accounts for 94% of the agreement between human observers searching for targets in over 900 scenes. In the people search task, the scene context model proves to be the single most important component driving the high performance of the combined source model. Butko et al. (2009) consider a method for improving the run-time of general-purpose object-detection algorithms. Their method is based on a model of visual search in humans, which predicts scanpaths to maximize the long-term information about the location of the target of interest. The approach is used to drive robot cameras that physically scan scenes and to improve the scanning speed for very large high resolution images.

Saliency-based object detection is also used in *medical applications*. Hong and Brady (2003) develop a segmentation method to detect salient regions in mammograms. Salient regions correspond to distinctive areas that may include the breast boundary, the pectoral muscle, candidate masses and some other dense tissue regions. The breast boundary and the pectoral muscle can be easily identified from the extracted salient regions using anatomical information. Parikh et al. (2010) present a portable wearable system that can be used in conjunction with a retinal prosthesis, to identify important objects that a retinal prosthesis patient may not be able to see due to implant limitations. Shen et al. (2013) propose a novel hierarchical moving target detection method based on spatiotemporal saliency. Temporal saliency is used to get a coarse segmentation, and spatial saliency is extracted to obtain the objects appearance details in candidate motion regions. Finally, by combining temporal and spatial saliency information, the method refines detection results.

*Object discovery* is the task of detecting unknown objects in images. Object discovery is a challenging task for machine. The reason behind is its ‘chicken-and-egg property’ of the problem: how to search for an object before knowing what it looks like? The task is of large interest in many fields of computer vision, ranging from the automatic analysis of web images to interpreting data of a mobile robot or a driver assistant system. Karpathy et al. (2013) present a method for discovering object models from 3D meshes of indoor environments. Their algorithm first decomposes the scene into a set of candidate mesh segments and then ranks each segment according to its “objectness” a quality that distinguishes objects from clutter. They use five intrinsic shape measures: compactness, symmetry, smoothness, and local and global convexity. The frequently occurring geometries are more likely to correspond to complete objects. Frintrop et al. (2014) present a new approach for object discovery, based on findings of the human visual system. Proto-objects are detected with a segmentation module, generating perceptually coherent image regions. In parallel, a saliency system detects regions of interest in images and serves to select segments, depending on their saliency. Roberts et al. (2012) use motion saliency and develop nonlinear image summary factors to keep computational complexity low while mapping relevant objects and maintaining accuracy.

### 3.3 Segmentation

Scene segmentation is an important step towards full scene understanding. Saliency is considered as a good cue for *figure/ground* segmentation. Maki et al. (2000) incorporate depth information obtained from stereopsis, the disparity and flow by local phase from the video for attention prediction. Donoser et al. (2009) introduce a fully unsupervised segmentation method, which is based on the idea of combining several figure/ground segmentations (each focussing on a different salient part of the image) into one composite segmentation result. Johnson-Roberson et al. (2010) extend traditional image segmentation techniques into a full 3D representation from a 3D point cloud. The image saliency techniques are applied to generate seed point for the proposed segmentation technique. The salient points provide a set of hypotheses that they project into the point cloud to begin the segmentation process. Li et al. (2011) use graph cuts (Kolmogorov and Zabih 2004; Boykov and Kolmogorov 2004) to find global optimal segmentation of an n-dimensional image. With the guidance of saliency, users do not have to select foreground object and background seeds.

As aforementioned, saliency map provides some hints about where salient objects locate in the input image. However, it cannot count how many salient objects or segment the salient object out. Therefore, there are some works on the figure/ground segmentation task which actually use saliency map as a cue to perform salient object segmentation. For example, Cheng et al. (2015) use the computed saliency map to assist in automatic salient object segmentation. This immediately enables automatic analysis of large internet image repositories. In particular, they make two enhancements to Rother et al. (2004): "iterative refining" and "adaptive fitting", which together handle considerably more noisy initializations. In another work, to extract the foreground of the image automatically, Qin et al. (2014) combine the region saliency based on entropy rate superpixel with the affinity propagation clustering algorithm to get seeds in an unsupervised manner, and use random walks method to obtain the segmentation results. In each saliency region, they apply the affinity propagation clustering to extract the representative pixels and obtain the seeds. A relabeling strategy is presented to ensure the extracted seeds inside the expected object. Scheier and Egner (1997) create a robot which visually approaches and selects objects. This is achieved by combining a segmentation and a selection mechanism. The segmentation mechanism uses synchronization of spiking neurons to bind image features corresponding to objects. The output of the segmentation serves as input to the selection mechanism which determines which object the robot will approach.

### 3.4 Tracking

To date, a vast number of tracking algorithms are developed for various applications. Many assumptions about objects, scenes and the camera movements are adopted to constrain tracking. The main advantage of using saliency is its ability to handle situations when an object appears in *different forms* and with *different background*.

Mahadevan and Vasconcelos (2009) propose a biologically inspired framework for visual tracking based on discriminant center surround saliency. The framework provides a principled unifying methodology to perform all three tasks involved in tracking: initialization, feature selection and target detection. At each frame, discrimination of the target from its background is posed as a binary classification problem. From a pool of feature descriptors for the target and the background, a subset that is most informative for classification between the two is selected using the principle of maximum marginal diversity. Using these features, the location of the target in the next frame is identified with saliency calculation, completing one iteration of the tracking algorithm. Frintrop and Kessel (2009) present a cognitive approach for visual object tracking from a mobile platform. The approach is based on a biologically motivated attention system that is able to detect regions of interest in images based on concepts of the human visual system. A top-down guided visual search module of the system enables to especially favor features that fit a previously learned target object. Here, the appearance of an object is learned online within the first image in which it is detected. In subsequent images, the attention system searches for the target features and builds a target-related saliency map. This enables to focus on the most relevant features with this object without knowing anything about a particular object model or scene in advance.

Klein et al. (2010) present a visual object tracker for mobile systems that is able to customize to individual objects during tracking. The core of their method is a novel observation model and the way it is automatically adapted to a changing object and background appearance over time. The system consists of a boosted ensemble of simple threshold classifiers built upon center-surround Haar-like features. Thus, the final algorithms are capable of processing video input at real-time. Borji et al. (2012) extend the works of Klein et al. (2010) and Frintrop et al. (2010) to deal with changing background by using a quick training phase with user interaction at the beginning of an image sequence. During this phase, some background clusters are learned along with foreground clusters. For the rest of the sequence the best fitting background cluster is determined for each frame and the corresponding object representation is used for tracking. The descriptor of an object is updated based on the cluster of the frame it appears in. Zhang et al. (2009) introduce a novel method of on-line object tracking with the static and motion saliency features extracted from the video frames locally, regionally and globally. Like the attention shifting mechanism of human vision, when the object being tracked disappears, their tracking algorithm can change its target to other objects automatically even without re-detection. Their algorithm has little dependence on the surface appearance of the object, so it can detect any category of objects as long as they are salient, and the tracking is robust to the change of global illumination and object shape. Stalder et al. (2012) propose dynamic objectness to sporadically re-discover the tracked object if it moves distinctly from its surroundings.

Li and Ngan (2008) introduce a method for human tracking. The method first generates a saliency map of the input video frame by using face tracking as the initial step for face segmentation in the subsequent frames. Next, a geometric model and an eye-map built from chrominance components are employed to localize the face region according to the saliency map. The final stage involves the adaptive boundary correction and the final face contour extraction. Later, Frintrop et al. (2010) introduce a component-based tracker. High contrast components in intensity and color channels are found and integrated in a descriptor. The descriptor captures the structure and appearance of a target in a flexible way. This descriptor can be learned quickly from a single training image and is easily adaptable to different objects. It is especially well suited to represent humans since they usually do not have a uniform appearance but, due to clothing, consist of different parts with different appearance.

For the task of vision-based autonomous driving, the goal is to control a robot vehicle by analyzing an image of the road ahead. Note that this task does not require prior landmarks as in localization task. Instead, the navigation should be chosen based on the location of important features like road edges. This is a difficult task since the scene ahead is often cluttered with distracting features such as other vehicles, pedestrians, trees, crosswalks, road signs and other objects that can appear or around a roadway. For the general task of autonomous navigation, these extra features are extremely important. Baluja and Pomerleau (1997) introduce the vision-based processing system for lane tracking that dynamically focuses only on the relevant inputs by masking out noise or distracting features. For lane marking detection, their algorithm is able to avoid being misdirected by distracting lane markings, passing cars, and other potentially confusing features.

### 3.5 Guided Attention Based Computer Vision Applications

There is a significant number of attentive systems exploiting guided attention, namely, task-driven saliency and human gaze in the computer vision area.

Task-driven saliency has shown to benefit many applications such as image question and answering (Yang et al. 2016), and person identification (Haque et al. 2016), handwriting synthesis (Graves 2013), machine translation (Bahdanau et al. 2014), image caption generation (Xu et al. 2015) and visual object classification (Mnih et al. 2014). There is a recent emerging trends on attention mechanisms in training neural networks, allowing models to learn alignments between different modalities, e.g., between visual features of a picture and its text description in the image caption generation task (Xu et al. 2015). In their work, as the model generates each word, its attention changes to reflect the relevant parts of the image. They propose two variant of attention, a "hard" attention mechanism and a "soft" attention mechanism. The soft attention refers to the global attention approach in which weights are placed "softly" over all patches in the source image. The hard attention, on the other hand, selects one patch of the image to attend to at a time. Similarly, Luong et al. (2015) extend this attention model to machine translation domain.

Inspired from the finding of Zhou et al. (2014) that object detector emerges in deep networks. Ren et al. (2017) propose an 'attention' Faster RCNN model for object detection. Using the recently popular terminology of neural networks with 'attention' (Xu et al. 2015) mechanisms, the proposed Region Proposal Network (RPN) module in Faster RCNN tells the state-of-the-art detector (Girshick 2015) module where to look. In particular, RPN takes an image (of any size) as input and outputs a set of rectangular object proposals, each with an objectness score.

Another application, person re-identification—an emerging trend in tracking applications, aims to re-identify human from different camera views. In this task, the human visual system can recognize person identities based on small salient regions, i.e., human saliency is distinctive and reliable in pedestrian matching across disjoint camera views. However, such valuable information is often hidden when computing similarities of pedestrian images with existing approaches. In Zhao et al. (2013a, b, 2015), saliency means distinct features that are (1) discriminative in making a person standing out from their companions, and (2) reliable in finding the same person across different views. For example, in Fig. 4, if most persons in the dataset wear similar clothes and pants, it is difficult to identify them. However, humans can easily identify the matching pairs due to the salient features, i.e., person (a1–b1) has a backpack with tilted blue stripes, person (a2–b2) has a red folder under her arms, and person (a3–b3) has a red bottle in his hand. Intuitively, if a body part is salient in one camera view, it is usually also salient in another camera view. In other words, if any region from person is so different from the others, its saliency value is very high. Thus these salient features are discriminative in distinguishing one from others and robust in matching them-

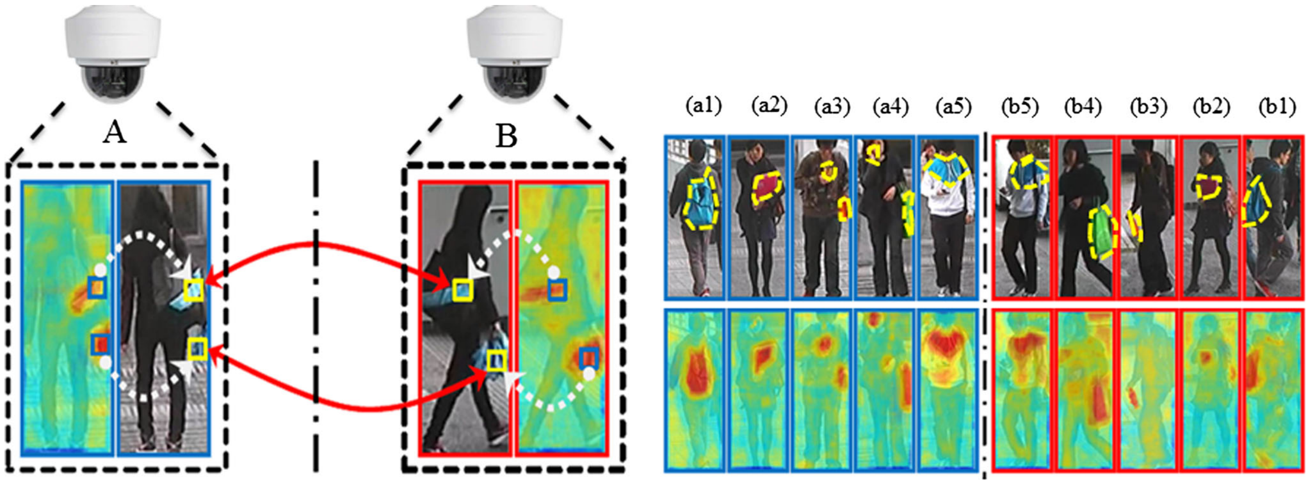

**Fig. 4** Examples of human image matching and saliency maps (image courtesy of Zhao et al. 2013a). Images on the left of the vertical dashed black line are from camera view **A** and those on the right are from camera view **B**. Upper part of the figure shows an example of matching based on dense correspondence and weighting with saliency values, and the lower part shows some pairs of images with their saliency maps

selves across different camera views. The authors figure out that clothes and trousers are generally the most important regions for person re-identification.

Regarding human-gaze based applications, Lee et al. (2012) develop region cues indicative of high-level saliency in egocentric video such as the nearness to hands, gaze, and frequency of occurrence and learn a regressor to predict the relative importance of any new region based on these cues. In a different work, Mishra et al. (2012) segment objects of interest by finding the "optimal" closed contour around the fixation point in the polar space. First, all visual cues are combined to generate the probabilistic boundary edge map of the scene; second, in this edge map, the "optimal" closed contour around a given fixation point is found. Recently, Xu et al. (2015) use gaze tracking information (such as fixation and saccade) significantly helps the summarization task. In particular, the gaze information allows meaningful comparison of different image frames and enables deriving personalized summaries (gaze provides a sense of the camera wearer's intent). Yun et al. (2013) find gaze to be a useful cue for image annotation, namely, outputing a set of object tags for an image. Papadopoulos et al. (2014) train object class detectors from eye tracking data in order to pursue the paradigm 'learning object detectors while watching TV.

## 4 Computer Graphics Applications

In this section, we review a variety of applications that utilize image/video manipulation under the saliency-based guidance. Here, visual attention implements a bottleneck mechanism, to focus resources on the most important part of images/videos. This is particularly helpful to handle the huge amount of image/video data.

### 4.1 Retargeting

Image retargeting sometimes is also referred as image cropping, thumbnailing, or resizing. The main idea of this saliency-based method is to remove indistinct regions and preserve the context with the most salient regions. Given the saliency map, Avidan and Shamir (2007) propose the Seam Carving method. Assume the given image is a landscape one where $n > m$, and the image is resized to the square size. The vertical seam $s$ is an 8-connected path in the saliency map $\boldsymbol{S}$ from the top to the bottom containing one pixel per row, is defined as below,

$$s = \{s_i\}_{i=1}^{m} = \{(x(i), i)\}_{i=1}^{m}, s.t.\forall i, |x(i) - x(i-1)| \leq 1.$$

The goal is to find the optimal seam that minimizes:

$$s^* = \min_{s} \sum_{i=1}^{m} \boldsymbol{S}(s_i),$$

where $\boldsymbol{S}(s_i)$ is one saliency pixel of the seam. This optimal seam can be found by dynamic programming. The process loops until the image reaches its expected square size. Figure 5 illustrates the general framework of image retargeting of Seam Carving process.

In a different approach, Setlur et al. (2005) propose using an importance map of the source image obtained from saliency and face detection. If the specified size contains all the important regions, the source image is simply cropped. Otherwise, the important regions are removed

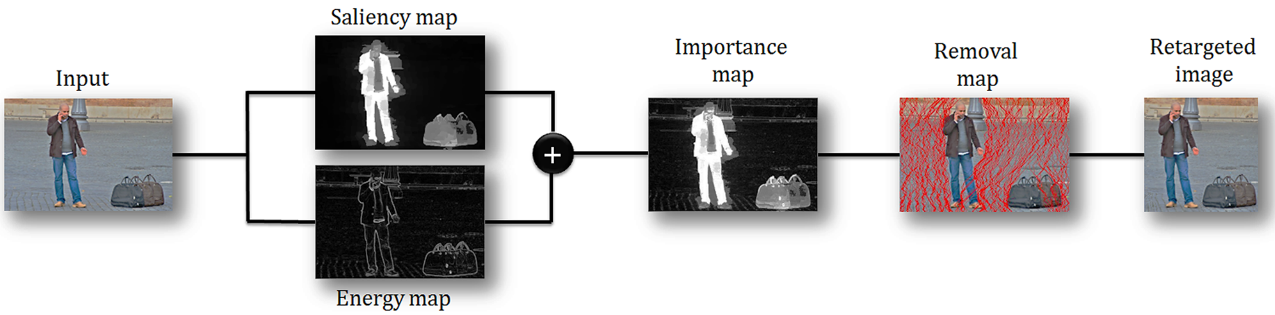

**Fig. 5** The flowchart of image retargeting. Given an input image, the importance map is first computed from the energy map and predicted saliency map. The removal map is later generated by seam carving operator, and the red lines are represented for the removal seams. The retargeted image is finally generated by removing the red lines (Color figure online)

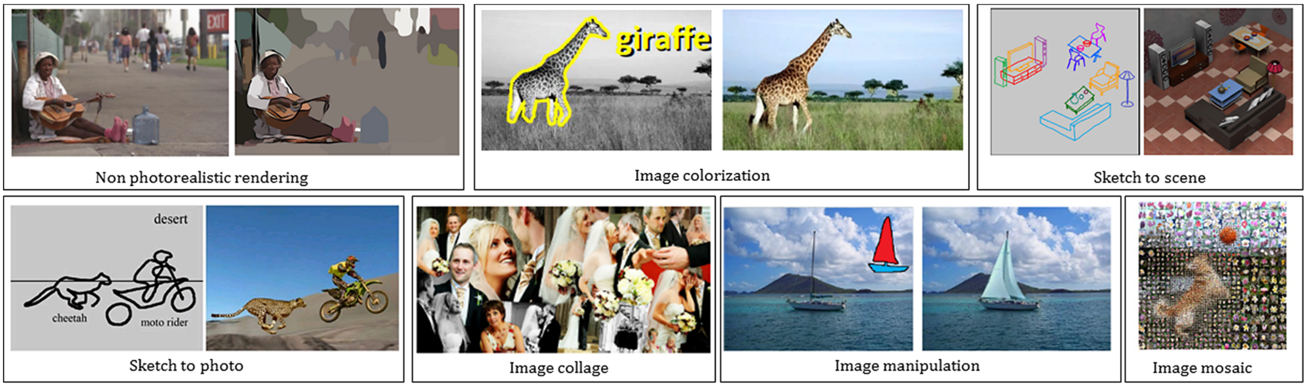

**Fig. 6** Sample applications of salient object detection. Images are credited from corresponding references (from left to right, top to bottom: DeCarlo and Santella 2002; Chia et al. 2011; Xu et al. 2013; Chen et al. 2009; Goferman et al. 2010; Goldberg et al. 2012; Margolin et al. 2013)

from the image, and fill the resulting “holes” using the background creation technique Later, Zhang et al. (2009) present an image resizing method that attempts to ensure that important local regions undergo a geometric similarity transformation, and at the same time, image edge structure is preserved.

While other works are dedicated to still pictures, Chamaret and Le Meur (2008) propose a video frame retargeting algorithm. The core of this algorithm consists of the extraction of a cropping window that is both related to the region of interest (ROI) and temporally smoothed in terms of location (center coordinates) and size by means of a strong temporal filtering. Temporal consistency is composed by two sequential steps: a Kalman filter is first applied in order to better predict the current samples. Then, a temporal filtering allows avoiding unlikely samples.

In order to facilitate image viewing on devices with limited display sizes, saliency map can be a very useful cue. Suh et al. (2003) propose a general thumbnail cropping method based on a saliency model that finds the informative portion of images and cuts out the non-core part of images. In Meur et al. (2006), the most important salient parts of the picture are cropped to fit the limited display size. Marchesotti et al. (2009) propose a framework for image thumbnailing based on visual similarity. The underlying assumption is that images sharing their global visual appearance are likely to share similar saliency values.

## 4.2 Image Manipulation

Image manipulation involves transforming or altering an image by using various methods and techniques to achieve desired results. Indeed, visual saliency can play a main role in this area. Some applications are shown in Fig. 6.

### *4.2.1 Image Montage*

Since a picture is said to be worth a thousand words, people often compose pictures to convey ideas. A common approach is to sketch a line drawing by hand, which is flexible and intuitive. An informative sketch, however, requires some artistic skill to draw, and line drawings typically limit the realism. An alternative approach known as photomontage uses existing photographs to compose a novel image to convey the desired concept. Chen et al. (2009) utilize online images as an enormous pool for image selection for photomontage. To achieve this, they search online for each scene item, and the background, using the text label. They only retain images with a clear and simple background, which greatly simplifies subsequent image analysis steps. This is achieved by the saliency filtering process to filter out images with a cluttered background. First, regions with high saliency values are computed for each image. Then, the process segments each image and counts the number of segments in a narrow band (of 30 pixels width) surrounding the highly salient region. If there are more than 10 segments in this band, the image is considered too complicated and discarded. During saliency filtering, each image is segmented to find scene elements matching items in the sketch. They then optimize the combination of the filtered images to seamlessly compose them, using the image blending technique. Meanwhile, Goldberg et al. (2012) present a framework for interactively manipulating objects in a photograph using related objects obtained from internet images. Given an image, the user selects an object to modify, and provides keywords to describe it. Objects with a similar shape are retrieved and segmented from online images matching the keywords, and deformed to correspond with the selected object. By matching the candidate object and adjusting manipulation parameters, their method appropriately modifies candidate objects and composites them into the scene. Supported manipulations include transferring texture, color and shape from the matched object to the target in a seamless manner. As in another fascinating application, Margolin et al. (2013) utilize saliency maps to do image mosaicing, which constructs an image using a dataset of images. In addition, they develop a cropping tool that automatically crops out the non-salient regions of an image.

#### *4.2.2 Image Collage*

Image collage is one type of visual image summary to arrange all input images on a given canvas, allowing overlay, to maximize visible visual information. The approach produce collages that are supposed to be informative, compact, and eye-pleasing. Wang et al. (2006) develop a picture collage method that considers the following properties. (1) Saliency maximization means that a picture collage should show as many visible salient regions (without being overlaid by others) as possible. (2) Blank space minimization indicates that a picture collage should make the best use of the canvas. (3) Saliency ratio balance means that each image in the collage has a similar saliency ratio (the percentage of visible salient region). (4) Orientation diversity illustrates that the orientations of the images are diverse. This property is used to imitate the collage style created by humans. In another work, instead of keeping the salient regions as rectangles, Goferman et al. (2010) use saliency for object cutout to embed the salient objects into the final collage. Instead of confining the collages in a regular canvas, Huang et al. (2011) present a novel approach for creating a fantastic collage artform, namely Arcimboldo[2]-like collage, which represents an input image with multiple thematically-related cutouts from the filtered internet images.

### 4.3 Rendering

Abstraction results in an image that directs users's attention to its most meaningful places and allows users to understand the structure there without conscious effort (Perazzi et al. 2012). Therefore, DeCarlo and Santella (2002) describe a computational approach to stylizing and abstracting photographs that explicitly responds to the design goal, to clarify the meaningful structure in an image. Their system transforms images into a line-drawing style using bold edges and large regions of constant color. It identifies the meaningful elements of this structure using saliency and a record of a user's eye movements in looking at the photo. The system renders a new image using transformations that preserve and highlight these visual elements. In another interesting work, El-Nasr et al. (2009) propose a system that adapts lighting specifically to direct participants' attention to important areas in real-time while maintaining visual continuity.

As image colorization can bring a grayscale photo to life, Chia et al. (2011) propose a method that utilizes internet photos and image filtering to minimize user effort and facilitate accurate color transfer. They first download a set of internet images with user-supplied text labels. They next select internet images using saliency filtering as done in Chen et al. (2009). Salient foreground objects are segmented automatically from these images by applying the saliency detector in Liu et al. (2011) and the Grabcut algorithm (Rother et al. 2004).

In computer graphics, image super-resolution refer to techniques that enhance the resolution of an image for a better rendering. Sadaka and Karam (2009) propose an attentive super-resolution technique that exploits the available saliency information of the active pixels to further reduce the computational complexity accompanied by imperceptible loss in the desired visual quality of the high-resolution image. During each iteration, only a subset of active pixels are selected for super-resolution processing based on a locally computed difference threshold criterion. The active pixels are further reduced by classifying them into background and foreground areas using visual attention information. The attended regions are further iterated upon in order to achieve a higher accuracy in these regions by setting a lower stopping threshold as compared to the background non-attended region. Jacobson et al. (2010) propose an algorithm for

[2] An Italian painter best known for creating imaginative portrait heads made entirely of objects such as fruits, vegetables, flowers, fish, and books.

improving both the objective and subjective quality by refining the motion vector field. They first utilize a discriminant saliency classifier to determine which regions of the motion field are most important to a human observer. These regions are refined using a multi-stage motion vector refinement that promotes motion vector candidates based on their likelihood given a local neighborhood. For regions that fall below the saliency threshold, a frame segmentation is used to locate regions of homogeneous color and texture via Normalized Cuts.

#### *4.3.1 Guided Attention Based Computer Graphics Applications*

Regarding task-driven saliency, Xu et al. (2013) introduce "Sketch2Scene", a framework that automatically turns a freehand sketch drawing inferring multiple scene objects to semantically valid, well arranged scenes of 3D models. This is enabled by summarizing functional and spatial relationships among models in a large collection of 3D scenes as structural groups. Object co-occurrence frequency is adopted to capture the reliability or saliency of a structural group. Lee et al. (2005) introduce the idea of mesh saliency as a measure of regional importance for graphics meshes. Mesh saliency at each rendering scale is computed as difference of Gaussian. The method generates and renders less detailed representations for small, distant, or unimportant portions of the scene. Luebke (2016) utilizes the human gaze information for foveated rendering in virtual reality. Basically the proposed method synthesizes the virtual environment with progressively less detail outside the eye fixation region. This aims to significantly speed-up for wide field-of-view displays, such as head mounted displays, where target frame rate and resolution is increasing faster than the performance of traditional real-time renderers.

It is also worth noting that human gaze is a powerful method by which emotions are expressed. There are many automatic approaches to transfer human gaze motion to animated characters. These approaches seek to analyse gaze motions in animated films to create animation models that can automatically map between emotions and gaze animation characteristics (Lance et al. 2004; Queiroz et al. 2007; Lance and Marsella 2010).

## 5 Multimedia Applications

In this section, we review applications in multimedia domain. We note that the difference between vision and graphics would be relatively clear, but the fundamental difference of multimedia from vision and graphics is not always obvious. In this section, we consider the applications that require the presentation of media in the combination of multi-modality, i.e., text, image, video, and audio. In addition, one notable point of multimedia applications is that most of those work require the subjective evaluation, i.e., user study with questionnaire or user-based assessment.

### 5.1 Multimedia Compression

As studied in Simoncelli (1996), when humans look at natural images or video clips, only a small region around the center of their eye fixation is captured at high resolution with logarithmic resolution falloff with eccentricity because of the nonuniform distribution of photoreceptor on the human retina. Ouerhani et al. (2001) use saliency maps to favor the preservation of perceptually important image details. Itti (2004) use saliency maps for MPEG-1 and MPEG-4 video compression.

Video summarization is considered as an application in multimedia compression. Ma et al. (2005) propose a feasible solution for video summarization, including key-frame selection and video skim extraction, based on user attention model, which does not require sophisticated heuristic rules or full semantic understanding. In Ji et al. (2013), representative frames are first selected at the shot level. The attention regions in representative frames are detected via a attention model. Finally, the visual features of attention regions are clustered in an online manner to reduce memory cost.

### 5.2 Multimedia Retrieval

An image retrieval system is a computer system for browsing, searching and retrieving images from a large database of digital images. Similarity measure is the main key in the content based information retrieval. Stentiford (2003) treats the similarity measure as a problem of distinguishing similar shapes in sets of black and white symbols. Feng et al. (2010) use both salient edges and regions extracted for images similarity comparison. Li et al. (2013) extracted the visually salient regions in the images as retrieval units. They represent each region using a bag-of-word model, and the method takes advantage of group sparse coding to encode the visual descriptor, achieving a lower reconstruction error and obtaining a sparse representation at the region level. Gao et al. (2015) integrate visual image re-ranking by exploiting saliency in the database. In particular, the bottom-up saliency mechanism computes the database saliency value of each image by hierarchically propagating a posterior probability in it, while the top-down saliency mechanism discriminatively expands the query from top-ranked images after the initial search.

## 5.3 Quality and Aesthetics Assessment

The main reason that we consider quality and aesthetics assessment in the ‘multimedia’ category is the natural user-based evaluation.

### *5.3.1 Quality Assessment*

Visual quality assessment, i.e, image or video quality assessment, is a critical issue in practical applications such as data acquisition, transmission, restoration, compression, and enhancement, etc. Ninassi et al. (2007) showed that applying the visual attention to image quality assessment is not trivial, even with the ground truth. In Liu and Heynderickx (2009), Liu and Heynderickx discover that visual saliency impacted the objective image quality. Li et al. (2013) propose a novel image quality assessment method using saliency. Different weights are assigned to extracted salient regions and non-salient regions.

### *5.3.2 Aesthetics Assessment*

According to statistics quoted by Facebook, an average of 350 million new photos are uploaded daily by its users. Thus, there is a great demand for multimedia applications to manage, assess and edit such content. Photo-quality assessment and improvement are two areas that have particularly attracted research attention (Bhattacharya et al. 2010) apply a saliency map to estimate visual attention distribution in photographs in order to infer the geometric context of a scene. With the help of the above methods, they extract aesthetic features that could be used to measure the deviation of a typical composition from ideal photographic rules of composition. These aesthetic features are subsequently used as input to two independent regressors in order to learn the visual aesthetic model. This learned model is then integrated into their photo-composition enhancement framework. Wong and Low (2009) present a saliency-enhanced method for the classification of professional photos and snapshots. First, they extract the salient regions from an image by utilizing Itti’s model (1998). The salient regions are assumed to contain the foreground objects/ humans. Then, in addition to a set of discriminative global image features, a set of salient features is extracted in order to characterize the subject and depict the subject-background relationship. Later, Wong and Wong (2012) present a semi-automatic photographic recomposition approach that employs a semantics-preserving warp of the input image to enhance the visual dominance of the main subjects. Their method uses the tearable image warping method to shift the subjects against the background (and vice versa), so that their visual dominance is improved, and yet preserve the desired spatial semantics between the subjects and the background. In a practical application, Gadde and Karlapalem (2011) build a robot that can replace a human in capturing quality photographs for publishing. The image quality assessment approach is based on few high level features of the image combined with some of the aesthetic guidelines of professional photography e.g., rule of thirds and golden ratio.

## 5.4 Advertisement-Driven Applications

Advertisement sense or virtual content insertion is an emerging application of video analysis and is used in video augmentation and advertisement insertion. Some applications, e.g., image sense, video sense, 3D sense, and subtitle embedding, seek for the least salient regions to embed contents such as advertisement. Liu et al. (2008) first propose a generic virtual content insertion system based on visual attention analysis. In Li et al. (2008), introduce a method to embed the advertisement into webpages. The ads are selected based not only on textual relevance but also visual similarity, so that the ads yield contextual relevance to both the text in the website/page and the image content. Basically, the ads are inserted into the non-salient positions and they are assumed to have similar appearance with the neighboring blocks around the insertion position. Given the image $I$ divided into square blocks, Li et al. (2008) proposed to find the position $R_c(b_i, a_j)$ between a candidate ad insertion block $b_i$ and ad $a_j$ as: $R_c(b_i, a_j) = (1-s_i)\times(1-d(B_i, a_j))$, where $s_i$ is the saliency value of the block $i$. $d(B_i, a_j) = \frac{1}{|B_i|}\sum_{b_i \in B_i} \| f_{b_i} - f_{a_j} \|_2$, $|B_i|$ denotes the number of blocks adjacent with $b_i$ , $f_{b_i}$ and $f_{a_j}$ denote the feature of block $b_i$ and ad $a_j$, respectively. Later, different systems are developed for different domains, e.g., image, game, video (Li et al. 2008, 2010a, b; Mei et al. 2012; Nguyen et al. 2012). Figure 7 illustrates some case studies of advertisement-targeted applications.

In addition, assisting the disabled people by applying computer vision/multimedia techniques consistently attracts the attention from many researchers. Recently, a technique for assisting hearing impairment patients in watching videos (Hong et al. 2010) is developed, which automatically inserts the dialogue near the talking persons to help the disabled understand who is talking and the content of the dialogue. However, there is often a need to insert the subtitle into the video without human appearance (i.e., only narration appears in the video), such as documentary and introductory films. Nguyen et al. (2013) introduce an application that automatically inserts the subtitle into such videos based on the video saliency map intelligently, in order to help the patients understand the content of the narration. The basic criteria of the subtitle insertion are twofolds. Firstly, the selected position to insert the subtitle should have a low saliency score. Otherwise, the inserted subtitle will overlap with the salient objects and disturb the watching experience of the audience. Second,

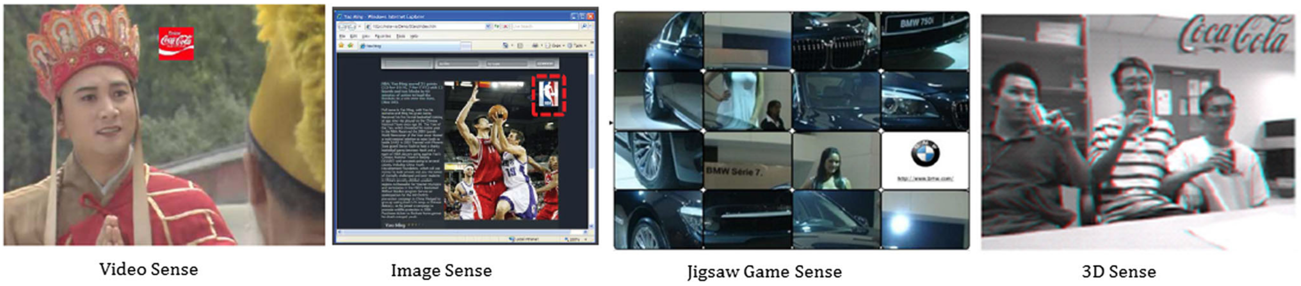
Video Sense

Image Sense

Jigsaw Game Sense

3D Sense

**Fig. 7** The exemplary advertising systems. The systems aim at seamlessly embedding the advertisements at an appropriate (non-intrusive) position within the webpage, image, or video frame. The additional contents are embedded into the least salient regions. Image courtesy of Liu et al. (2008), Li et al. (2008, 2010a), Nguyen et al. (2012)

the selected position should be near to the high saliency position. Thus the inserted subtitle will not distract the audience's attention.

### 5.5 Guided Attention Based Multimedia Applications

There exist some task-driven saliency based multimedia applications in literature. Gautier et al. (2012) consider the depth map as the important map for encoding algorithm. The method aims at exploiting the intrinsic depth maps properties since depth images indeed represent the scene surface and are characterized by areas of smoothly varying grey levels separated by sharp edges at the position of object boundaries. Preserving these characteristics is important to enable high quality view rendering at the receiver side. Muratov et al. (2012) utilize saliency detection as a support for image forensic. They assume the images used for the forged creation have different JPEG compression qualities, then there exist inconsistencies within and outside the tampered region. Therefore the tampered region with different image compression (e.g., JPEG) qualities can be detected by analyzing the differences between the original image and its JPEG-compressed versions. Gupta et al. (2013) also propose a novel video compression architecture, incorporating saliency, to save significant amount of computation. This architecture is based on thresholding of mutual information between successive frames for flagging frames requiring re-computation of saliency, and use of motion vectors for propagation of saliency values.

The human gaze information is usually used in multimedia applications in order to to better understand user's intentions, as implicit input in gaming or as automatic tagging and context recognition tool during everyday life (Ishiguro etal. 2010). From human gaze data, Shen and Zhao (2014) experiment that human attention usually focuses on large texts, logos, faces and objects that near the center or the top-left regions in the webpage. In Alkan and Cagiltay (2007), Alkan et al. use the gaze information to study how computer gamers explore a computer game that they do not know how to play, in a naturalistic manner. Drewes et al. (2007) found that eye-gaze interaction for mobile applications is attractive to users and that the gaze gestures are an alternative method for eye-gaze based interaction.

## 6 Miscellaneous Applications

In this section, we review the attentive systems unclassified in preceding three categories, namely, computer vision, computer graphics and multimedia.

### 6.1 Human Robot Interaction

The first kind of application is human robot interaction, which is the study of interactions between humans and robots. With the advances of technologies, the autonomous robots could eventually have more proactive behaviors. By embedding a saliency based attentional model, the robot is able to "see the scene as the way human sees" and engage in an interaction with a human. Muhl et al. (2007) present an interesting sociological study in which the interaction of a human with a robot simulation is investigated. The user interface shows a robot face which people can communicate. The robot face interacts with a human partner by changing its gaze direction as well as facial expression in response to visual input. The gaze direction is controlled so that the partners are able to perceive that the robot is looking at an interesting location in the environment. The qualitative analysis revealed that people established a communicative space with their robot and accepted it as a proactive agent. Meger et al. (2008) develop Curious George, an intelligent system that attempts to perform robust object recognition in a realistic scenario, where a mobile robot moving through an environment must use the images collected from its camera directly to recognise objects. To perform successful recognition, they choose a combination of techniques including a peripheral-foveal vision system, an attention system combining bottom-up visual saliency with structure from stereo, and a localisation and mapping technique. The result is a highly capable object recognition system that can be easily trained to locate the objects of interest in a particular region, and to subsequently build a spatial-semantic map of the region.

Dankers et al. (2007) develop a synthetic active visual system capable of detecting and reacting to unique and dynamic visual stimuli, as well as capable of being tailored to perform basic visual tasks. The system is able to direct its attention towards previously unattended salient objects/regions. Upon saccading to a new target, it extracts the object that attracts attention whereas maintaining stereo fixation on that object, regardless of its shape, colour or motion. Belardinelli (2008) presents a robot that learns visual scene exploration by imitating human gaze shifts. Nagai (2009) develop an action learning model for robots based on spatial and temporal continuity of bottom-up features. The proposed system can extract key actions from human action demonstrations so that robots can imitate. Frintrop (2011) envision how future ways to obtain attentive robots might look like. Coury and Marchand (2003) develop a simulation of the visual perception of a synthetic actor. Breazeal et al. (1999) present a visual attention system based on a model of human visual search behavior from Wolfe (1994). The attention system integrates perception inputs (i.e., motion detection, color saliency, and face popouts) with habituation effects and influences from the robot's motivational and behavioral state to create a context dependent attention activation map. This activation map is used to direct eye movements and to satiate the drives of the motivational system. Vijayakumar et al. (2001) investigate the interplay between oculomotor control, visual processing, and limb control in humans and primates by exploring the computational issues of these processes with a biologically inspired artificial oculomotor system on an anthropomorphic robot. Stimuli in the environment excite a dynamical neural network that implements a saliency map, i.e., a winner-take-all competition between stimuli while simultenously smoothing out noise and suppressing irrelevant inputs. In real-time, this system computes new targets for the shift of gaze, executed by the head-eye system of the robot. The redundant degrees-of-freedom of the head-eye system are resolved through a learned inverse kinematics with optimization criterion. For humans, an important capability for joint attention is to follow the pointing gesture with fingers. Approaches to endow robots with a similar capability are proposed by Heidemann et al. (2004). They analyze the direction of a pointing finger and fuse this cue with bottom-up saliency maps. They present a system which uses an attention map as a representation of focus of attention. Also, the attention map allows the future integration of symbolic information from speech recognition systems.

### 6.2 Attention Retargeting

Rosenholtz et al. (2011) find that users typically first pick out more regions with high salient values then look for what they correspond when using a user interface. This raises a new line of applications that aim at changing saliency to modulate human attention for the pre-defined tasks, i.e., aesthetic enhancement, advertisement attraction, etc. Saliency retargeting aims at changing image saliency for enhancing image aesthetics. Wong and Low (2011) propose saliency retargeting as a means for image aesthetics enhancement as shown in Fig. 8. Given an image $I$ and a set of $N$ object segments with target importance value $T_i$ for each object segment $i$, they aim at enhancing the aesthetics of the image by applying a set of low-level image modifications $x$ to the input image $I$ to produce an output image with saliency value $S_i$ that matches the target importance value $T_i$ for each object segment $i$. The saliency retargeting is formulated as a constrained optimization problem as: $\min_x f(x) = \sum_{i=1}^{N} |q(T_i) - q(S_i)|$, $x = \{v_i, s_i, \sigma_i | i = 1, 2, \ldots, N\}$, $v_i, s_i, \sigma_i$ is the increase of average luminance, color saturation, sharpness in segment $i$, $s_i$ is the increase of average color saturation in object segment $i$, respectively, and $q(.)$ is the normalization function. The saliency retargeting modifies only the low-level image features that correspond directly to the features used in saliency computation.

Bailey et al. (2009) deploy subtle modulations to the peripheral regions of the field of view to draw the viewer's foveal vision to the modulated region. The modulations are

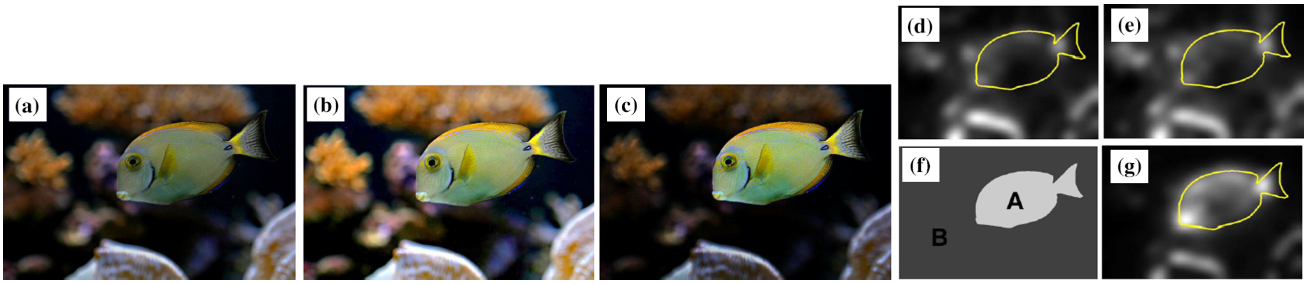

**Fig. 8** **a**, **d** Original image and its saliency map. **b**, **e** Globally-enhanced image and its saliency map. **c**, **g** Image enhanced by saliency retargeting and its saliency map. **f** Object segments, where Objects **A** and **B** are in the reverse order of importance. Images courtesy of Wong and Low (2011)

simply alternating interpolations of the pixels in the predetermined area of interest with a warm and a cool color. Similarly, Tanaka et al. (2015) propose a method to induce users to look at a selected point in virtual space during uninterrupted viewing by shifting the virtual angular direction. Nguyen et al. (2013) propose a new framework to alter human attention by re-coloring the desired regions. Later, Mateescu and Bajić (2014) propose a method that modifies the color of a selected region in an image to increase its saliency and draw attention towards it. They describe the hue content of a ROI and its surroundings using a polar representation of a perceptually uniform color space, which allows them to easily determine the optimal hue adjusment to maximize the dissimilarity between the ROI and its surroundings.

### 6.3 Saliency-Based Eye Tracker Calibration

In literature, there are many research works on saliency-based eye tracker calibration (Chen and Ji 2001; Choi et al. 2014; Sugano et al. 2010; Nguyen et al. 2013; Perra et al. 2015). Sugano et al. (2010) propose a calibration-free gaze sensing method using visual saliency maps. Their goal is to construct a gaze estimator only using eye images captured from a person watching a video clip. To efficiently identify gaze points from saliency maps, they aggregate saliency maps based on the similarity of eye appearances. Mapping between eye images to gaze points is established by Gaussian process regression. Similarly, Chen and Ji (2001), Chen and Ji (2015) introduce a probabilistic approach to online eye gaze tracking without explicit personal calibration. Meanwhile, Choi et al. (2014) and Nguyen et al. (2013) propose using GMM-based saliency aggregation and particle filter for calibration-free gaze tracking, respectively. In another work, Perra et al. (2015) develop a calibration scheme allows a headworn device to calculate a locally optimal eye-device transformation on demand by computing an optimal model from a local window of previous frames.

## 7 Discussions and Conclusion

In a nutshell, we intensively review attentive systems, i.e., applications that exploit visual saliency analysis. We review a large body of works relating to saliency applications and discuss the role of saliency in the applications. The attentive systems are categorized into four areas including: computer vision, computer graphics, multimedia, and miscellaneous. We believe this survey offers a comprehensive overview and suggests important insights for the next generation of applications based on visual saliency analysis. In the following, we summarize our main findings as follows:

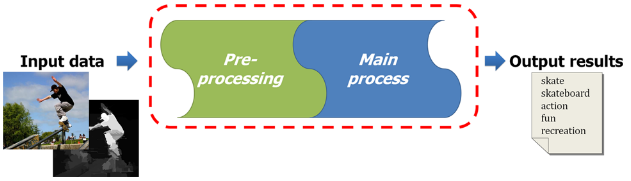

**Fig. 9** The flowchart of one attentive system pipeline where the saliency can be exploited. The red rectangle highlights the contributions of saliency into the traditional pipeline

1. Saliency maps in general can be considered as a reliable cue to applications that process important or distinctive regions in the images. They can be seamlessly integrated into the conventional pipeline of different problems. Figure 9 depicts the flowchart of one image recognition pipeline where the saliency can be exploited. Saliency can be used in the pre-processing and the main process of the system. Table 1 highlights some case studies of saliency in different attentive systems. The obvious advantages can be multi-fold, for example, building efficient systems, removing background noise, and guiding user attention. Since image retargeting is one popular application to demonstrate the effectiveness of a new saliency prediction model, we are interested in investigating the impact of each type of saliency maps in this problem. We perform image retargeting as described in Sect. 4.1 on MSRA-1000, ECSSD, and iCoSeg datasets (Achanta et al. 2009; Yan et al. 2013; Batra et al. 2010) (with the computational models introduced in Sect. 2). As seen in Fig. 10, the retargeting results from salient object detection methods well preserve the main salient objects without distortion. Therefore, the explanation of using a certain saliency prediction model may provide good practice for further applications. While fixation prediction is in general biologically plausible and suggests important regions the same way as humans look at, salient object detection can be used for more task-specific applications that requires full cutout of the salient objects.
2. Saliency can be used to overcome the speed bottleneck problem, for example, the sliding window approach in object detection task can be facilitated by using a generic attention operator to quickly select a few interest regions in the image.
3. Saliency-based descriptor extraction is able to remove the noisy data from the dense sampling. For example, in action recognition task, a large amount of densely extracted descriptors is indeed unnecessary and may even be harmful for the recognition performance.
4. The computation of attention regions will also improve the performance of other modules since more processing resources can be provided to essential parts of the sensory input.

**Table 1** The summary of exemplary saliency usage in different applications

| Pipeline component | Saliency integration | Human fixation prediction | Salient object detection |
|---|---|---|---|
| Pre-processing | Feature selection criterion (Kadir and Brady 2001; Frintrop and Jensfelt 2008; Borji et al. 2012) | ✓ | ✓ |
| | Seed generation for segmentation (Johnson-Roberson et al. 2010; Qin et al. 2014) | | ✓ |
| | Hypotheses generation for object detection (Viola and Jones 2004; Fritz et al. 2004) | ✓ | ✓ |
| | Segment selection in object discovery (Karpathy et al. 2013; Frintrop et al. 2014) | | ✓ |
| | Saliency filtering in visual computing (Chia et al. 2011; Chen et al. 2009) | ✓ | ✓ |
| | Cue for data compression (Ouerhani et al. 2001; Itti 2004) | ✓ | |
| | Cue for content embedding (Li et al. 2008, 2010a, b; Mei et al. 2012) | ✓ | |
| Main process | Feature pooling for image/video recognition (Chen et al. 2012; Nguyen et al. 2015) | ✓ | ✓ |
| | Feature matching (Borji and Itti 2011; Rutishauser et al. 2004) | | ✓ |
| | Scanpath prediction in object detection (Butko et al. 2009) | ✓ | ✓ |
| | Dynamic objectness in object tracking (Stalder et al. 2012) | | |
| | Rule of third to automatically take photos (Gadde and Karlapalem 2011) | ✓ | |
| | Input data of image retargeting (Avidan and Shamir 2007; Achanta et al. 2009) | | ✓ |
| | Saliency ratio for image collage (Goferman et al. 2010) | | ✓ |
| | Input data of image aesthetic prediction (Wong and Wong 2012; Wong and Low 2009) | ✓ | |

We only include some notable works for each type of application

5. The viewpoint of using saliency maps is different from different problems. For example, the image resizing task aims to preserve the most salient regions. In contrast, the advertisement embedding task looks for and leverage the least salient regions. Meanwhile, the feature pooling attempts to separate the most/least regions in order to pool features into different channels in terms of saliency values (Table 2).
6. The view of using salient features in person re-identification is novel since the solution focuses on the high-level saliency, i.e., bags or clothes which makes a person standing out from their companions.
7. Besides adopting saliency maps, an emerging trend is saliency retargeting. This task aims to modify saliency/attention in order to benefit certain applications such as advertisement-oriented applications.
8. We are interested in the contribution of visual saliency to attentive systems. As shown in Table 3, we summarize some of the roles of visual saliency in performance, speed, and subjective experiences. Although using saliency maps for feature extraction was indeed expected to improve the performances (e.g., scene recognition task), most of the state-of-the-art recognition models are not taking such an approach. For holistic image recognition, it would be more common understanding that densely sampling visual features from the whole image region is a better strategy. Therefore, we also conduct an experiment where we extend the state-of-the-art scene recognition (Zhou et al. 2014) by extracting deep learned features from foreground/background regions and the whole image. In the original work, Zhou et al. (2014) use the implementation of Jia et al. (2014) with the learned model to extract features from the layer just before the final classification layer (often referred as fc7), resulting in a feature dimension of 4096. In our reproduced work, we extracted learned features from

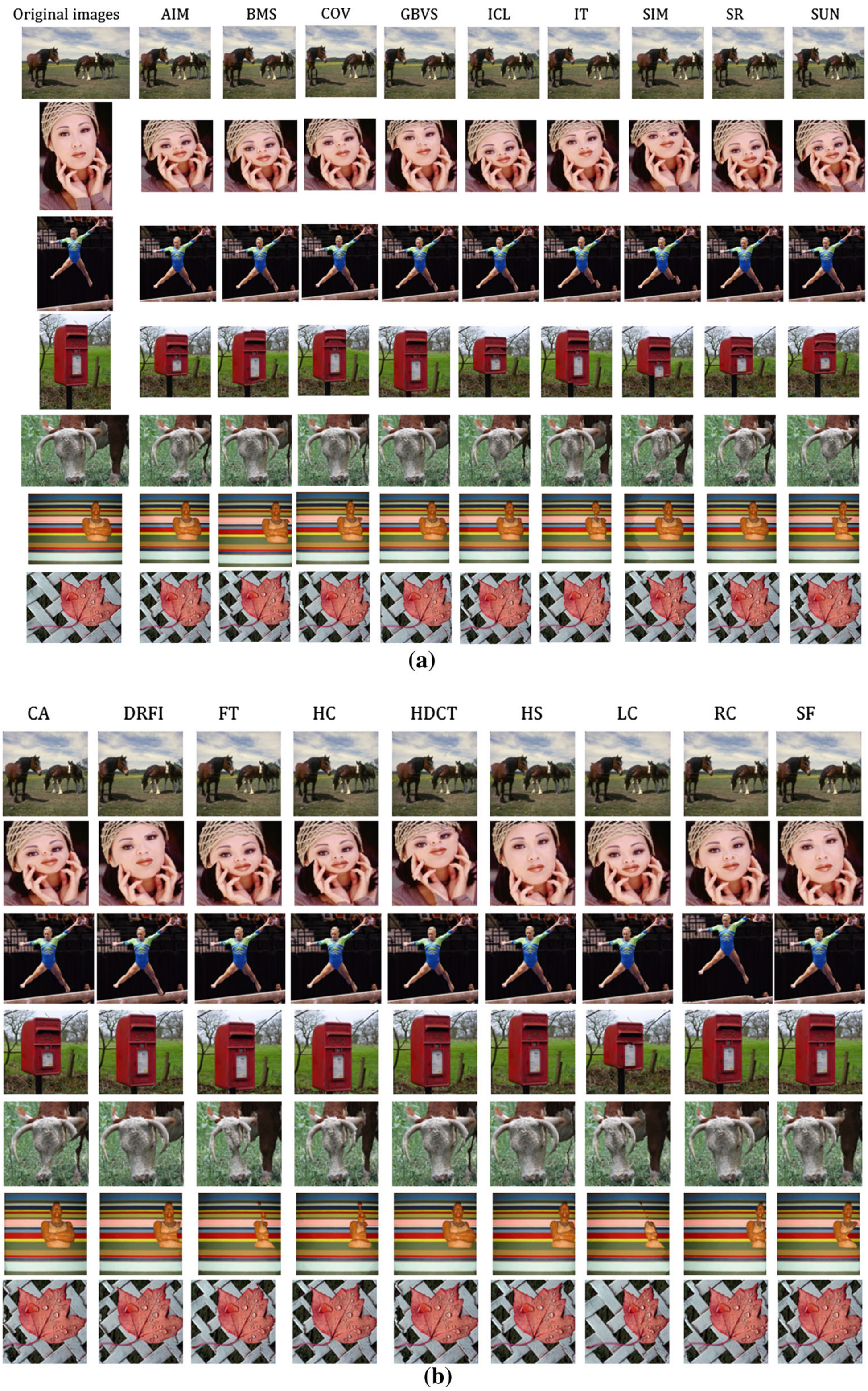

**Fig. 10** Visual comparison of retargeting images on MSRA-1000, ECSSD, and iCoSeg datasets (Achanta et al. 2009; Yan et al. 2013; Batra et al. 2010). Salient object detection-based results preserve the main salient objects, which is suitable for image retargeting application (Please view in high 400% resolution for best visual effect). **a** The orignal images and the retargeting results of 9 fixation prediction methods **b** The retargeting results of 9 salient object detection methods

**Table 2** Comparison of running times in the benchmark (Achanta et al. 2009)

| Method | IT | GB | SR | SUN | AIM | CA | FT | RC | HC | SF | HS | DRFI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Time (s) | 0.24 | 1.61 | 0.06 | 1.12 | 4.28 | 51.2 | 0.01 | 0.14 | 0.01 | 0.15 | 0.43 | 0.18 |
| Code | M | M | M | M | M | M | C++ | C++ | C++ | C++ | C++ | C++ |

M indicates the code is written in MATLAB

**Table 3** The contribution of visual saliency to attentive systems

| | Performance | | | Speed-up | Subjective experiences |
|---|---|---|---|---|---|
| | Object recognition on PASCAL VOC 2010 (Everingham et al. 2010) (mAP) (Chen et al. 2012) (%) | Action recognition on YouTube dataset (accuracy) (Nguyen et al. 2015) (%) | Scene recognition on 15-scene dataset (Lazebnik et al. 2006) (accuracy) (Zhou et al. 2014) (%) | Face detection (Butko et al. 2009) | Saliency retargeting (correlation coefficient) (Wong and Low 2011) |
| Without saliency | 74.5 | 84.2 | 90.2 | 1× | 0.284 |
| With saliency | 77.2 | 87.9 | 92.5[a]<br>(92.9) | 1.93× | 0.546 |

mAP: mean of Average Precision, is calculated by taking the mean of the average precision of each object category. Accuracy is actually the average accuracy over all action classes. Correlation coefficient: correlation between the number of fixations on each object segment. [a]: our reproduced results with saliency is used as a side information for feature pooling. Note that visual saliency leads to remarkable improvement in performance, speed and subjective experiences

the salient regions. Note that the learned model and the feature dimensionality are exactly the same with the one from Zhou et al. (2014). The corresponding performance is 92.5% leading the performance of the features extracted from the global image (90.2%). In other words, saliency-based feature extraction is still very important to improve the performance of state-of-the-art methods. In addition, we also did another experiment with learned features extracted from salient (foreground) and non-salient (background) regions, respectively. We observed that pooling from both foreground and background obtains better results, 92.9%, than that from the foreground only (92.5%). This observation is consistent with the finding in Chen et al. (2012), Nguyen et al. (2015).

However, from the survey listed above, there are still some critical limitations within the existing saliency-based applications.

1. To date, there is still no saliency prediction model fitting all of the applications. From the survey, different saliency detection methods suit for different applications.
2. We are interested in the most frequently used saliency models in existing attentive systems. Therefore, we list how many times the most famous saliency models (e.g., IT, GB, CA, SR, SF, DRFI) are used in existing attentive systems in Fig. 11. Itti's pioneering method (Itti et al. 1998) has been widely adopted in the early stage of attentive systems. In the past few years, some newly introduced models such as SF (Perazzi et al. 2012), DRFI (Jiang et al. 2013) have been favored due to their high performance as studied in Borji et al. (2015).
3. Last but not least, the computational time of the saliency models needs to be taken into consideration especially when embedding visual saliency into practical applications that require real-time processing. Table 2 compiles the average running time of the state-of-the-art methods on the benchmark images (Achanta et al. 2009) (the

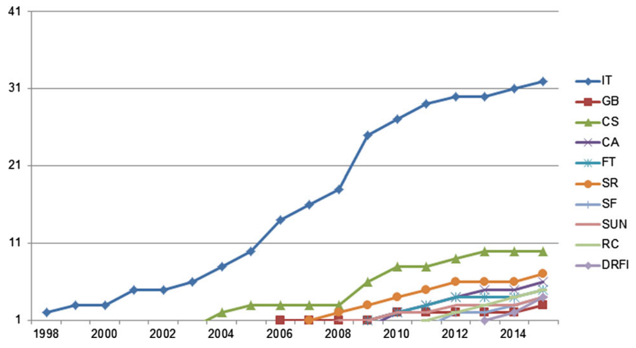

**Fig. 11** The accumulative usage of various saliency models in attentive systems

image's size is roughly 400 × 300). Currently most of the available code of saliency detection is in MATLAB or yet optimized C++, therefore, it requires a lot of engineering work to realize the research work into the practical systems.

In future, the following research directions may play important roles in practice:

1. Visual saliency will be adapted into new applications, i.e., autonomous driving car, digital image forensics. Indeed visual saliency can be used in many other systems not restricted to the aforementioned areas.
2. While good results are obtained in some areas, it is still a long way to obtain a perfect attentive system. Among the parts that are still missing is certainly a close interaction between different modules. In computer vision, the common tasks such as object detection, segmentation, tracking, and categorization benefit strongly from each other if the modules collaborate and share information. Similarly, future attentive systems will strongly benefit from interacting modules. Contextual information and prior knowledge from other modules can enable an attentive system to obtain better, more useful regions of interest as discussed in Jiang et al. (2015b).
3. Visual saliency could be employed in different modalities apart from image or video, i.e., auditory perceptions, speech recognition, touch behavior among others.
4. An emerging trend of smart-glasses integrating eye gaze detector [3,4,5] promises to facilitate predicted saliency in the complex scenes where there are multiple objects in a complex background.
5. The field of visual attention still lacks computational principles for task-driven attention. A promising direction for future research is the development of systems that take into account time varying task demands, especially in interactive, complex, and dynamic environments.

## References

Achanta, R., Hemami, S. S., Estrada, F. J., & Süsstrunk, S. (2009). Frequency-tuned salient region detection. In *IEEE conference on computer vision and pattern recognition* (pp. 1597–1604).

Alexe, B., Deselaers, T., & Ferrari, V. (2012). Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligenc*, *34*(11), 2189–2202.

Alkan, S., & Cagiltay, K. (2007). Studying computer game learning experience through eye tracking. *BJET*, *38*(3), 538–542.

Avidan, S., & Shamir, A. (2007). Seam carving for content-aware image resizing. *ACM Transactions on Graphics*, *26*(3), 10.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR* (abs/1409.0473).

Bailey, R., McNamara, A., Sudarsanam, N., & Grimm, C. (2009). Subtle gaze direction. *ACM Transactions on Graphics*, *28*(4), 100.

Baluja, S., & Pomerleau, D. A. (1997). Expectation-based selective attention for visual monitoring and control of a robot vehicle. *Robotics and Autonomous Systems*, *22*(3), 329–344.

Batra, D., Kowdle, A., Parikh, D., Luo, J., & Chen, T.(2010). icoseg: Interactive co-segmentation with intelligent scribble guidance. In *IEEE conference on computer vision and pattern recognition* (pp. 3169–3176).

Belardinelli, A. (2008). Salience features selection: Deriving a model from human evidence. Ph.D. thesis, Sapienza Universita di Roma, Rome, Italy.

Bhattacharya, S., Sukthankar, R., & Shah, M. (2010). A framework for photo-quality assessment and enhancement based on visual aesthetics. In *ACM multimedia conference* (pp. 271–280).

Borji, A., Cheng, M., Jiang, H., & Li, J. (2015). Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, *24*(12), 5706–5722.

Borji, A., Frintrop, S., Sihite, D. N., & Itti, L. (2012). Adaptive object tracking by learning background context. In *IEEE conference on computer vision and pattern recognition workshops* (pp. 23–30).

Borji, A. & Itti, L. (2011). Scene classification with a sparse set of salient regions. In *IEEE international conference on robotics and automation* (pp. 1902–1908).

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 185–207.

Borji, A., Sihite, D. N., & Itti, L. (2012). Salient object detection: A benchmark. In *European conference on computer vision* (pp. 414–429).

Boykov, Y., & Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*, 1124–1137.

Breazeal, C., & Scassellati, B. (1999). A context-dependent attention system for a social robot. In *International joint conference on artificial intelligence* (pp. 1146–1153).

Bruce, N., & Tsotsos, J. (2005). Saliency based on information maximization. In *Advances in neural information processing systems*.

Butko, N.,& Movellan, J. (2009). Optimal scanning for faster object detection. In *IEEE conference on computer vision and pattern recognition* (pp. 2751–2758).

Chamaret, C., & Le Meur, O. (2008). Attention-based video reframing: Validation using eye-tracking. In *International conference on pattern recognition* (pp. 1–4).

Chen, J., & Ji, Q. (2011). Probabilistic gaze estimation without active personal calibration. In *IEEE conference on computer vision and pattern recognition, CVPR* (pp. 609–616).

Chen, J., & Ji, Q. (2015). A probabilistic approach to online eye gaze tracking without explicit personal calibration. *IEEE Transactions on Image Processing*, *24*(3), 1076–1086.

Chen, Q., Song, Z., Hua, Y., Huang, Z., & Yan, S. (2012). Hierarchical matching with side information for image classification. In *IEEE conference on computer vision and pattern recognition* (pp. 3426–3433).

Chen, T., Cheng, M.-M., Tan, P., Shamir, A., & Hu, S.-M. (2009). Sketch2photo: Internet image montage. *ACM Transactions on Graphics*, *28*, 124.

Chen, Y., Nguyen, T., Kankanhalli, M. S., Yuan, J., Yan, S., & Wang, M. (2014). Audio matters in visual attention. *IEEE Transactions on Circuits and Systems for Video Technology*, *24*(11), 1992–2003.

[3] http://www.google.com/glass.

[4] http://www.eyetracking-glasses.com/.

[5] http://www.tobii.com/en/eye-tracking-research/global/landingpages/tobii-glasses-2/.

Cheng, M., Mitra, N. J., Huang, X., Torr, P. H. S., & Hu, S. (2015). Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(3), 569–582.

Cheng, M.-M., Zhang, Z., Lin, W.-Y., & Torr, P. H. S. (2014). BING: Binarized normed gradients for objectness estimation at 300 fps. In *IEEE conference on computer vision and pattern recognition* (pp. 3286–3293).

Chia, A., Zhuo, S., Gupta, R. K., Tai, Y.-W., Cho, S., Tan, P., et al. (2011). Semantic colorization with internet images. *ACM Transactions on Graphics*, *30*, 1–7.

Choi, J., Ahn, B., Park, J., & Kweon, I. (2014). Gmm-based saliency aggregation for calibration-free gaze estimation. In *IEEE international conference on image processing* (pp. 1096–1099).

Choi, J., Oh, T., & Kweon, I. (2016). Human attention estimation for natural images: An automatic gaze refinement approach. *CoRR* (bs/1601.02852).

Courty, N., & Marchand, E. (2003). Visual perception based on salient features. In *International conference on intelligent robots and systems* (Vol. 1, pp. 1024–1029).

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE conference on computer vision and computer vision* (pp. 886–893).

Dankers, A., Barnes, N., & Zelinsky, A. (2007). A reactive vision system: Active-dynamic saliency. In *International conference on computer vision systems*.

DeCarlo, D., & Santella, A. (2002). Stylization and abstraction of photographs. *ACM Transactions on Graphics*, *21*(3), 769–776.

Desingh, K. Krishna, K. M., Rajan, D., & Jawahar, C.(2013). Depth really matters: Improving visual salient region detection with depth. In *British machine vision conference*.

Donoser, M., Urschler, M., Hirzer, M., & Bischof, H. (2009). Saliency driven total variation segmentation. In *IEEE 12th international conference on computer vision* (pp. 817–824).

Drewes, H., Luca, A. D., & Schmidt, A. (2007). Eye-gaze interaction for mobile phones. In *Proceedings of international conference on mobile technology, applications, and systems* (pp. 364–371).

Ehinger, K., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modeling search for people in 900 scenes. *Visual Cognition*, *17*, 945–978.

El-Nasr, M. S., Vasilakos, A., Rao, C., & Zupko, J. (2009). Dynamic intelligent lighting for directing visual attention in interactive 3-d scenes. *IEEE Transactions on Computational Intelligence and AI in Games*, *1*(2), 145–153.

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, *8*(3), 3–3.

Erdem, E., & Erdem, A. (2013). Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision*, *13*(4), 1–20.

Everingham, M., Gool, L. J. V., Williams, C. K. I., Winn, J. M., & Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, *88*(2), 303–338.

Feng, S., Xu, D., & Yang, X. (2010). Attention-driven salient edge (s) and region (s) extraction with application to cbir. *Signal Processing*, *90*(1), 1–15.

Frintrop, S. (2006). *VOCUS: A visual attention system for object detection and goal-directed search* (Vol. 3899).

Frintrop, S. (2011). Towards attentive robots. *Paladyn*, *2*(2), 64–70.

Frintrop, S., Garcia, G. M., & Cremers, A. B. (2014). A cognitive approach for object discovery. In *International conference on pattern recognition* (pp. 2329–2334).

Frintrop, S., & Jensfelt, P. (2008). Attentional landmarks and active gaze control for visual SLAM. *IEEE Transactions on Robotics*, *24*(5), 1054–1065.

Frintrop, S., & Kessel, M. (2009). Most salient region tracking. In *IEEE international conference on robotics and automation* (pp. 1869–1874).

Frintrop, S., Königs, A., Hoeller, F., & Schulz, D. (2010). A component-based approach to visual person tracking from a mobile platform. *International Journal of Social Robotics*, *2*(1), 53–62.

Fritz, G., Seifert, C., Paletta, L., & Bischof, H. (2004). Attentive object detection using an information theoretic saliency measure. In *International workshop on attention and performance in computational vision* (pp. 29–41).

Gadde, R. & Karlapalem, K. (2011). Aesthetic guideline driven photography by robots. In *International joint conference on artificial intelligence* (Vol. 22, pp. 2060).

Gao, D., Han, S., & Vasconcelos, N. (2009). Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(6), 989–1005.

Gao, D. & Vasconcelos, N. (2004). Discriminant saliency for visual recognition from cluttered scenes. In *Advances in neural information processing systems* (pp. 481–488).

Gao, Y., Shi, M., Tao, D., & Xu, C. (2015). Database saliency for fast image retrieval. *IEEE Transactions on Multimedia*, *17*(3), 359–369.

Gautier, J., Le Meur, O., & Guillemot, C. (2012). Efficient depth map compression based on lossless edge coding and diffusion. In *Picture coding symposium* (pp. 81–84).

Girshick, R. B. (2015). Fast R-CNN. In *IEEE international conference on computer vision, ICCV* (pp. 1440–1448).

Goferman, S., Tal, A., & Zelnik-Manor, L. (2010). Puzzle-like collage. *Computer Graphics Forum*, *29*, 459–468.

Goferman, S., Zelnik-Manor, L., & Tal, A. (2010). Context-aware saliency detection. In *IEEE conference on computer vision and pattern recognition* (pp. 2376–2383).

Goldberg, C., Chen, T., Zhang, F., Shamir, A., & Hu, S. (2012). Data-driven object manipulation in images. *Computer Graphics Forum*, *31*, 265–274.

Graves, A. (2013). Generating sequences with recurrent neural networks. *CoRR* (abs/1308.0850).

Gupta, R., Khanna, M. T., & Chaudhury, S. (2013). Visual saliency guided video compression algorithm. *Signal Processing: Image Communication*, *28*(9), 1006–1022.

Han, S., & Vasconcelos, N. (2010). Biologically plausible saliency mechanisms improve feedforward object recognition. *Vision Research*, *50*(22), 2295–2307.

Haque, A., Alahi, A., & Fei-Fei, L.(2016). Recurrent attention models for depth-based person identification. In *IEEE conference on computer vision and pattern recognition* (pp. 1229–1238).

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In *Advances in neural information processing systems* (pp. 545–552).

Heidemann, G., Rae, R., Bekel, H., Bax, I., & Ritter, H. (2004). Integrating context-free and context-dependent attentional mechanisms for gestural object reference. *Machine Vision and Applications*, *16*(1), 64–73.

Hong, B., & Brady, M. (2003). A topographic representation for mammogram segmentation. In *Medical image computing and computer-assisted intervention* (pp. 730–737).

Hong, R., Wang, M., Xu, M., Yan, S., & Chua, T. (2010). Dynamic captioning: Video accessibility enhancement for hearing impairment. In *ACM multimedia*.

Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In *IEEE conference on computer vision and pattern recognition*.

Hou, X., & Zhang, L. (2008). Dynamic visual attention: Searching for coding length increments. *Advances in Neural Information Processing Systems*, *21*, 681–688.

Huang, H., Zhang, L., & Zhang, H.-C. (2011). Arcimboldo-like collage using internet images. *ACM Transactions on Graphics*, *30*, 1–7.

iLab, C., (2010). *Neuromorphic vision*. Toolkit.

Ishiguro, Y., Mujibiya, A., Miyaki, T., & Rekimoto, J. (2010). Aided eyes: Eye activity sensing for daily life. In *Proceedings of augmented human international conference* (p. 25).

Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, *13*(10), 1304–1318.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

Jacobson, N., Lee, Y., Mahadevan, V., Vasconcelos, N., & Nguyen, T. Q. (2010). A novel approach to fruc using discriminant saliency and frame segmentation. *IEEE Transactions on Image Processing*, *19*(11), 2924–2934.

Ji, Q., Fang, Z., Xie, Z., & Lu, Z. (2013). Video abstraction based on the visual attention model and online clustering. *Signal Processing: Image Communication*, *28*(3), 241–253.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *ACM multimedia* (pp. 675–678).

Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., & Li, S. (2013). Salient object detection: A discriminative regional feature integration approach. In *IEEE conference on computer vision and pattern recognition* (pp. 2083–2090).

Jiang, M., Huang, S., Duan, J., & Zhao, Q. (2015a). Mouse saliency-a new method for low-cost large-scale attentional data collection. *Journal of Vision*, *15*(12), 221–221.

Jiang, M., Huang, S., Duan, J., & Zhao, Q. (2015b). SALICON: Saliency in context. In *IEEE conference on computer vision and pattern recognition*.

Johnson-Roberson, M., Bohg, J., Björkman, M., & Kragic, D. (2010). Attention-based active 3d point cloud segmentation. In *International conference on intelligent robots and systems* (pp. 1165–1170).

Kadir, T., & Brady, M. (2001). Saliency, scale and image description. *International Journal of Computer Vision*, *45*(2), 83–105.

Kanan, C., & Cottrell, G. W. (2010). Robust classification of objects, faces, and flowers using natural image statistics. In *IEEE conference on computer vision and pattern recognition* (pp. 2472–2479).

Karpathy, A., Miller, S., & Fei-Fei, L. (2013). Object discovery in 3d scenes via shape analysis. In *IEEE international conference on robotics and automation* (pp. 2088–2095).

Kim, J., Han, D., Tai, Y., & Kim, J. (2014). Salient region detection via high-dimensional color transform. In *IEEE conference on computer vision and pattern recognition* (pp. 883–890).

Kläser, A., Marszalek, M., & Schmid, C. (2008). A spatio-temporal descriptor based on 3D-gradients. In *British machine vision conference*.

Klein, D. A., Schulz, D., Frintrop, S., & Cremers, A. B. (2010). Adaptive real-time video-tracking for arbitrary objects. In *International conference on intelligent robots and systems* (pp. 772–777).

Koch, C., & Ullman,S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, *4*, 219–227.

Kolmogorov, V., & Zabih,R. (2004). What energy functions can be minimized via graph cuts? volume 26, pages 147–159.

Krähenbühl, P., & Koltun,V. (2014). Geodesic object proposals. In *European conference on computer vision* (pp. 725–739).

Lance, B., & Marsella, S. (2010). The expressive gaze model: Using gaze to express emotion. *IEEE Computer Graphics and Applications*, *30*(4), 62–73.

Lance, B., Marsella, S., & Koizumi, D. (2004). Towards expressive gaze manner in embodied virtual agents. In *AAMAS workshop on empathic agents* New-York.

Lang, C., Nguyen, T., Katti, H., Yadati, K., Kankanhalli,M.S., & Yan,S. (2012). Depth matters: Influence of depth cues on visual saliency. In *European conference on computer vision* (pp. 101–115).

Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld,B. (2008). Learning realistic human actions from movies. In *IEEE conference on computer vision and pattern recognition*.

Lazebnik, S., Schmid, C., & Ponce,J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer vision and pattern recognition* (pp. 2169–2178).

Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(5), 802–817.

Lee, C. H., Varshney, A., & Jacobs, D. W. (2005). Mesh saliency. *ACM Transactions on Graphics*, *24*, 659–666.

Lee, Y. J. , Ghosh, J., & Grauman, K. (2012). Discovering important people and objects for egocentric video summarization. In *IEEE conference on computer vision and pattern recognition* (Vol. 2, p. 6).

Li, A., She, X., & Sun,Q. (2013). Color image quality assessment combining saliency and fsim. In *International conference on digital image processing*.

Li, H., & Ngan, K. N. (2008). Saliency model-based face segmentation and tracking in head-and-shoulder video sequences. *Journal of Visual Communication and Image Representation*, *19*(5), 320–333.

Li, L., Jiang, S., Zha, Z.-J., Wu, Z., & Huang, Q. (2013). Partial-duplicate image retrieval via saliency-guided visual matching. *IEEE MultiMedia*, *20*(3), 13–23.

Li, L., Mei, T., & Hua, X.-S. (2010a). Gamesense: Game-like in-image advertising. *Multimedia Tools and Applications*, *49*(1), 145–166.

Li, L., Mei, T., Hua, X.-S., & Li, S. (2008). Imagesense. In *ACM multimedia* (pp. 1027–1028).

Li, L., Mei, T., Niu, X., & Ngo, C.-W. (2010b). Pagesense: Style-wise web page advertising. In *International conference on world wide web* (pp. 1273–1276).

Li, Q., Zhou, Y., & Yang, J. (2011). Saliency based image segmentation. In *International conference on multimedia technology* (pp. 5068–5071).

Li, Y., Hou, X., Koch, C., Rehg, J. M., & Yuille, A. L. (2014). The secrets of salient object segmentation. In *IEEE conference on computer vision and pattern recognition* (pp. 280–287).

Liu, H., & Heynderickx, I. (2009). Studying the added value of visual attention in objective image quality metrics based on eye movement data. In *IEEE international conference on image processing* (pp. 3097–3100).

Liu, H., Jiang, S., Huang, Q., & Xu, C. (2008). A generic virtual content insertion system based on visual attention analysis. In *ACM multimedia* (pp. 379–388).

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., et al. (2011). Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(2), 353–367.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *IEEE international conference on computer vision* (pp. 1150–1157).

Luebke, D. (2016). Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics*, *35*(6), 179:1–179:12.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025.

Ma, Y.-F., Hua, X.-S., Lu, L., & Zhang, H.-J. (2005). A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia*, *7*(5), 907–919.

Mahadevan, V., & Vasconcelos, N. (2009). Saliency-based discriminant tracking. In *IEEE conference on computer vision and pattern recognition* (pp. 1007–1013).

Maki, A., Nordlund, P., & Eklundh, J. (2000). Attentional scene segmentation: Integrating depth and motion. *Computer Vision and Image Understanding*, *78*(3), 351–373.

Marchesotti, L., Cifarelli, C., & Csurka, G. (2009). A framework for visual saliency detection with applications to image thumbnailing. In *IEEE conference on computer vision* (pp. 2232–2239).

Margolin, R., Zelnik, L., & Tal, A. (2013). Saliency for image manipulation. *The Visual Computer*, *29*(5), 381–392.

Martín-Martín, A., Ayllón, J. M., Orduña-Malea, E., & López-Cózar, E. D. (2014). Google scholar metrics 2014: A low cost bibliometric tool. arXiv preprint arXiv:1407.2827.

Mateescu, V. A., & Bajić, I. V. (2014). Attention retargeting by color manipulation in images. In *International workshop on perception inspired video processing* (pp. 15–20).

Mathe, S., & Sminchisescu, C., (2012). Dynamic eye movement datasets and learnt saliency models for visual action recognition. In *European conference computer vision* (pp. 842–856).

Mathe, S., & Sminchisescu, C., (2013). Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *CoRR* (abs/1312.7570).

Meger, D., Forssén, P.-E., Lai, K., Helmer, S., McCann, S., Southey, T., et al. (2008). Curious george: An attentive semantic robot. *Robotics and Autonomous Systems*, *56*(6), 503–511.

Mei, T., Li, L., Hua, X.-S., & Li, S. (2012). Imagesense: Towards contextual image advertising. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *8*(1), 6.

Mertsching, B., Bollmann, M., Massad, A., & Schmalz, S., (1998). Recognition of complex objects with an active vision system. In *Symposium on neural computation* (pp. 469–475).

Mishra, A. K., Aloimonos, Y., Cheong, L. F., & Kassim, A. A. (2012). Active visual segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(4), 639–653.

Mitri, S., Frintrop, S., Pervölz, K., Surmann, H., & Nüchter, A., (2005). Robust object detection at regions of interest with an application in ball recognition. In *IEEE international conference on robotics and automation* (pp. 125–130).

Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K., (2014). Recurrent models of visual attention. In *Advances in neural information processing systems* (pp. 2204–2212).

Moosmann, F., Larlus, D., & Jurie, F., (2006). Learning saliency maps for object categorization. In *ECCV workshop on the representation and use of prior knowledge in vision*.

Muhl, C., Nagai, Y., & Sagerer, G., (2007). On constructing a communicative space in hri. In *KI 2007: Advances in artificial intelligence* (pp. 264–278).

Muratov, O., Dang, T., Boato, G., & De Natale, F., (2012). Saliency detection as a support for image forensics. In *International symposium on communications control and signal processing* (pp. 1–5).

Murray, N., Vanrell, M., Otazu, X., & Párraga, C. A. (2011). Saliency estimation using a non-parametric low-level vision model. In *IEEE conference on computer vision and pattern recognition* (pp. 433–440).

Nagai, Y. (2009). From bottom-up visual attention to robot action learning. In *International conference on development and learning* (pp. 1–6).

Navalpakkam, V., & Itti, L. (2006). An integrated model of top-down and bottom-up attention for optimizing detection speed. In *2006 IEEE conference on computer vision and pattern recognition* (pp. 2049–2056).

Nguyen, P., Fleureau, J., Chamaret, C., & Guillotel, P. (2013). Calibration-free gaze tracking using particle filter. In *IEEE international conference on multimedia and expo* (pp. 1–6).

Nguyen, T., Li, L., Tan, J., & Yan, S. (2012). 3DME: 3d media express from rgb-d images. In *ACM multimedia* (pp. 1331–1332).

Nguyen, T., Ni, B., Liu, H., Xia, W., Luo, J., Kankanhalli, M., et al. (2013). Image re-attentionizing. *IEEE Transactions on Multimedia*, *15*(8), 1910–1919.

Nguyen, T., Song, Z., & Yan, S. (2015). STAP: spatial-temporal attention-aware pooling for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *25*(1), 77–86.

Nguyen, T., Xu, M., Gao, G., Kankanhalli, M. S., Tian, Q., & Yan, S. (2013). Static saliency vs. dynamic saliency: A comparative study. In *ACM multimedia* (pp. 987–996).

Nguyen, T. V. (2015). Salient object detection via objectness proposals. In *Proceedings of AAAI conference on artificial intelligence* (pp. 4286–4287).

Nguyen, T. V., & Liu, L. (2017). Salient object detection with semantic priors. In *International joint conference on artificial intelligence*.

Nguyen, T. V., & Sepulveda, J. (2015). Salient object detection via augmented hypotheses. In *International joint conference on artificial intelligence* (pp. 2176–2182).

Ni, B., Xu, M., Nguyen, T., Wang, M., Lang, C., Huang, Z., et al. (2014). Touch saliency: Characteristics and prediction. *IEEE Transactions on Multimedia*, *16*(6), 1779–1791.

Ninassi, A., Le Meur, O., Le Callet, P., & Barbba, D. (2007). Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric. In *IEEE international conference on image processing* (Vol. 2, pp. 2–169).

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

Ouerhani, N., Bracamonte, J., Hugli, H., Ansorge, M., & Pellandini, F. (2001). Adaptive color image compression based on visual attention. In *International conference on image analysis and processing* (pp. 416–421).

Ouerhani, N., Bur, A., & Hügli, H. (2005). Visual attention-based robot self-localization. In *European conference on mobile robots* (pp. 8–13).

Papadopoulos, D. P., Clarke, A. D., Keller, F., & Ferrari, V. (2014). Training object class detectors from eye tracking data. In *European conference on computer vision* (pp. 361–376).

Parikh, N., Itti, L., & Weiland, J. (2010). Saliency-based image processing for retinal prostheses. *Journal of Neural Engineering*, *7*(1), 1–10.

Perazzi, F., Krähenbühl, P., Pritch, Y., & Hornung, A. (2012). Saliency filters: Contrast based filtering for salient region detection. In *IEEE conference on computer vision and pattern recognition* (pp. 733–740).

Perra, D., Gupta, R. K., & Frahm, J. (2015). Adaptive eye-camera calibration for head-worn devices. In *IEEE conference on computer vision and pattern recognition* (pp. 4146–4155).

Qin, C., Zhang, G., Zhou, Y., Tao, W., & Cao, Z. (2014). Integration of the saliency-based seed extraction and random walks for image segmentation. *Neurocomputing*, *129*, 378–391.

Queiroz, R. B., Barros, L. M., & Musse, S. R. (2007). Automatic generation of expressive gaze in virtual animated characters: From artists craft to a behavioral animation model. In *Intelligent virtual agents, 7th international conference* (pp. 401–402).

Ren, S., He, K., Girshick, R. B., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149.

Ren, Z., Gao, S., Chia, L., & Tsang, I. W. (2014). Region-based saliency detection and its application in object recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *24*(5), 769–779.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*, 1019–1025.

Roberts, R., Ta, D.-N., Straub, J., Ok, K., & Dellaert, F. (2012). Saliency detection and model-based tracking: A two part vision system for small robot navigation in forested environment. In *SPIE defense, security, and sensing* (pp. 83870S–83870S).

Rosenholtz, R., Dorai, A., & Freeman, R. (2011). Do predictions of visual perception aid design? *ACM Transactions on Applied Perception*, *8*(2), 12.

Rother, C., Kolmogorov, V., & Blake, A. (2004). Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, *23*(3), 309–314.

Rutishauser, U., Walther, D., Koch, C., & Perona, P. (2004). Is bottom-up attention useful for object recognition? In *IEEE conference on computer vision and pattern recognition* (pp. 37–44).

Sadaka, N., & Karam, L. (2009). Efficient perceptual attentive super-resolution. In *IEEE international conference on image processing* (pp. 3113–3116).

Salah, A., Alpaydin, E., & Akarun, L. (2002). A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(3), 420–425.

Scheier, C., & Egner, S. (1997). Visual attention in a mobile robot. In *IEEE international symposium on industrial electronics* (Vol. 1, pp. 48–52).

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1.

Setlur, V., Takagi, S., Raskar, R., Gleicher, M., & Gooch, B. (2005). Automatic image retargeting. In *International conference on mobile and ubiquitous multimedia* (pp. 59–68).

Shen, C., & Zhao, Q. (2014). Webpage saliency. In *European conference on computer vision* (pp. 33–46).

Shen, H., Li, S., Zhu, C., Chang, H., & Zhang, J. (2013). Moving object detection in aerial video based on spatiotemporal saliency. *Chinese Journal of Aeronautics*, *26*(5), 1211–1217.

Shiffrin, R., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127.

Siagian, C., & Itti, L. (2007). Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(2), 300–312.

Siagian, C., & Itti, L. (2009). Biologically inspired mobile robot vision localization. *IEEE Transactions on Robotics*, *25*(4), 861–873.

Simoncelli, E. (1996). Foundations of vision .

Srivatsa, R. S., & Babu, R. V. (2015). Salient object detection via objectness measure. In *International conference on image processing* (pp. 4481–4485).

Stalder, S., Grabner, H., & Gool, L. J. V. (2012). Dynamic objectness for adaptive tracking. In *Asian conference on computer vision* (pp. 43–56).

Stentiford, F. (2003). Attention-based image similarity measure with application to content-based information retrieval. In *Electronic imaging* (pp. 221–232).

Sugano, Y., Matsushita, Y., & Sato, Y. (2010). Calibration-free gaze sensing using saliency maps. In *IEEE conference on computer vision and pattern recognition* (pp. 2667–2674).

Suh, B., Ling, H., Bederson, B. B., & Jacobs, D. W. (2003). Automatic thumbnail cropping and its effectiveness. In *ACM symposium on user interface software and technology* (pp. 95–104).

Tanaka, R., Narumi, T., Tanikawa, T., & Hirose, M. (2015). Attracting user's attention in spherical image by angular shift of virtual camera direction. In *ACM symposium on spatial user interaction* (pp. 61–64).

Tatler, B., Hayhoe, M., Land, M., & Ballard, D. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5), 5.

Vig, E., Dorr, M., & Cox, D. D. (2012). Space-variant descriptor sampling for action recognition based on saliency and eye movements. In *European conference on computer vision* (pp. 84–97).

Vijayakumar, S., Conradt, J., Shibata, T., & Schaal, S. (2001). Overt visual attention for a humanoid robot. In *Proceedings 2001 IEEE/RSJ international conference on intelligent robots and systems* (Vol. 4, pp. 2332–2337).

Viola, P., & Jones, M. (2004). Robust real-time face detection. *International Journal of Computer Vision*, *57*(2), 137–154.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*(9), 1395–1407.

Wang, H., Kläser, A., Schmid, C., & Liu, C. (2011). Action recognition by dense trajectories. In *IEEE conference on computer vision and pattern recognition* (pp. 3169–3176).

Wang, H., & Schmid, C. (2013). Action Recognition with Improved Trajectories. In *IEEE international conference on computer vision* (pp. 3551–3558).

Wang, H., Ullah, M., Kläser, A., Laptev, I., & Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. In *British machine vision conference*.

Wang, J., Quan, L., Sun, J., Tang, X., & Shum, H.-Y. (2006). Picture collage. In *IEEE conference on computer vision and pattern recognition* (Vol. 1, pp. 347–354).

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238.

Wong, L., & Low, K. (2011). Saliency retargeting: An approach to enhance image aesthetics. In *IEEE workshop on applications of computer vision* (pp. 73–80).

Wong, L.-K., & Low, K.-L. (2009). Saliency-enhanced image aesthetics class prediction. In *IEEE international conference on image processing* (pp. 997–1000).

Wong, L.-K., & Wong, K.-L. (2012). Enhancing visual dominance by semantics-preserving image recomposition. In *ACM multimedia* (pp. 845–848).

Xu, J., Mukherjee, L., Li, Y., Warner, J., Rehg, J. M., & Singh, V. (June 2015). Gaze-enabled egocentric video summarization via constrained submodular maximization. In *IEEE conference on computer vision and pattern recognition*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning* (pp. 2048–2057).

Xu, K., Chen, K., Fu, H., Sun, W.-L., & Hu, S.-M. (2013). Sketch2scene: Sketch-based co-retrieval and co-placement of 3d models. *ACM Transactions on Graphics*, *32*(4), 1–12.

Yan, Q., Xu, L., Shi, J., & Jia, J. (2013). Hierarchical saliency detection. In *IEEE conference on computer vision and pattern recognition* (pp. 1155–1162).

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. J. (2016). Stacked attention networks for image question answering. In *IEEE conference on computer vision and pattern recognition* (pp. 21–29).

Yun, K., Peng, Y., Samaras, D., Zelinsky, G. J., & Berg, T. L. (2013). Studying relationships between human gaze, description, and computer vision. In *IEEE conference on computer vision and pattern recognition* (pp. 739–746).

Zhai, Y., & Shah, M. (2006). Visual attention detection in video sequences using spatiotemporal cues. In *ACM international conference on multimedia* (pp. 815–824).

Zhang, G., Yuan, Z., Zheng, N., Sheng, X., & Liu, T. (2009). Visual saliency based object tracking. In *Asian conference on computer vision* (pp. 193–203).

Zhang, G.-X., Cheng, M.-M., Hu, S.-M., & Martin, R. R. (2009). A shape-preserving approach to image resizing. *Computer Graphics Forum*, *28*, 1897–1906.

Zhang, J., & Sclaroff, S. (2013). Saliency detection: A boolean map approach. In *IEEE international conference on computer vision* (pp. 153–160).

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, *8*(7), 32.

Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2015). Appearance-based gaze estimation in the wild. In *IEEE conference on computer vision and pattern recognition* (pp. 4511–4520).

Zhao, R., Ouyang, W., & Wang, X. (2013a). Person re-identification by salience matching. In *IEEE international conference on computer vision* (pp. 2528–2535).

Zhao, R., Ouyang, W., & Wang, X. (2013b). Unsupervised salience learning for person re-identification. In *IEEE conference on computer vision and pattern recognition* (pp. 3586–3593).

Zhao, R., Ouyang, W., & Wang, X. (2015). Person re-identification by saliency learning. *CoRR* (abs/1412.1908).

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2014). Object detectors emerge in deep scene cnns. In *International conference on learning representations*.

Zhou, B., Lapedriza, À., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Advances in neural information processing systems* (pp. 487–495).

Zitnick, C. L., & Dollár, P. (2014). Edge boxes: Locating object proposals from edges. In *European conference on computer vision* (pp. 391–405).