PhD Dissertation

International Doctorate School in Information and Communication Technologies

DISI - University of Trento

Visual Saliency Detection and Its Application to Image Retrieval

Oleg Muratov

Advisor:

Assist prof. Giulia Boato

Università degli Studi di Trento

Co-advisor:

Prof. Francesco G. De Natale

Università degli Studi di Trento

March 2013

Abstract

People perceive any kind of information with different level of attention and involvement. It is due to the way how our brain functions, redundancy and importance of the perceived data. This work deals with visual information, in particular with images. Image analysis and processing is often requires running computationally expensive algorithms. The knowledge of which part of an image is important over other parts allows for reduction of data to be processed. Besides computational cost a broad variety of applications, including image compression, quality assessment, adaptive content display and rendering, can benefit from this kind of information. The development of an accurate visual importance estimation method may bring a useful tool for image processing domain and that is the main goal for this work. In the following two novel approaches to saliency detection are presented. In comparison to previous works in this field the proposed approaches tackles saliency estimation on the object-wise level. In addition, one of the proposed approach solves saliency detection problem through modelling 3-D spatial relationships between objects in a scene. Moreover, a novel idea of the application of saliency to diversification of image retrieval results is presented.

Keywords

[visual saliency, attention, diversity, image retrieval]

Contents

List of Tables

List of Figures

Chapter 1

Introduction

While analysing the content of images and videos one can notice that importance of the content within the frame is not equal. We are living in the age of information revolution. Information is everywhere in our lives. Television, radio, newspapers, social networks. This work deals with very particular part of information - images, mainly photographs. Images may represent different aspects of our live: everyday routine, events, holiday trips and arts. Like in any means of information exchange only some parts of an image contain the desired information. Indeed, due to the way images are created there is no way of full control over their content. Consider for example you want to make a picture of some monument in a city. This monument may be surround by some building, in between the monument's face surface and camera plane there could be people passing over, depending on the orientation of the camera with respect to the monument sky or ground plane may also appear in the frame. Although it is possible to avoid these objects by zooming very close to the monument, this is rarely done, because then the visualization of the monument will be less informative. Thereby inclusion of two types of content (foreground and background) in images is inevitable. For instance, in the aforesaid toy example the monument is the foreground and the rest of the frame can be considered as background. It is natural to think that the foreground part of an image is to be perceived as an important peace of information.

The same problem can be viewed from observer's point of view. Starting from our eyes the visual information is perceived unequally. The resolution of retina depend on location of objects in the visual field. Rays falling onto the center of the retina are perceived with higher resolution. Receptive fields in the cortex span for over 30 degrees, however their placement is not uniform at the mass of the fields are located at the center of the gaze. Each higher perception level of neural network performs a more complex information and capacity of each level decreases while ascending from bottom to the top. This peculiarities leads to the competition of information streams. Thus the way how our brain is functioning orders incoming visual information by its importance. The saccade search and selectivity process are guided by bottom-up and top-down stimulus. Top-down stimulus usually represent selection based on knowledge, for example, a subject is looking for a picture of an animal. Bottom-up stimulus are driven by properties of perceived visual information, such as high contrast, difference in orientation, etc.

Ability of automatic detection of important regions can be a priceless tool for a broad variety of multimedia applications. A wide spectrum of application may benefit from separative processing of important and less important regions of an image. Thus development of a robust method for automatic detection of important regions in image may lead to significant progress in multimedia processing. As it will be shown later there already exist a number of approaches to automatically determine important regions, or as it is often called saliency. However, as it is will be shown in Chapter 2 there is still room for improvements in this direction and in this work two novel approaches to saliency detection are presented.

Although it has passed over 15 years since first works on saliency detection were presented, still a few number of works were dedicated to the application of saliency information for improving multimedia processing algorithms and addressing new challenging problems. Current applications include gentle advertising in images [37], where a saliency map is used to insert advertisement above unimportant region of an image, thus preventing advertisement block from occlusion of foreground object. An approach to use of saliency detection in image retrieval was presented in [23]. The authors proposed to use saliency value as a weighting parameter for SIFT key-points. Thus similarity of two images is defined by number of matched SIFT key-points only from their salient areas. A similar approach was presented in [39]. Though instead of using key-points for similarity measure, features employed in saliency detection are used. A quite similar idea can be found in [61], where saliency and color features are used for measuring similarity of images. Another way of using saliency for image retrieval was presented in [19]. Here, the authors apply saliency to interactive retrieval. At each iteration feedback from a user is used to construct affinity matrix that is formed by salient regions of positive images that is further used as query. Alternative way for the same problem was proposed in [62]. The authors proposed using saliency information to drive a semantics model of an image that is further used for retrieval.

Another common application of saliency is adaptive content display. For instance, in [40] a saliency based image re-targeting approach was presented. The authors proposed using unimportant regions of images as areas for continuous seam carving. Thereby, scale compression and corresponding spatial distortion does not affect important regions of an image and thus perceptional content of an image is not destroyed. In [42] the authors proposed using saliency maps for thumbnail generation. A thumbnail is generated in a way such that a corresponding bounding box is generated around salient areas. A similar approach was presented in [32].

Saliency detection has also found its application in object recognition. For instance, in [22] saliency is employed for image classification. Here, the authors targeted categorization of objects in images to classes like bicycles, bus, cars, etc. Saliency is used to emphasize features in important areas. As a features several key-points extractors were used. In [48] a survey on using saliency for object categorization was presented. Here, the authors proposed to use saliency as one of the feature together with key-points both during training and classification steps.

Image compression can also benefit from using saliency detection. For example, in [16] the authors proposed to use saliency as a parameter defining compression ration of different parts of a frame for MPEG compression. A similar idea was proposed in [28], where saliency define compression ratio for JPEG2000 compression. Another interesting application of saliency is rendering. For instance, in [13, 34] the detail level of rendering is guided by saliency value. Another work applying saliency to arts can be found in [43]. Here, the authors showed how automated photo-manipulation techniques can benefit from saliency information. The manipulation techniques include rendering, cropping and mosaic. The common approach is that saliency information guides the level of details for each manipulation technique. For instance, in mosaic application the accuracy of blob color representation for salient regions is kept higher with respect to other regions by means of using tiles of smaller size. Finally application of visual saliency to image forensics can be found in [45]. Here, the main idea is that manipulation is detected by computing JPEG-ghost effect and comparing its value in salient and non-salient areas. The assumption is that if manipulation is done it most probably effects the main object of an image. Thus the difference in output of JPEG-ghost for salient and non-salient areas indicates that the presence of manipulation.

For the purpose of expanding this list and contributing to saliency detection in Chapter 3 diversity of image retrieval results is addressed through the use of saliency. A single word may have several semantic meaning, e.g. the word jaguar most often refers to a mammal and car manufacturer. In the same way a single concept may have different representation, e.g. a car can have different body color, body type, etc. The inclusion of these variations by retrieval is what is here understood as diversity. Another contribution of this work is two novel saliency detection methods. While most of the works propose detection of salient regions in pixel-wise domain, in this work the detection is done in object-wise domain. The intuition is that separate pixels, even compared globally, cannot represent object properties. Classification of data in object-wise domain allows capturing and comparison of objects properties and thus detection of higher-level information. Moreover, the method described in Section 2.5 defines saliency by modelling probability of an object through its spatial relationship in 3D space with respect to other objects.

The rest of the work is structured as follows: Chapter 2 is devoted to the problem of saliency detection, the state of the art in saliency detection is described in Section 2.2, then Sections 2.4 and 2.5 present segment-based and depth-based approaches respectfully, then in Section 2.7 the evaluation of the two proposed approach on the database described in Section 2.6 is done. Chapter 3 is devoted to the problem of using saliency detection in diversification of content-based image retrieval. The state of the art on diversification is given in Section 3.2. The description of the proposed approach is presented in Section 3.3. Its evaluation is then done in Section 3.4. Finally, the conclusion of this work is done in Chapter 4.

Chapter 2

Visual saliency detection

2.1 Introduction

There exist different approaches to saliency detection. Before describing state-of-art methods it is good to introduce main approaches to saliency detection. Figure 2.1 presents ontology of possible saliency detection schemes. Saliency detection methods can be grouped according to the model inspiration source. For instance, Itti's approach is referred as a biological inspired method. Such methods explore percularities of human vision and attention operation and try to mimic the processes taking place while a human observes a scene. Another group of methods explore natural statistics found in images. For instance, in Hou et al proposed spectral residual approach [31] that exploits spectral histogram singularities to detect salient regions in images. Another common approach of saliency is computational. These methods exploit information domain properties to detect salient regions. Another grouping can be made considering what type of task the authors addressed in their works. Here, two groups are possible: human fixations and region of interest. Although these two tasks may sound similarly, there is a noticeable difference in output maps. Human vision system works with very sparse data and fixations are usually found only on a small portion of object's area. Thus a method trying to predict human fixations would highlight edges and contrast spots of an object. Methods aimed at detection of salient regions should provide a map highlighting the whole object of interest. Often, in region-based methods human fixations map is an interim product that is further developed into region-based map by means of region growing or segmentation. Output map can also differ in their representation. Here two options are possible: binary and grayscale. The former draws salient pixels/regions in white and non-salient in black. The later usually represents probability of a region to be salient by different tones of gray. For region of interest detection methods output maps can also differ in how the salient region is highlighted. Some authors prefer to use rectangular windows, others use segmentation mask and paint different segments according to their value of saliency and lastly region masks can be represented by raw pixels.

Another important grouping considers what features are used to detect saliency. Bottom-up approaches use low-level features such as color contrast, orientation and luminosity to detect saliency. Top-down approaches instead use high-level features such as faces and objects. Fusion of bottom-up and top-down features is also exploited by some works. Most of methods extract more than one feature from an image. There are different approaches on how these features are fused. The most simple approach is when values of feature for a pixel or a region are combined linearly with some weighting. Another possible option is to take value of a feature with the maximum output. Some methods use fusion technique known as ”winner take all”, where a feature with maximum value farthest from the mean value is selected. Other options include probabilistic inference, support vector machine (SVM) classification, etc. In addition, models can be grouped according to how their parameters are selected. Most models require supervised parameter tuning either by means of learning on training database or by setting parameters manually. In rare cases, there is no need for parameter tuning.

In next section recent advances in saliency detection are presented. Then Section 2.3 highlights main contributions proposed segments-based and depth-based approaches presented in Section 2.4 and Section 2.5 respectfully. Section 2.6 describes a dataset created for the training and evaluation. Finally, in Section 2.7 the evaluation of these two methods and their comparison to state of the art methods is given.

2.2 State of the art on visual saliency detection

The saliency detection model proposed by Itti et al [33] resulted in an avalanche of different saliency detection algorithm. One of the most recent and well-excepted extensions of Itti's saliency detection approach was presented by Judd et al in [34]. Although Itti's approach was taken as a basis the authors have combined a much broader set of features. The authors proposed to use three levels of features: low level, mid-level and high-level. Low level is formed by intensity, orientation and color contrast features as they were defined in the Itti's work. In addition, the authors included distance to center, local energy pyramids, and probability of color channels computed from 3D color histograms with median filter at 6 scales. The mid-level is formed by horizon line detector. Finally high level features are formed by Viola-Jones face and person detectors. The classification is done using SVM with linear kernel. Before features are extracted from an image it is resized to 200×200 px, thus original aspect ratio of the image is loosen. For the training and evaluation of their model the authors collected a database of 1003 images from photosharing services and collected human fixations from 15 users. It is worth mentioning the training settings the authors used. Instead of direct parsing training images data and ground truth labels to SVM, the authors performed selection of data

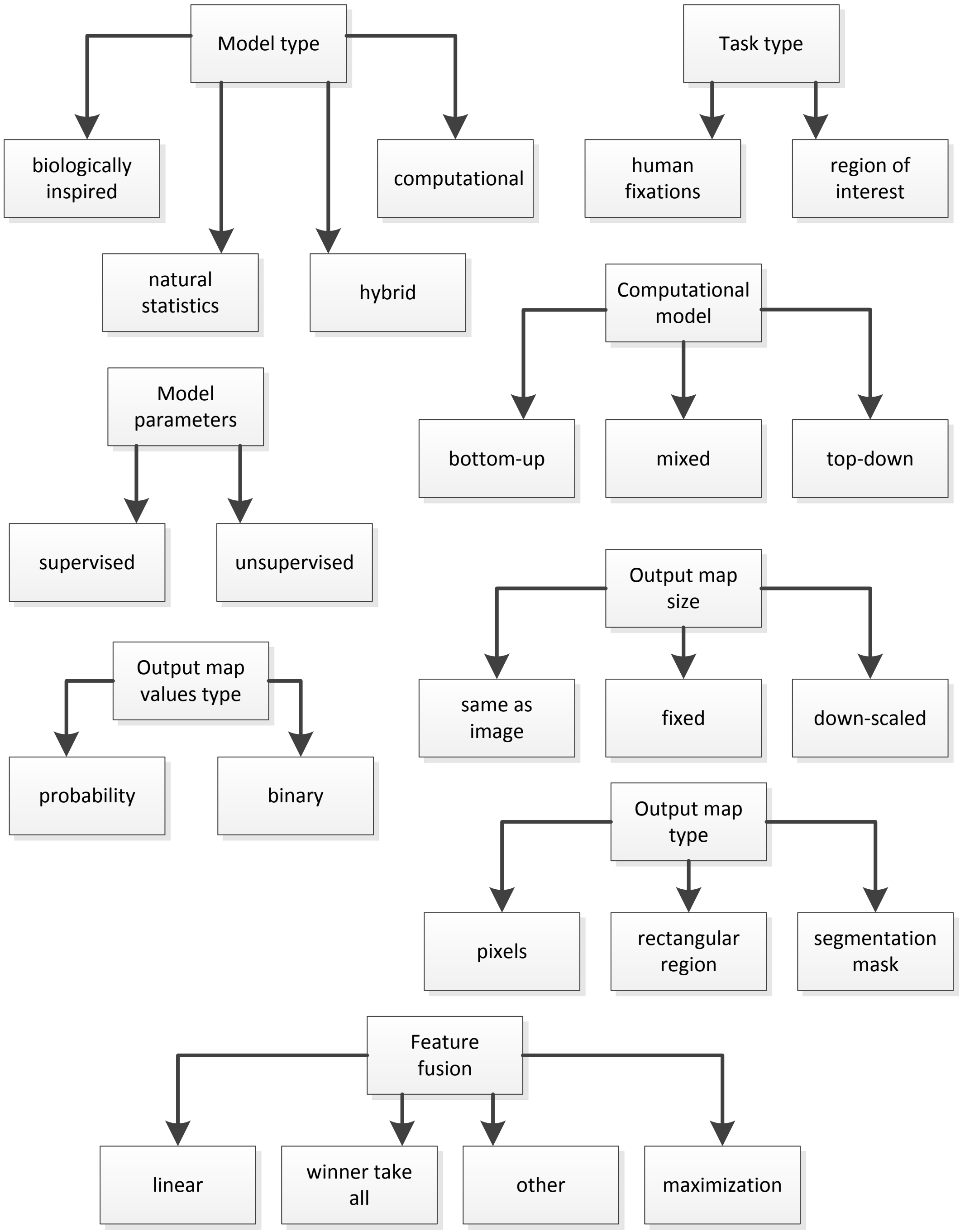


Figure 2.1: Saliency detection methods ontology

for training. Due to relatively high number of observers and time variation of fixation points it was possible to rank fixation locations by time subjects gazed at it. From each test image 10 pixels ranked from the top 20% salient locations were picked as positive examples and 10 non-salient pixels from bottom 70% salient regions as negative examples. In their evaluation the authors did provide how each feature contributes to the final result. It is of interest to point-out that use of distance to center feature allows higher performance as using all features except it. It is not clear why this feature alone resulted in such high performance. Possible answer could be high bias in the test images or weak filtering of eye-tracker data. Another interesting observation is that object detectors perform just a bit higher than chance baseline, that can be explained by difference in region-based and fixation-based maps. That result leaves unclear why object detectors are used in an eye fixation prediction method. Nevertheless, their evaluation has shown a noticeable increase in accuracy compared to Itti's model, however, there is no comparison to any other saliency detection method.

Biological motivated feature integration theory proposed in [35] is exploited in many works. For instance, in [36] a coherent detection method was presented. The features employed are similar to many other biologically inspired methods and include intensity, colors, orientations and contrast. The color value of an input image are converted into Krauskopf's color space. Then decomposition of three image channels in spectrum domain is done, that corresponds to activity of visual cells during perception of signals with specific 2D frequency and orientation. Contrast sensitivity function is then used to perform filtering, with anisotropic filter applied to the chroma component of an image and two low pass filters performing sinusoidal color grating applied to color components of an image. This step is also motivated by how human eyes perceive light signals. The next step is adaptive thresholding of feature outputs that mimics masking effects within inter-feature and intra-feature spaces. This step is followed by color enhancement that is increase of saliency for an achromatic region surrounded by high-contrast area. Then difference of Gaussians is used to mimic center-surround property of visual cells. Next butterfly filter is employed to perform contour grouping with motivation that if structures within center and surround stimuli are iso-oriented and co-aligned then perceptional grouping mechanism is launched. Then final map is obtained by linear combination of feature outputs.

Another work attracted much attention in saliency detection community was presented by Liu et al [38]. Here the authors based their model on center-surround processing, like it is done in many other computational-based approaches. Saliency of a pixel is modelled via CRF modelling. The pairwise term represents penalty as: $|a_x - a_{x'}| \bullet exp(-\beta d_{x,x'})$, where the primer term denotes difference in saliency value of two pixels, β is normalization term and $d_{x,x'}$ denotes L2 norm of the color difference. The pairwise term allows for spatial smoothness of saliency labels. Singleton term includes a number of low-level features. Among them is multi-scale local contrast measured in 9×9 neighbourhood. Center-surround histogram is another feature employed for singleton saliency detection. This feature is based on observation that a histogram of a window drawn around an object has higher extend compared to that of a window of the same area drawn around that window. In addition, color spatial distribution is counted to consider global context of an image. For this purpose spatial distribution and variance of each color is modelled via Gaussian mixture models. For training and evaluation the authors collected two databased overall consisting of over 20000 images. The ground-truth data in this dataset is represented by bounding boxes. The final labelling is computed through MAP assignment of the CRF network. The output map is then obtained by thresholding saliency values.

An approach based on information maximization was proposed by Bruce et al [9]. The authors proposed saliency measure based on self-information of each local image patch. This is done by calculating Shannon's self-information measure applied to joint likelihood statistics over the image. For this purpose, a basis function matrix was learned by running independent component analysis on a large sample of patches drawn from natural images. Given an image distribution of basis coefficients is calculated. Probability of each image patch is then calculated as probability of basis coefficients over the image it described with. Image patches are drawn at a very pixel location, the final map is calculated by accumulating saliency value a pixel received. Output maps are to large extend region-based.

A relatively large number of works are devoted to modelling saliency through color distribution and color cues. For instances, in [57] the authors proposed to detect saliency by means of isocentric curvedness and color. For this purposes the input image is converted into isophotes coordinate space - such that a line connects points of equal intensity. The main idea is that the bend of the isophotes indicate where an object these isophotes belong to is located. Thus by clustering and accumulating the votes for coordinates of the center of the circle estimated from bends of isophotes it is possible to determine location of objects present on an image. In addition, a parameter related to the slope of gradients around edges that is called curvedness. Another feature included performs saliency estimation via color boosting that is a transformation of image derivatives into spheres. These features are computed at multiple scales and then combined linearly. For generation of region-based maps graph-cut segmentation is used.

Another work exploiting color distribution over an image for saliency detection was proposed by Gopalakrishnan et al in [26]. The authors proposed to search saliency through color compactness and distinctiveness in hue saturation domain. The assumption here is that if a color cluster is distinct with respect to the other colors in the image then it is probably a salient color. For this purpose colors a modelled via Gaussian mixture models. Through iterative expectation-maximization (EM) process colors are assigned to the corresponding Gaussian cluster. A cluster with higher mean and smaller variations is an indicator of saliency. Numerically it is done by multiplying isolation by compactness of a cluster. In addition, orientations are included into the model. Unlike many other works here, orientations are computed not through filters, but by analysis of complex Fourier transform coefficients. The output map is obtained through selection process. For this purpose for each map its saliency index is computed, that treats connectivity of saliency regions and spatial variance. Only the map with highest saliency index is used as a final map. Even though no segmentation is used due to the clustering maps a region-based.

Similar idea of color-based saliency was proposed by Cheng et al in their global contrast based salient region detection approach [10]. Specifically, the authors defined pixel's saliency as a sum of its color contrast to all other pixels. This contrast is measured as a color distance in L*a*b* colorspace. In addition to global contrast, the authors include region-based contrast. Given two regions their contrast is defined as a sum of product over probabilities of each of their colors times distance between colors. Then saliency map is obtained by linear combination with further thresholding of global and region contrast maps. This saliency map is further used for initialization of GrabCut segmentation. Segmentation is done in iterative way with results of each iteration used for reinitialization after dilation and erosion. Thus output maps are region-based.

Another work exploiting global contrast was proposed by Vikram et al [60]. They proposed a method based on random window sampling. Specifically, from an input image a number of random sub-windows is generated and saliency of each window is defined as the absolute difference of the pixel intensity value to the mean intensity value of this windows. Input image is converted into L*a*b* colorspace with each color channel processed separately. The final map in obtained by linear combination of saliency value of sub-windows with further median filtering and contrast enhancement via histogram equalization. Due to sub-windows the output map is rather region-based.

An interesting approach to saliency detection was proposed by Marchesotti et al [42]. They addressed the problem of saliency detection via visual vocabulary. From a database with manually selected salient and non-salient regions they created a visual vocabulary. Features used for construction of visual vocabulary include SIFT-like gradient orientation histograms and local RGB mean and variations. Each visual word is parametrised as a Gaussian mixture model. Each image is then described as a pair of Fisher vectors - one for salient and other for non-salient parts. Once a new image is given for the classification k most similar images are retrieved. Similarity of two images is computed as distance between their Fisher vectors. From retrieved images a model of salient and non-salient regions is constructed. The input image is divided into 50×50 px patches and for each patch its distance to salient and non-salient areas of retrieved images is computed via Fisher vectors. A patch is considered as salient if its distance to salient regions is lower that to non-salient regions. A gaussian filter is applied to get smoothed saliency map from saliency values of patches. This map is further used for initialization of Graph-cut segmentation algorithm. For the evaluation the authors used PASCAL database, however, they indicated their primary application as thumbnail generation.

Another alterative approach to saliency detection was presented by Avraham et al [5]. Their extended saliency approach detects saliency through modelling distribution of labels within an image. The authors based their model on several observations, namely, that the number of salient objects is small, visually similar objects attract same amount of attention and finally natural scenes consist of several clustered structural objects. Thus the problem of classification is set as finding optimal joint assignment of labels to objects considering their number and similarity. The assumptions are exploited through soft clustering of the possible joint assignment. For this purpose image feature space is clustered into 10 clusters using Guassians mixtures using EM algorithm. For each cluster initial probability of its saliency is set according to the assumption of small number of expected salient objects. For each candidate its visual similarity and label correspondence to the rest of the candidates is measured. The correspondence between two labels is defined as their covariance divided by the product of their variances. This data is further used to represent dependencies between candidate labels using a Bayesian network, which is then inferences to find N most likely joint assignments. Final labels are obtained by marginalization of each candidate. Although the method targets prediction of human fixation points due to clustering to some extend maps can be considered as region-based.

Harel et al presented Graph-based saliency detection algorithm [29]. The authors proposed to represent an image as a fully-connected graph. The weight of each node is defined as feature distance between two pixels. Hence, higher difference in pixels appearance results in larger weight value. Then the equilibrium distribution over this graph reflecting time spent by a random walker highlight nodes with high dissimilarities with respect to other nodes. This activation measure reflects saliency of an image. Features include orientation maps obtained using Gabor filter for four orientations and luminance contrast map in a local neighborhood of size 80×80 px measured at two spatial scales. These twelve maps are then downsampled to a 28×37 feature maps. The authors performed evaluation of their method on Doves database [58] that contains natural grayscale images.

Among statistical approaches there is a group of methods exploring properties of images in frequency domain for detection of salient regions. A good example of such an approach is Spectral residual approach presented by Hoe et al [31]. Here, the authors represent an image as a superposition of two parts $H(image) = H(innovation)+H(priorknowledge)$ and explain the task of saliency detection as extraction of the innovation part. To do that the authors explore log Fourier spectrum of an image. They show that natural images share the same spectrum behaviour and deviations from this spectrum may indicate presence of a distinctive content in a particular part of an image. Through convolution the averaged spectrum of an image is computed: $A(f) = h_n(f) * L(f)$, where $L(f)$ is a log spectrum and $h_n(f)$ is a normalization matrix. Then spectral residual is obtained by subtracting averaged spectrum from log spectrum. Inverse Fourier Transform applied to the spectral residual returns a map, that is further smoothed with a Gaussian filter. The final saliency map is obtained by thresholding. This threshold is defined as average intensity of the interim saliency map, hence this method is completely unsupervised. The spectrum is computed from a down-sampled image with heigh or width equals 64 px, the final map has similar dimensions.

An extension of spectral residual approach targeting salient object detection was proposed by Fu et al in [21]. Here, the authors combined saliency maps with graph-cut interactive segmentation tool by Boykov and Jolly [8]. Instead of using human labels to perform segmentation, the authors proposed to use saliency data to perform human-like region labelling. Boykov's algorithm requires a user to draw curves on near the boundary of foreground and background regions. The authors point out that saliency maps produced by spectral residual approach highlight mostly edges of objects of interest thus there is need to perform estimation of where the center of object of interest is located before providing labels to the segmentation tool. To overcome this problem the authors proposed to perform iterative segmentation each time updating the location of seed labels. The assumption here is that once the first seed locations are provided the segmentation algorithm starts region growing in directions where an actual object is located. Then the results of the segmentation are used to move seed labels into the region of interest. The authors did not perform quantitative evaluation nor did they provide comparison to other methods.

Another work using segmentation can be also found in [44]. Here, the graph-cut segmentation is used for refinement of the final saliency map. The authors proposed to model saliency via joint-optimization of saliency labels computed on superpixels. Thus the first step of their approach is superpixel segmentation. From the input image color and texture information is extracted. In addition, for each superpixel its size and location are computed. Once features are computed saliency of each superpixel is estimated using appearance model. Then graph-cut optimization is used in order to refine the model. The smoothness term includes intensity difference. The evaluation is performed on Berkeley segmentation dataset.

Another method exploiting spectral domain properties of images for saliency detection was presented by Achanta et al in their frequency-tuned approach [2]. Instead of performing Fourier transform, the authors proposed to use a stack of filters to catch spectral properties of an image. Image color channels are passed though difference of gaussians (DoG) filters that are a bandpass type filters. Similarly to spectral residual approach the authors detect saliency by subtracting some averaged data from frequency response of an image. In this case, mean pixel value of pixels is used as average data. Unlike spectral residual approach the authors included color information into their approach. The input image is transformed into L*a*b* colorspace and saliency map produced by each color channel is then combined using Euclidean distances. Parameters of filters are selected by manual tuning. To obtain region-based map the authors use mean-shift segmentation in L*a*b* colorspace.

Alternative approach to saliency detection in spectral domain was presented in quaternion saliency detection [51]. Here the authors propose discrete cosine transform (DCT) representation to find singularity regions. For the purpose of decreasing computational time the authors propose to transform an RGB image into quaternion matrix: $I_Q = I_4 + I_1 i + (I_2 + I_3)j$ that allows faster DCT and inverse DCT (IDCT) computation. The key idea is that saliency can be computed from DCT signatures: $S_{DCT}(I) = g * \sum\limits_c [T(I_c) \circ T(I_c)]$, where $T(I_c) = IDCT(sgn(DCT(I_c)))$, where I_c is c'th image channel, and sgn is signum function and g is a Gaussian smoothing filter. The authors use quaternion "direction" as a signum function. In addition, the authors include Viola-Jones face detection. The final map is constructed by linear combination of quaternion DCT saliency and the face conspicuity map. The size of the output map is 64 × 48 px. The main advantage of this work is very high computational speed. The authors report raw map inference time about 0.4 ms excluding time necessary for resizing and anisotropic filtering.

An approach based on DCT representation for saliency detection can be found in discriminant saliency method by Gao et al described in [22]. This approach is based on marginal diversity of Kullback-Leiber divergence. The authors compute DCT with different basis functions including detectors of corners, edges, t-junctions, etc. The coefficients of DCTs is then used as a function. Saliency of each pixel is then defined as: $S(x,y) = \sum\limits_i^{2n} w_i R_i^2(x,y)$, where w_i is a marginal diversity of a feature output F_i and $R_i = max[I * F_i(x,y), 0]$. The saliency map generation process is iterative. After a map is computed feature output near detected salient region is set to zero and the process is repeated again. The output maps of this method are region-based. For the evaluation of the algorithm the authors used PASCAL database [17].

There is a group of methods aimed at detection of saliency via depth information. For instance, Ouerhani et al [46] presented a depth-based method extending Itti's salient detector by using depth information as one of information cues. The authors proposed three different depth features, namely depth, mean curvature and depth gradient. However, in their experiment they utilize only depth and color features. A quite similar approach with some minor differences can be found in [20] with application to robotics.

In [6] a method of saliency detection based on cloud point data has been reported. The authors proposed an unsupervised hand-tuned model for images acquired from a time-of-flight camera. The method relies on multi-scale local surface properties feature that is linearly combined with distance of pixels to the camera plane. The authors based their work on the assumption the closest object to the camera is salient. The authors performed evaluation of their approach on grayscale synthetic images obtained by rendering.

Another work performing fusion of 3-D data with visual features for saliency detection has been presented in [15]. This work was aimed at proving a solution for guiding focus of blind people by using information from wearable stereo cameras. They addressed this problem by utilizing saliency as one of features helping to inform blind people about objects around them. The authors proposed to use depth-information for weighting saliency of objects. They define two-thresholds d_{min} and d_{max} within which objects are more likely to be salient. Also the authors compute depth gradient over time to estimate the motion information that is after being processed using Difference of Gaussians to construct conspicuous map. For visual feature the authors utilize illumination intensity, red/green opposition and blue/yellow opposition being passed through multi-scale Gaussian filters. The final saliency map is obtained by linearly summing feature outputs.

2.3 The proposed innovations

In the following sections two saliency detection methods are proposed. Both proposed methods address the problem of finding salient regions, since region-based maps have a broader set of possible applications. Although both methods have the same task, the approaches are different. One method addresses saliency detection via visual features, while the other uses estimated depth information as a primary cue. The first method will be referred as a segment-based and the later will be referred as a depth based method.

The proposed segment-based approach addresses the problem of salient region detection by estimating saliency at segment-wise level not at pixel-wise level. Idea of using segmentation representation is not novel. However, normally segmentation is done after saliency is computed like it was done in [1], [57], [21], [2] and [10]. The most related work to the proposed segment-based approach can be found in [44]. However, in the proposed approach the set of features is more broad and includes higher level relations. Another difference is that [44] used a superpixel segmentation, the proposed approach uses a mean-shift based segmentation, that results in larger segments size and thus closer approximation to objects.

The proposed depth-based approach models saliency through 3-D spatial relationships of objects. In comparison with [46], [20] and [15] the proposed approach aims at prediction salient regions rather than separate pixels (or eye-fixations). In addition, a broader set of depth features is synthesized with respect to other models. The above mentioned methods obtain depth maps using special hardware, whereas in the proposed method depth maps are estimated form a 2-D image. The proposed depth-based approach models labels using CRMs thus allowing better treatment of neighbourhood context thus spatial dependencies of objects are taken into account.

2.4 Segment-based saliency detection

Examining the state of the art methods it is clear that the requirements in terms of saliency in multimedia applications, like image retrieval, scene detection and others, are not fully satisfied. The main problem is that as a rule the output map produced by a saliency detector highlights only small parts of objects of interest like edges and high-contrast points. This kind of maps sufficiently matches with maps obtained from experiments with eye trackers. The human vision system (HVS) has unattainable performance thus is able to recognize objects having very sparse data. However, when we deal with multimedia applications it is essential to extract the whole object of interest instead of few parts of it.

The most feasible solution for the extraction of objects from images is to perform image segmentation. In case of saliency detection there are two possible ways of applying segmentation: i) by computing a saliency map and deriving an average saliency value over a segment, and ii) by computing directly the saliency value of each segment. The advantage of the first method is that we can use any available method of saliency detection and simply apply segmentation to the output map. The advantage of the second method is that a more accurate estimation of saliency could be achieved due to the consideration of relationships of segments/objects rather than pixels. For this reason we have employed EDISON segmentation tool [11].

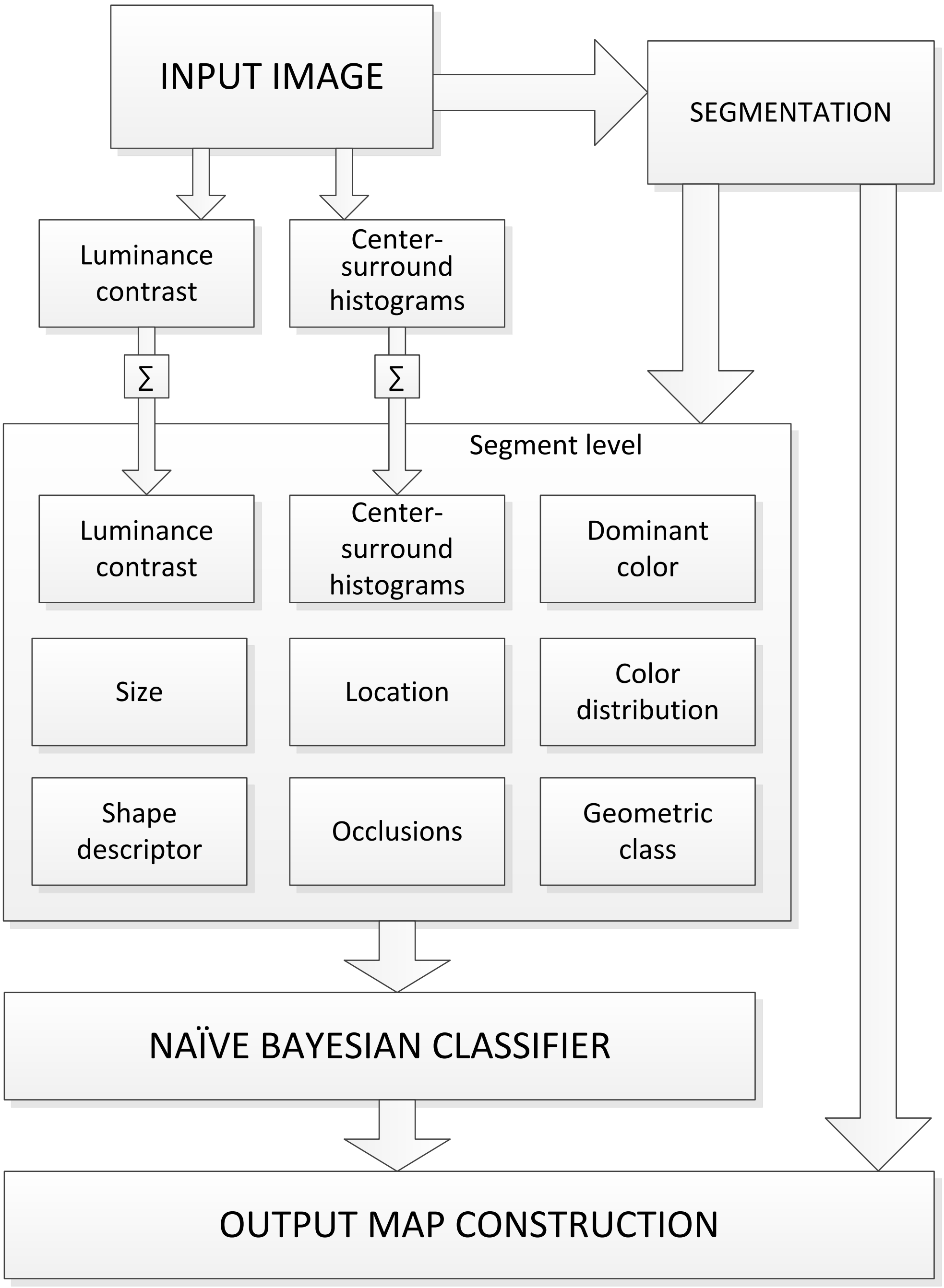


Figure 2.2: Segment-based saliency detection scheme

This choice was due to publicly available source code and satisfactory performance both from accuracy point of view and computational time. This segmentation tool is based on mean-shift segmentation. To achieve better object shape estimation the default parameters were tuned, their values are reported in Table 2.1.

Table 2.1: Segmentation parameters

Minimum region area	$im_{height} \times im_{width} \times 0.005$
Spatial bandwidth	10
Range bandwidth	7.5
Gradient window radius	2
Mixture parameter	0.3
Edge strength threshold	0.7

In this model saliency is detected mostly due to visual features that are described below. Some of the features used cannot be extracted directly from segment data. Their values are computed first on the whole image, and segment-wise level is then obtained by averaging value of a feature on pixels of that segment.

Colors have a great impact on the perception of objects. Gelasca et al. in [24] described an experiment they conducted to discover colors impact on saliency. Their study proved that some colors are more likely to attract attention than overs. Figure 2.3 shows dependency of saliency likelihood on the color of an object. In the original work the authors proposed to convert image color values into CIELab colorspace. However, I have found out that better color conversion is achieved with HSV colorspace. In the original work, the authors assigned a weight to each color equal to the fraction of attraction caused by this color during the experiments. Unlike that, in this work prior value of saliency for each color is defined by its position in the table of saliency likelihood, such that dark green has prior value of 0 and for red this value is 1.

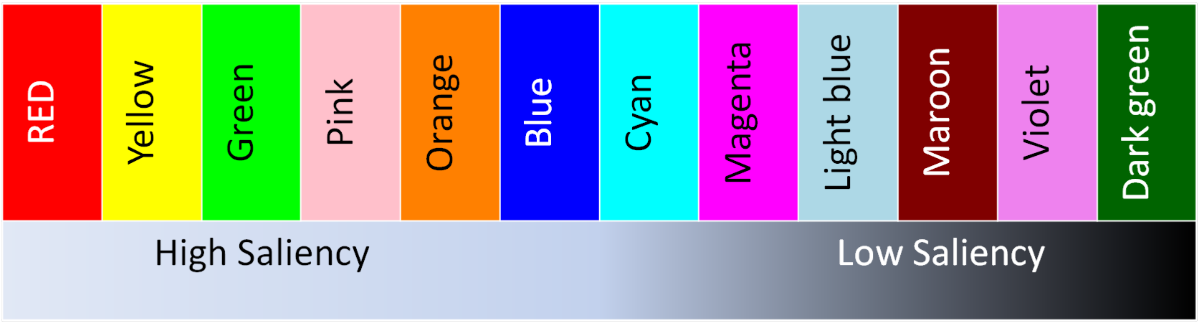


Figure 2.3: Saliency dependency on colors

In addition, a feature taking into account colors distribution over an image is included. This feature is based on observation that the dominant color is very likely to belong to background and thus is unlikely to be salient. For each segment its average color is computed and that match with our colorspace. For each color tone a corresponding weight is computed as follows:

$$\begin{aligned} w'_c &= \frac{1}{\sum_{i \in S_c} a_i}, \\ w_c &= \frac{w'_c - min(w'_i)}{max(w'_i) - min(w'_i)}, \end{aligned} \tag{2.1}$$

where S_c is a set of all segments assigned to color c, a_i is the area of segment i, w'_c and w_c are the unnormalized and normalized weights correspondingly.

As it was mentioned above human attention is sensitive to contrast. In order to exploit this property luminance contrast is included into the proposed model. This feature is measured on a downscaled by factor 8 version of an image. The motivation is that maximum contrast value is usually observed on edges and glare spots, while downscaling allows to catch global contrast changes. The luminance contrast LC is computed as follows:

$$LC(x,y) = \sum_m \sum_n \frac{|L(x,y) - L(x+m, y+n)|}{\sqrt{m^2+n^2}}, \tag{2.2}$$

where $L(x,y)$ is the luminance value of the pixel with coordinates (x,y),

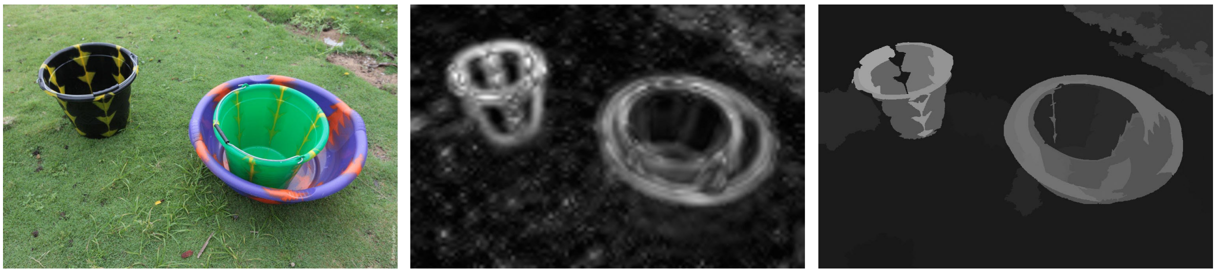


Figure 2.4: Luminance contrast feature output. From left to right: input image, luminance contrast without segmentation, luminance contrast after segmentation.

and $m, n = \{-2, -1, 1, 2\}$ denote relative coordinates of neighbor pixels. Example of luminance contrast output is shown in Figure 2.4.

The idea to measure the distance between foreground and background for saliency detection was used in several previous works. The underlying idea is that usually the histogram of the foreground object has a larger extent than its surroundings. In our work we employ center-surround histogram filter from [38] with slight modifications. The input image is scanned by two rectangular windows R_f and R_s, both having a similar area and R_s encloses R_f (thus R_f is a notch inside the window R_s). We use the following size ratios of windows: [0.3, 0.7], which were defined experimentally with respect to the minimum image dimension, as well as the following three aspect ratios: [0.5 1 1.5]. Specifically the distance of foreground and surrounding histograms is computed as follows:

$$dist(R_s, R_f) = \frac{1}{2} \sum \frac{(R_f^i - R_s^i)^2}{R_f^i + R_s^i}, \qquad (2.3)$$

where R_s^i, R_f^i are surrounding and foreground histograms, respectively. In addition, unlike the original work, windows are moved with overlap of 0.1 with respect to the corresponding size of the window to eliminate boundary effects within the scanning windows. Histogram distances are computed at each scale and aspect ratio. Then, they are normalized and summed into a single map. Finally, after the computation of these features, we assign an

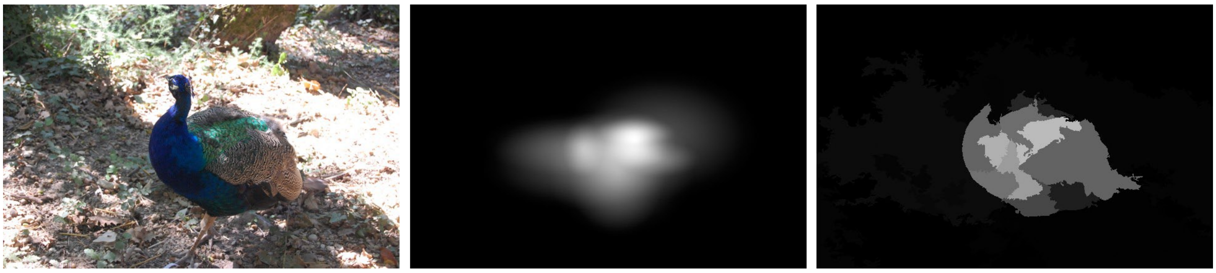


Figure 2.5: Center-surround histogram filter. From left to right: input image, center-surround histogram output before segmentation, center-surround histogram output after segmentation.

average value of each global feature to each segment of the input image. Figure 2.5 shows the principle of this feature.

In addition to visual feature, geometric features have been included. The location feature has been included into the scheme due to the fact that photographers generally place the object of interest to the center of images. The location M_S of the segment S is computed as follows:

$$M_S = ((\sum_x \sum_y f(x,y)p(x))^2 + (\sum_x \sum_y f(x,y)q(y))^2)^{\frac{1}{2}}, \tag{2.4}$$

with

$$f(x,y) = \begin{cases} 1 & \text{if } x \in S \text{ and } y \in S \\ 0 & \text{otherwise,} \end{cases} \tag{2.5}$$

$$p(x) = \frac{mx}{2} - x, \tag{2.6}$$

$$q(y) = \frac{my}{2} - y, \tag{2.7}$$

and mx, my are the corresponding dimensions of the image. Considering size, the object of interest usually occupies a significant portion of the image. Thus it is unlikely that a very small segment is salient. On the other hand natural background like sky, ground, and forest usually occupy large portion of area; thereby it is very unlikekly that a salient objects size exceeds some threshold. Therefore, the relative size of a segment could give relevant information about its saliency. Another feature exploiting geometric properties of a scene detects if a segment is occluded with others. The intuition here is that normally, the main object of a scene is placed in front of some background. Thus the background region becomes occluded by the main object. There exist quite accurate methods for occlusion detection, however most of them require a lot of computation. Therefore, here I propose a simple method of occlusion detection based on the segmentation map. For computation of occlusion, firstly its necessary to compute spread of each segment. Spread shows numerical extend of a segment relative to its center of mass. One can think of spread as of a rectangle that is a rough representation of an actual region occupied by a segment. Spread of a segment is computed as follows:

$$mx_s = \frac{\sum\limits_{y \in Y_s} \parallel x \in [X_s^{<y>} < cx_s] \parallel}{\parallel Y_s \parallel}, \tag{2.8}$$

$$px_s = \frac{\sum\limits_{y \in Y_s} \parallel x \in [X_s^{<y>} > cx_s] \parallel}{\parallel Y_s \parallel}, \tag{2.9}$$

$$my_s = \frac{\sum\limits_{x \in X_s} \parallel y \in [Y_s^{<x>} < cy_s] \parallel}{\parallel X_s \parallel}, \tag{2.10}$$

$$py_s = \frac{\sum\limits_{x \in X_s} \parallel y \in [Y_s^{<x>} < cy_s] \parallel}{\parallel X_s \parallel}, \tag{2.11}$$

where (cx_s, cy_s) are coordinates of the center of mass of a segment s, X_s and Y_s are vectors holding all rows and colons of the segment s lies in, (mx_s, my_s) and (px_s, py_s) are coordinates of top left and bottom right points of the rectangle representing the spread. Then, if two segments have overlapping regions occlusion is detected by thresholding the area of their intersection. Once occlusion is detected foreground and background segments are detected by comparing their sizes. A segment with large size becomes background, and the other segment is treated as foreground.

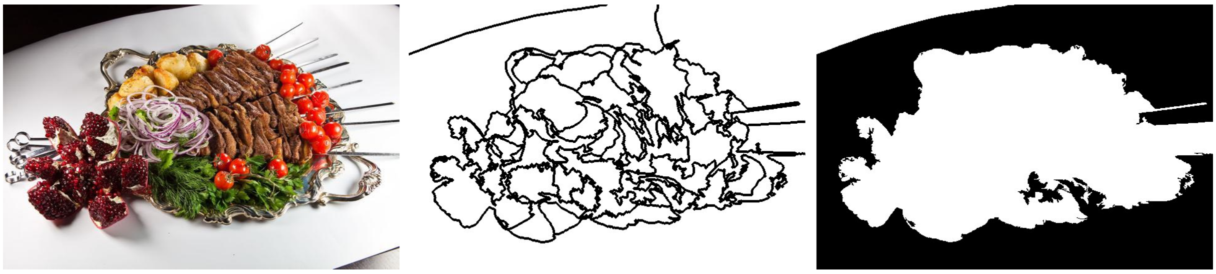


Figure 2.6: Occlusion feature output example. From left to right: input image, its segmentation map, occluded segments (drawn in black).

The background and foreground segments receive values of -1 and 1 respectfully. Thus for each segment there is an $(n-1)$-element vector of occlusions, where n is the number of all segments. Final value of this feature is calculated by taking mean value of this vector. Example output of this feature is shown in Figure 2.4. In addition to occlusion, a feature counting number of neighbour segments is included into the model. The intuition here, is that normally an object of interest is composed from several segments. These segments become neighbours to each other. On opposite background objects are represented by a few segments. Thus counting the number of neighbour segments adds some knowledge on the structure of the scene.

Moreover, another feature responsive for geometric properties of a scene is geometric class detector. It is based on the method described in [30]. This method performs geometric classification of surfaces found in 2-D images. As an output the method provides such classes as sky, ground, vertices. This classification is quite relevant for the task of saliency prediction - rarely one would make a picture to capture ground, or sky, with some obvious exceptions like a picture of a sunset or sunrise. Also experiments have shown there is slight correlation between class vertices and ground-truth saliency data. The geometric classifier performs classification using

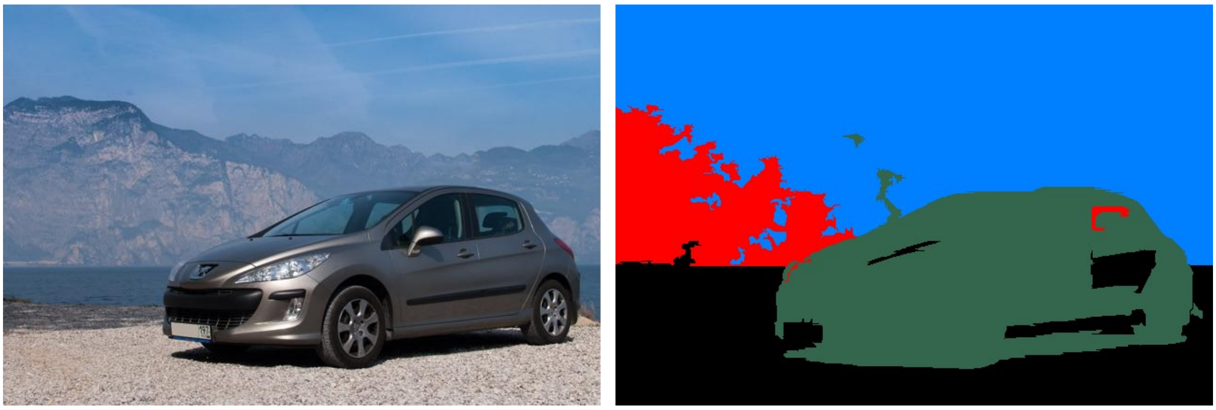


Figure 2.7: Geometric class detector output. On the left the input image detector output is on the right. Different geometric classes are shaded with different colors. It is evident that vertices class (drawn in green) is related with the salient object of this image.

internal segmentation map. In the original work the authors used super-pixel segmentation. Segmentation is quite expensive operation in terms of time. Thereby, in order to avoid running two independent segmentation operations, an interim segmentation map obtained while running initial image segmentation is provided to the geometric classification method. This interim segmentation map is oversegmented with respect to the final segmentation map, thus to some extend its properties are close to that of super-pixel segmentation. Since segments in interim and final segmentation maps are different matching is needed. Once geometric classes are computed the 2-D map with pixel values equal to geometric class of corresponding segments is constructed. This map is then used to compute the geometric class of a segment in the final segmentation map by finding the most frequent pixel value for the segment.

The shape of the object also can contribute to visual importance. For instance, skewed objects are unlikely to be important parts of a scene. Another example is a rectangular objects that are likely to be a picture of an information board or sign. There exist a number of chain-like methods to describe the shape of an object, however, most of them require large computational resource, or results in a high-order vector of data. For the case of saliency detection there is no need to have very precise shape information, on opposite only some properties of object's shape are necessary. For this reason, here I propose a shape coding method that allows for very compact description of shape properties. The visualization of feature computation is shown in Figure 2.4. Each segment is fitted into 5x5 grid. If a segment has line-like shape, then its shape is preserved and it occupies only on row or column. The occupancy of grid cells is then used for three descriptors: horizontal, vertical and center. Each of this descriptors counts number of certain cells occupied by a segment according to the descriptor's map. The output of horizontal and vertical descriptors is then multiplied element-by-element wise and summed up into one value that is later is summed with the output of center descriptor. Once this feature computed over the whole dataset the range of values of this feature is divided into six regions. The index of this region is used as a final output of the feature. Although this method seems to be naive the experiments have shown that it is able to encode different types of shapes and moreover there is correlation between output of this descriptor and ground-truth saliency data as it is shown in Section 2.7.

Features described above represent each segment by a 12 element vector. For discrete features, namely for geometric context and occlusions values are represented by multivariate indicator functions. Other features are represented using gaussians. Modelling is done using probabilistic framework. In this case it is Naive Bayesian classifier. Although some of features used have correlation thus ruining the assumption of feature independence, the performance of this classifier is satisfactory. Classification is done per segment. Thus the output of the classifier holds saliency estimation for each segment, that are further used to reconstruct the corre-

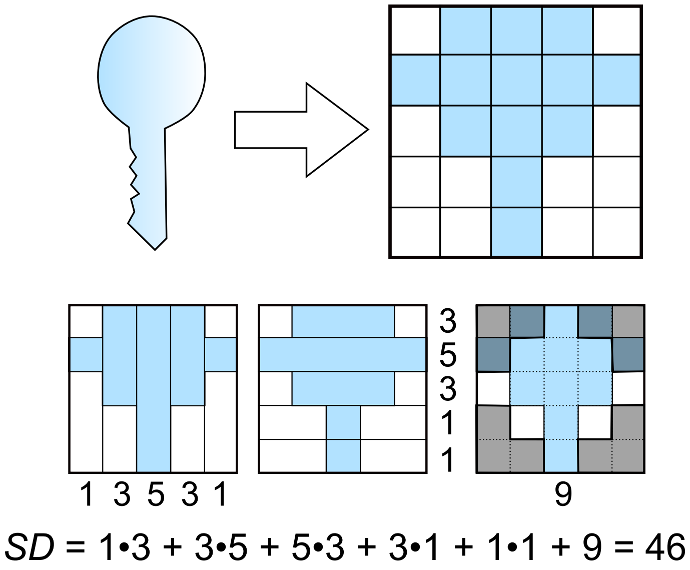


Figure 2.8: Shape descriptor

sponding saliency map using segmentation data. Learning was done using expectation-maximization (EM) learning method. The overall scheme of the method is shown in Figure 2.2. The evaluation results can be found in Section 2.7.

2.5 Depth-based saliency detection

In this section a depth-based saliency detection method is presented. Here, the main idea is that saliency to some extend can be estimated through analysis of spacial layout of the scene. Unlike previous works in this topic, depth information is not acquired using dedicated hardware, but instead is estimated from input 2-D image. Recently, there have been achieved some progress in the area of depth estimation that made this method possible. One of the most important components in this method is depth estimation that is done using the algorithm proposed by Saxena et al. in [50]. This method performs depth estimation modelling Markov Random Field (MRF) using texton and gradient-based features. The image is divided

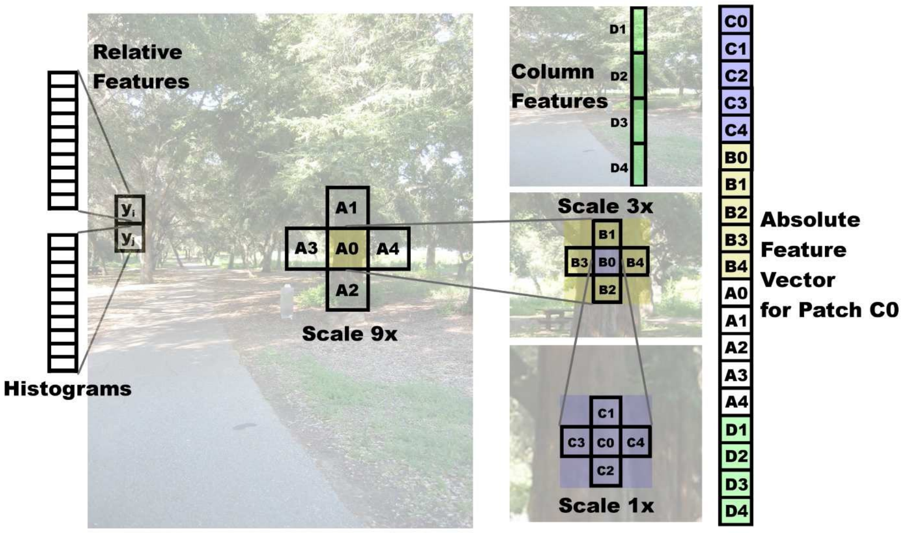


Figure 2.9: Depth extraction [50]. Depth features are extracted at 3 scales. Relative and absolute feature vectors are composed from texton feature filter output.

into patches and a single depth value is then estimated for each patch. For each patch two types of features are calculated: absolute and relative (see Figure 2.9). Absolute feature estimates how far a patch is from the camera plane, while the relative feature estimates whether two adjacent patches are physically connected to each other in 3-D space. Both absolute and relative features are calculated using texture variations, texture gradients and color. Original work applied depth estimation to outdoor scenes, mainly including rural-like pictures. However, the test have shown that even on other scenes depth estimation is quite reasonable. In my implementation the grid of MRF is set to 50x50. Since it uses probababilistic framework there is need to train its parameters. That has been done by learning parameters on the database the original work was using. In addition, 50 indoor images acquired using stereo cameras were added into the database in order to improve performance for indoor scenes. Figure 2.10 shows an example of depth estimation.

Similarly to the method described in Section 2.4 the detection is made in object-wise domain with the same motivation. Likewise, this is done

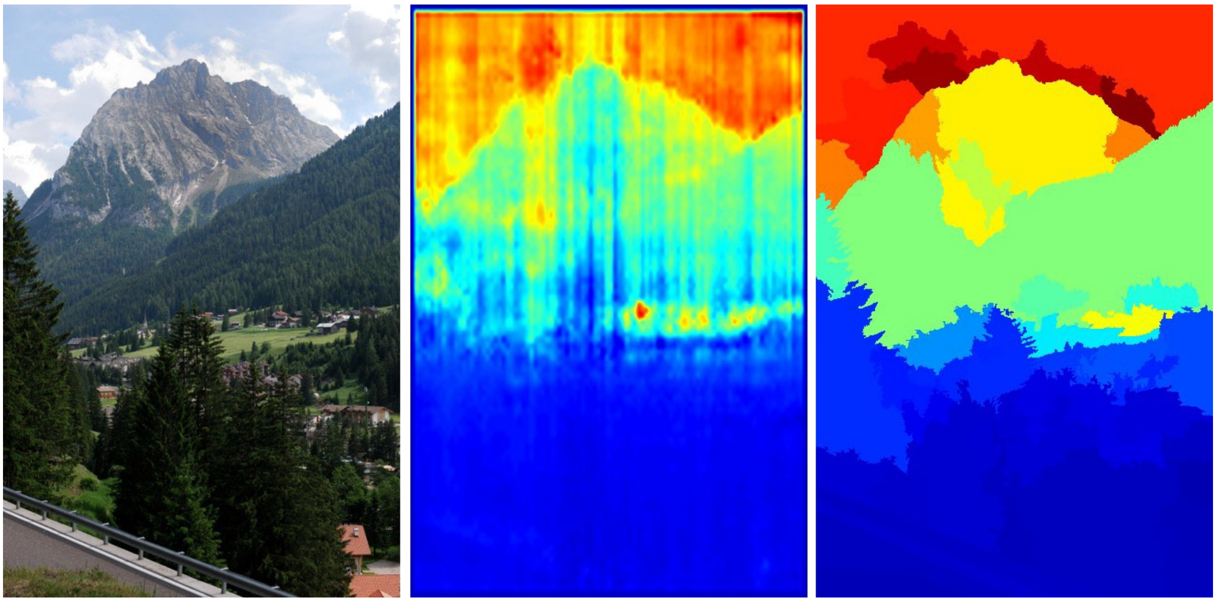


Figure 2.10: Depth estimation example. From left to right: input image, depth estimation map, depth estimation after segmentation. As one can notice after the segmentation procedure the estimation results in a quite realistic depth layout.

using Edison segmentation method [11]. Since the depth-estimation implementation used works with MRF with the grid of 50x50 size, its output is presented also as a grid. The depth of each segment then is defined as the spatial average value of depth of the cells it is lying on.

Another common component with the visual-based method is geometric-class detector [30]. However, here the motivation for its inclusion is different. In the previous method this feature was responsible for layout detection, whereas here it adds some semantics to segments. Even though the geometric-class detector provides classification over seven classes of surface types, here these classes are merged into 3 final values: sky, ground and others.

Although depth by itself provides some important information about scene geometric properties, it is more of interest to exploit depth relationship of different object in the scene. There are several features responsible for that. One of them describes how far (in z-axis) an object is from the farthest object in the scene. This feature returns distance to rear-plane. Another feature describes sum of absolute difference of an object to its adjacent objects divided by the number of neighbours. This feature is responsible for depth contrast, with intuition that if an object is distant from its neighbours than probably it can be an object of interest. Likewise, if two objects are close to each other it is likely they are both two parts of a foreground or background scene. Figure 2.11 illustrates this principle.

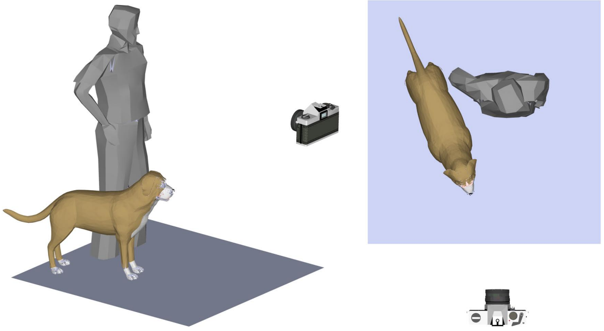


Figure 2.11: Spatial closeness. Objects of the same context are placed spatially close to each other.

Analysing images of landscapes one can notice that they mostly consist of flat surfaces. Such surfaces may have different orientation in z-y and z-x axes with respect to the camera plane. For example consider shooting a building in a city. It is obvious that the best view is achieved when the elevation of the building is parallel to the camera plane. Thereby it is of interest to measure the angle of an object with respect to the camera plane. For this purpose a tangent-sensing feature is included. It is computed as follows:

$$\tan_y(r) = \frac{1}{\parallel X_r \parallel} \bullet \sum_{X_r} (d_x - d_{x-1}) \tag{2.12}$$

$$\tan_y(obj) = \frac{1}{\parallel R_{obj} \parallel} \bullet \sum_{R_{obj}} (\tan_y(r)) \tag{2.13}$$

where X_r is a set of all pixel coordinates of object *obj* in row r and R_{obj}

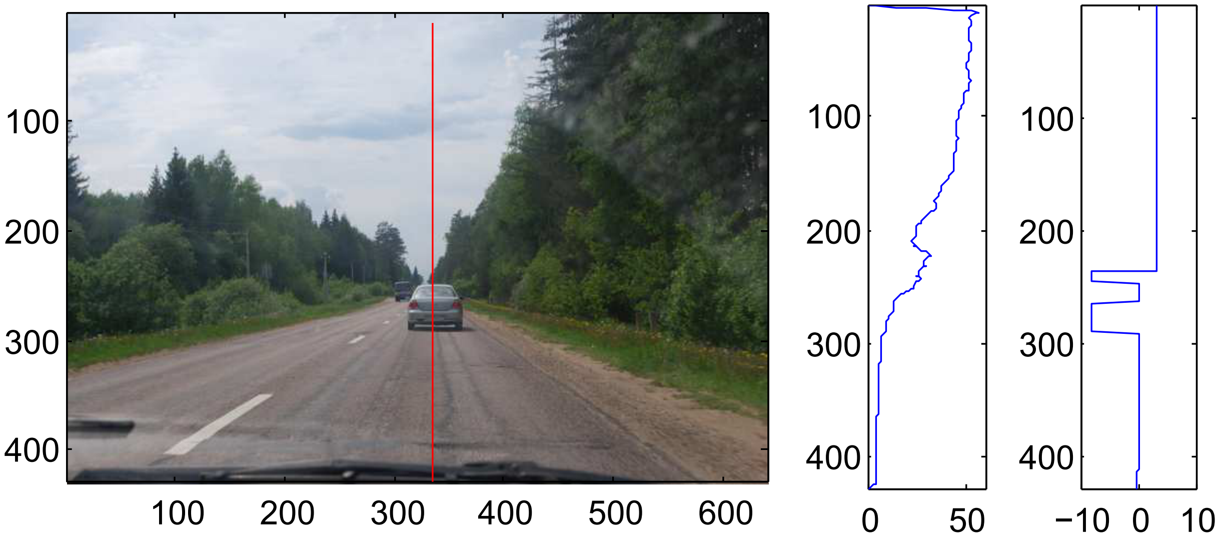


Figure 2.12: Tangent-like feature output example. From left to right: input image with the red line denoting the test column, estimated depth within the column, ZY-angle estimation per segment. It is clear visible how the rapid changes of angles in rows 200-300 corresponds to the body of the car. Although the tangent-like feature does not provide accurate angle estimation due to limitations of the depth estimation, still relative variance in plane orientation is detected.

is a set of all rows belonging to the object. As can be seen from (2.12) tangent is computed in pixel-wise fashion. In the same way tangent in X direction is computed by averaging column tangents.

To obtain more complete information about the scene the coordinates of object's geometric center of mass in X and Y axes are included into the feature set. Here, the intuition is that normally a person would place an object of interest close to the center of the frame rather on its boundary. These features are computed in the same way as it was done in segment-based method using Equations 2.6 and 2.7.

Portraits and group photos usually form a large portion of personal albums. Our test have shown that depth estimation turns out to work poor predicting z-coordinate for faces interfering with overall performance of saliency estimation for such images. The same drawback is present when an image contains very flat objects, and depth estimation provides little information. Thereby, to allow better performance two additional visual features are employed. For each object its average color is measured and matched to the closest representation from a fixed set of colors. This fixed set consists of 9 tones, namely red, yellow, green, cyan, violet, pink, white, grey and black. The colors are defined by splitting the HSV color space. Color tones are obtained by dividing H component of HSV color space into 6 equal regions. Monotones are obtained by dividing V space into 3 equal regions. Input color matched to grey tones if its S and V component satisfy the following condition:

$$V < 0.1 \vee S < (0.1 + \frac{0.01}{V^2}) \tag{2.14}$$

Another visual feature is color contrast computed based on object's dominant color difference with respect to all other objects as it is described in 2.4. To sum up the feature vector for each segment consists of ten variables. In addition, it is of interest to exploit pairwise labelling information between adjacent objects. This is done through measuring similarity of adjacent segments. Similarity is computed based on difference in depth. The intuition behind this is that if two segments have close depth estimation then it is likely that they are two parts of one object composed from several segments and thus their saliency estimation should be the same. Besides depth difference similarity is also defined by whether two segments have the same dominant color and geometric class.

$$sim(i,j) = \begin{cases} (d_i - d_j)^2 + \sum\limits_{k=1}^{9} (c[k]_i - c[k]_j)^2, & \text{if } gc_i = gc_j \\ 0, & \text{otherwise,} \end{cases} \tag{2.15}$$

where d_i is the average depth of segment i, c_i is its color histogram and gc_i its geometric class. Here color histogram is obtained by counting the number of pixels matched to the set of colors described above.

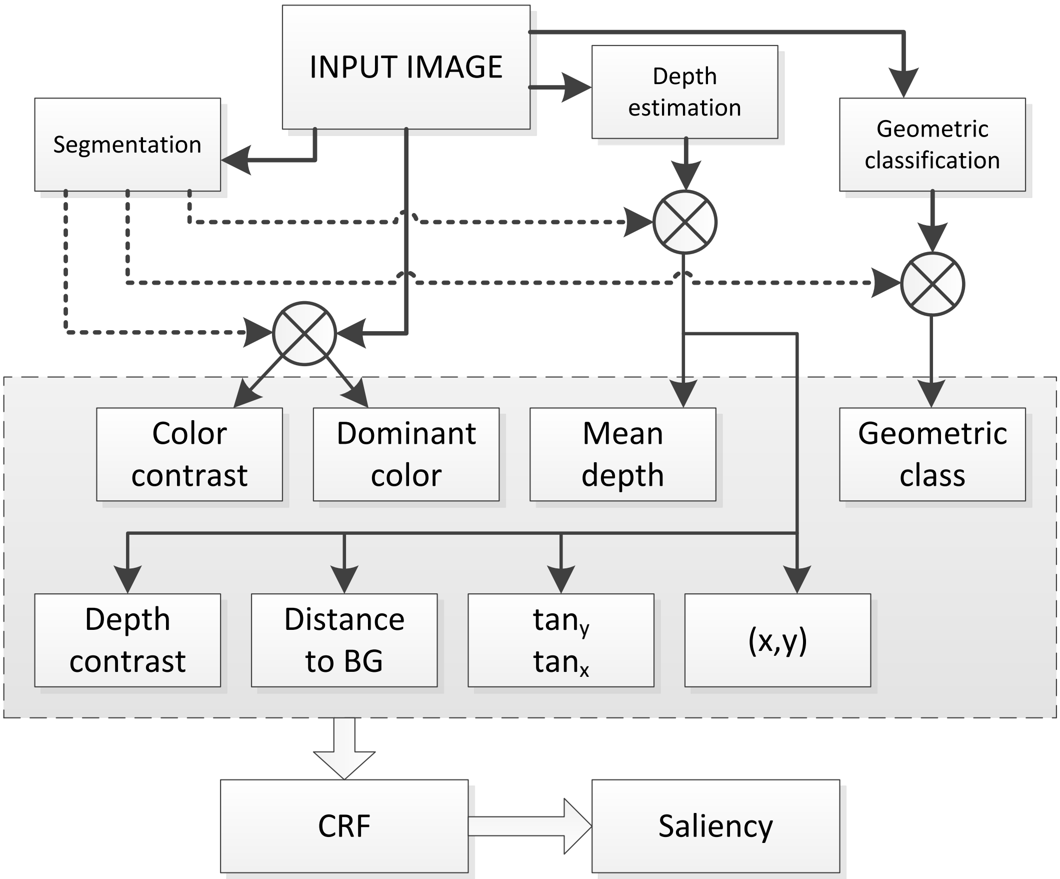


Figure 2.13: Depth-based saliency detection scheme

The modelling of saliency in this case is done using conditional random field (CRM). Thus the optimization energy is described as follows:

$$E(\boldsymbol{sl}, \boldsymbol{x}, \boldsymbol{w}) = \sum_{i \in S} w_s \phi_i^s(sl_i, \boldsymbol{x}) + \sum_{(i,j) \in A} w_p \phi_{ij}^p(sl_i, sl_j, \boldsymbol{x}) \tag{2.16}$$

where the first term describes singleton energy and the later is pairwise energy; S is the set of all segments and A is the set of all adjacent segments. The resulting graphical model is a loopy tree. Inference in this model is done using graph-cut optimization for binary labelling [7]. Both singleton and pairwise features are modelled using linear logistic regression. Singleton and pairwise parameters are jointly learnt using stochastic gradient descent. The overall detection flow is shown in Figure 2.13.

2.6 Saliency evaluation dataset

Both for evaluation and training of the proposed models a ground-truth database was created. Although there exists some datasets for evaluation of saliency, there are not suitable for the proposed models. These databases were collected using eye-trackers (for instance [34][1] and [58][2]) thereby ground truth data is represented by fixation points. However, the methods proposed operate with segments rather than with separate pixels. Thus there is need to perform matching of fixation points to segments. Performing this task automatically is not possible due to limitations of existing segmentation tools and sparse nature of fixation points. It is common with current segmentation tools that a real object is being represented by several segments, and it often happens that fixation points can be absent in several object's segments. In addition, tasks given to humans while collecting these databases were different. For example, in [33] the authors

[1] http://people.csail.mit.edu/tjudd/WherePeopleLook/index.html

[2] http://live.ece.utexas.edu/research/doves

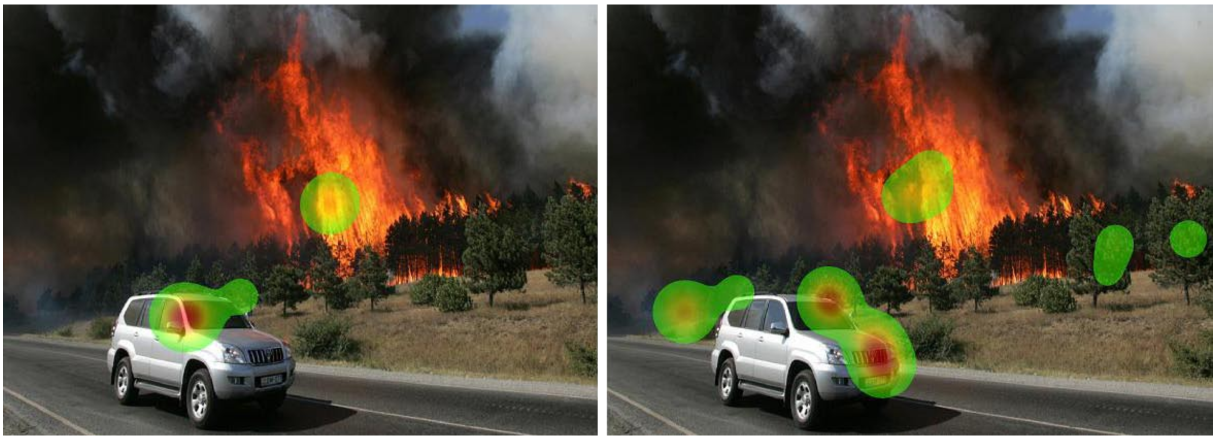


Figure 2.14: Fixation points over time [49]. On the left fixations over first 2 seconds, on the right fixations over next 5 seconds. At the first moment human eyes capture main objects of the scene, then discovering some surrounding to gather complete scene information.

investigated the order of objects causing attention in time. People tend to expand the grasp over the image in time (see Figure 2.14). Thus due to the long exposure of images to users almost the whole area of images were covered by fixation points. Another problem is that the very first fixation point usually lies in the center of an image, due to the prior people have about saliency. These problems makes it necessary to perform some kind of filtration of fixation points and supervision of matching results. Another option is to use a database where objects were selected by users by hands, rather than with eye trackers, like it was done in [38][3]. However, in this work saliency objects were selected via rectangles, that again leads to the problem of matching.

Thereby to avoid these problems and have better control over ground-truth data content I created new database. The first important task was to select proper images. For example, there is no sense to use artistic images, since the saliency will be to subjective for this kind of media. The database should consist of images with unambiguous content to diminish the effect of

[3] http://research.microsoft.com/en-us/um/people/jiansun/SalientObject/salient-object.htm

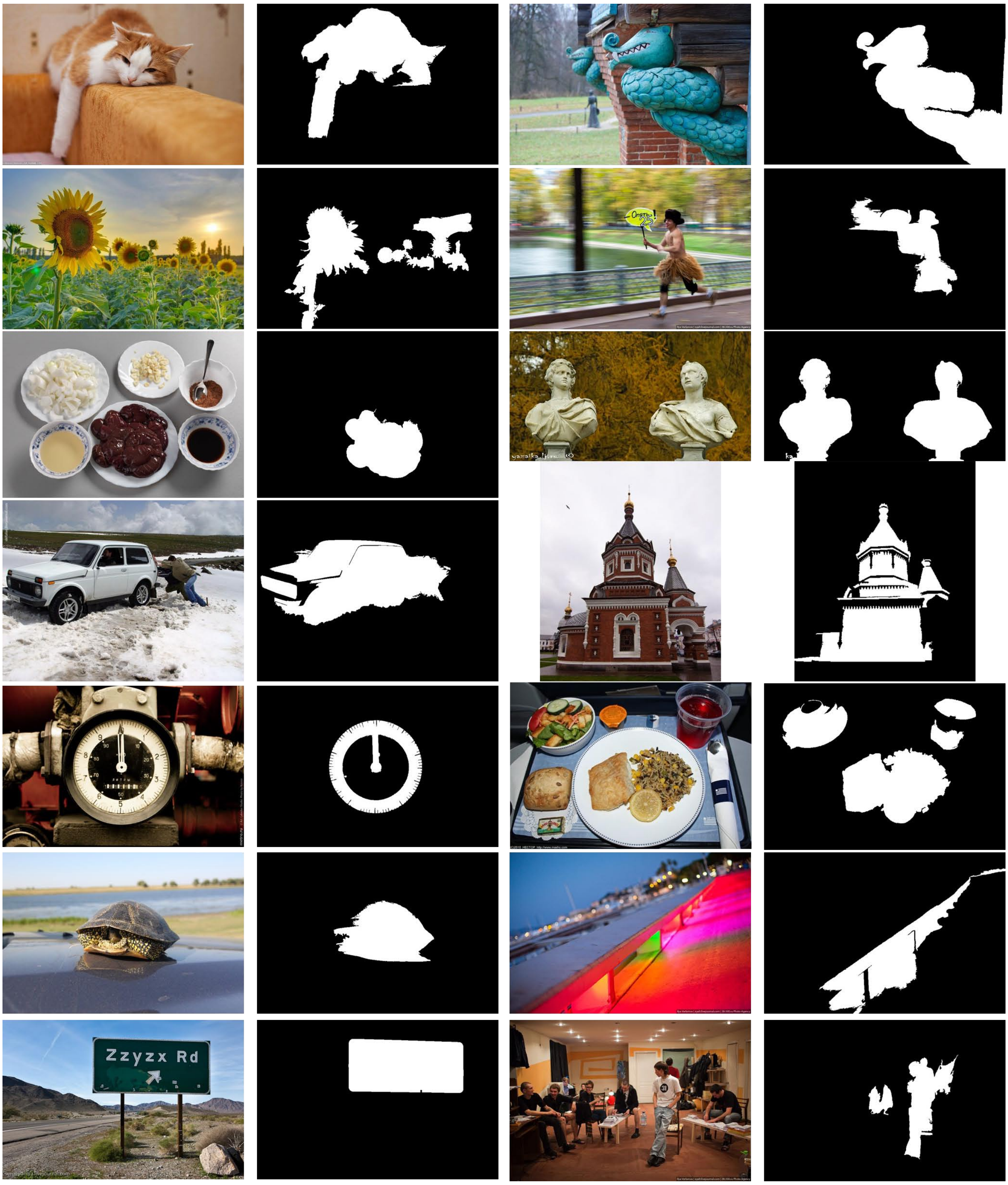
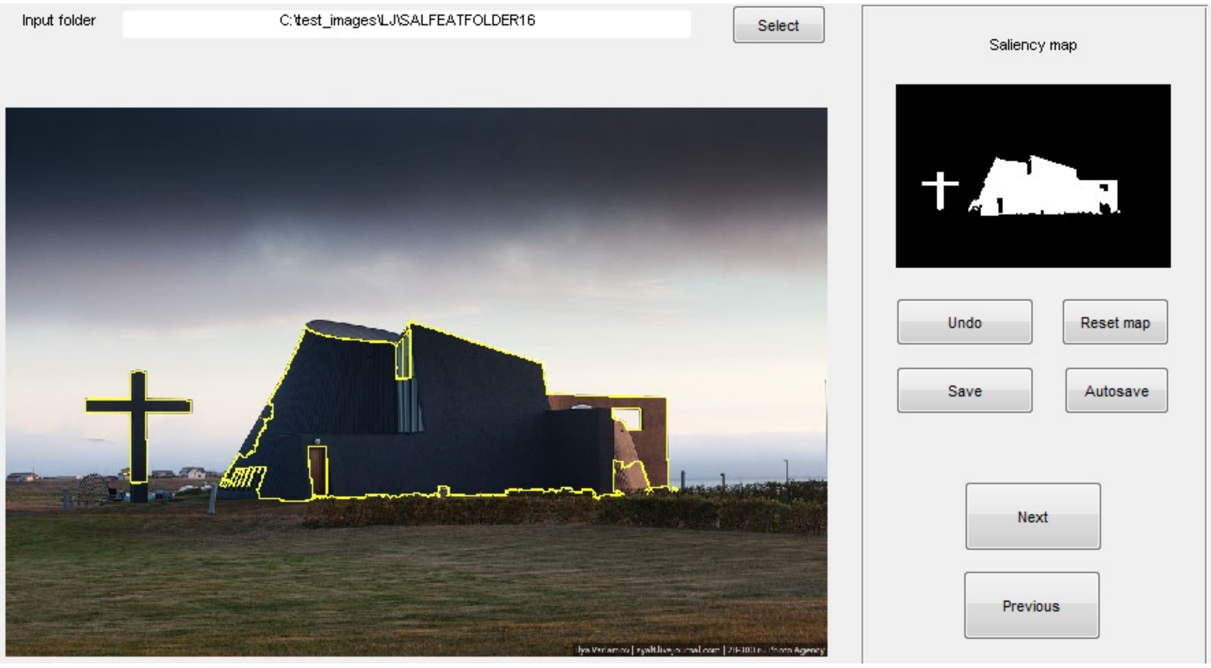


Figure 2.15: Example images from the dataset. From left to righ: input image, final ground-truth map.



Input folder C:\test_images\LJ\SALFEATFOLDER16

Figure 2.16: Ground truth collection UI. Selected areas are outlined with yellow contour, the view on the right highlight selected areas on the selection map.

subjectivity of the ground truth data. The decision here was simple. Since the methods proposed here are aimed at dealing with media available in the Internet and shared through social networks the best source is there. Thereby the database consists of images one would capture during their everyday life, travelling and attending some events. The dataset contains images of categories like: landscapes, building, monuments, cars, airplanes, trains, food, souvenirs, flowers, animals, pets, people and sports. Totally the dataset consists of 800 images. Figure 2.15 shows example images from the dataset with the corresponding ground-truth data.

The collection of ground-truth dataset was done by means of a simple user interface shown in Figure 2.16. Users were asked to click onto parts of an image they consider most important. Once a user clicks the image a segment where this click falls into highlight on the selection map view thus allowing users control over what has been selected. Once the user has finished selecting salient regions next image is shown to him. There is no time limitation on the image exposure, due to the task driven selection approach. The ground-truth dataset was collected from three users. Since saliency is a subjective property ground-truth data from users are different. Final decision on saliency of a particular segment was obtain by marks agreement at least of two users. On average each image has 23 segments with 5 of them marked as salient. The proposed methods require parameter learning, thereby the dataset was divided into two parts: 600 images for training and 200 images for evaluation.

2.7 Evaluation

In this section the evaluation of the of methods described in Sections 2.4 and 2.5 is presented. Besides numerical evaluation, qualitative comparison to state-of-the-art methods is given.

The first part is the quantitative evaluation of the proposed methods on the test dataset described in Section 2.6. This test allows for numerical evaluation of the proposed approaches due to the ground truth data. Ground truth data is a list of segments selected by users as salient. Thus the most straight-forward method to compute accuracy is to measure how many segments in a predicted map match to the corresponding ground-truth map. However, this approach neglects the size of the segments. While in cases where segments of an image are of close size this problem is not critical, often segment size varies greatly over the image interfering with the perceptual estimation of saliency. Thereby, instead of matching segment numbers, the area of intersection of estimated and ground-truth maps is used for evaluation. Another advantage of this approach is that when changing parameters of segmentation algorithm thus creating another segmentation map still it is possible to compare results since only intersection area is needed. The performance is measured using two parameters: F-score and accuracy. Likewise to many other works F_1 score is used:

$$precision = \frac{\sum\limits_i p_t^i \bullet w_i}{\sum\limits_i p_t^i \bullet w_i + \sum\limits_i p_f^i \bullet w_i}$$

$$recall = \frac{\sum\limits_i p_t^i \bullet w_i}{\sum\limits_i p_t^i \bullet w_i + \sum\limits_i n_f^i \bullet w_i}$$

$$w_i = \frac{area(i)}{\sum\limits_{j \in I} area(j)}$$

$$F_1 = 1.5 \bullet \frac{precision \bullet recall}{precision + recall},$$

where p_t and p_f are true and false positive labels respectfully, n_t and n_f are true and false negative labels respectfully, and w_i is the weight of the label i. Accuracy is computed as follows:

$$accuracy = \frac{p_t + n_t}{p_t + p_f + n_t + n_f}$$

The evaluation results are reported in Table 2.2. As it can be seen from the evaluation, both segmented-based and depth-based methods show high performance on the test dataset. The higher performance of segment-based method can be explained by failure of depth estimation on some images.

Table 2.2: Performance of the proposed methods

	segment-based	depth-based
F-score	0.69	0.59
Accuracy	0.85	0.71

It is of interest to see how each feature contributes to the final estimation of saliency. Here the correlation of saliency data and feature values are shown. Figure 2.17 shows dependencies of color distribution, luminance contrast, center-surround histogram filter and location feature values on saliency. For color-distribution it is evident if a segment is assigned color distribution value 5 and higher it is more likely that this segment is a part of foreground scene. Thus less distributed color in a scene is (higher value corresponds to less segments have this color), more likely it belongs to salient part. Likewise luminance contrast, center-surround histogram filter, location and number of neighbours feature (Figure 2.18) after a certain value indicate that it is more likely a segment to be salient. For geometric-class feature it is evident that it is more likely that a segment with geometric class 6, that corresponds to vertices class, is salient. Thus another geometric classes indicate opposite information. That is why in the depth-based method geometric classes are merged into 3 classes. Shape descriptor likelihood is more close to Gaussian distribution with mean value between 3 and 4. Figure 2.19 shows dependencies of depth-features on saliency. If to inspect mean depth value versus saliency dependency it is evident that salient objects are more likely to be displaced in range of 8 to 40 estimated units. Objects placed closer are more likely to be a part of ground surface, and more distant objects usually belong to background. Interesting observation can be done considering tangent-like features. While in X-Z plane it is more likely that the facet of an object of interest is orientated parallel to the camera plane, in Y-Z plane the facet of an object of interest is on average has some pitch.

Another part of the evaluation shows comparison of the proposed methods to 14 concurrent methods. Among concurrent methods there are both pixel-based and segment-based methods. There is no comparison to concurrent depth-based methods since to the best of my knowledge all of them require real depth data acquired by means of hardware, thus requiring construction of a database combining saliency and depth data. Although it

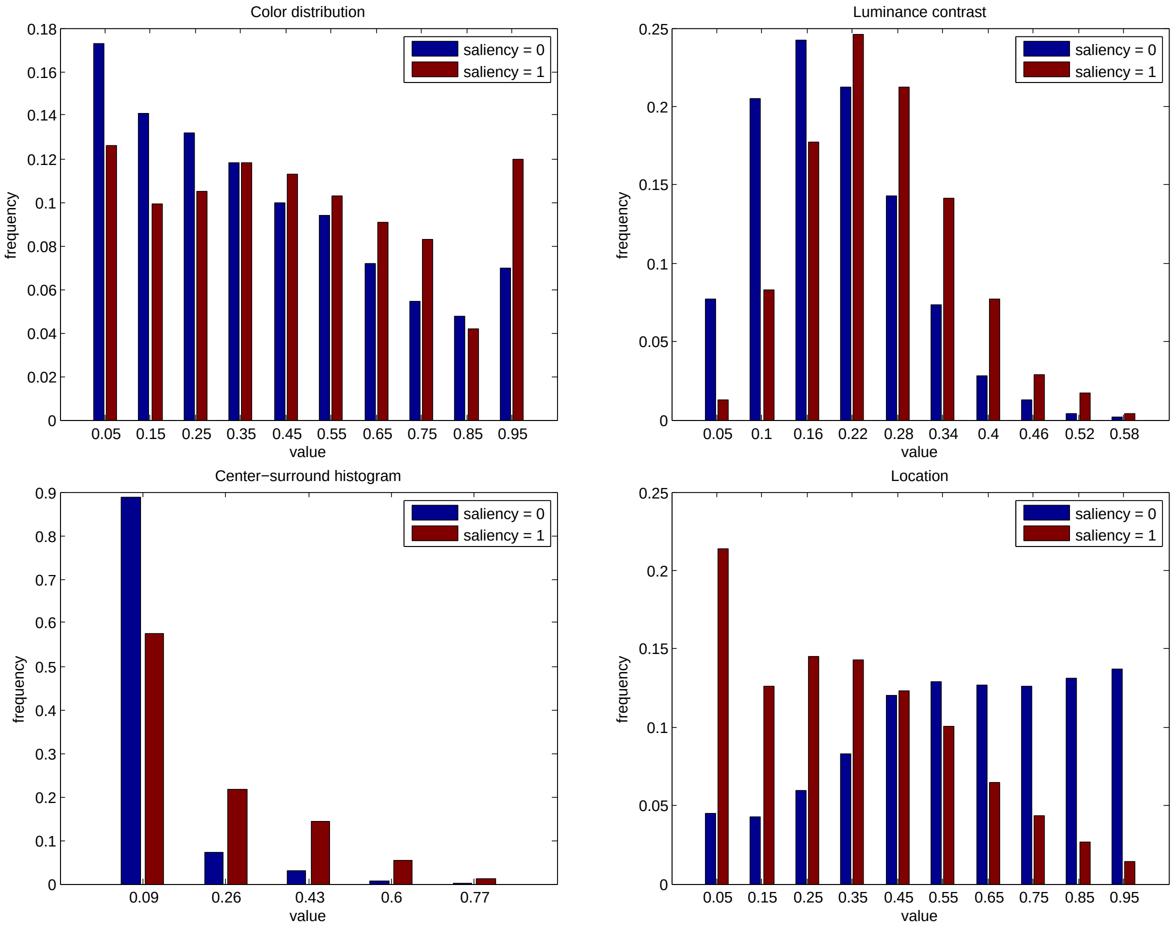


Figure 2.17: Features analysis. Here the dependencies of color distribution, luminance contrast, center-surround histogram filter and location features on saliency are shown.

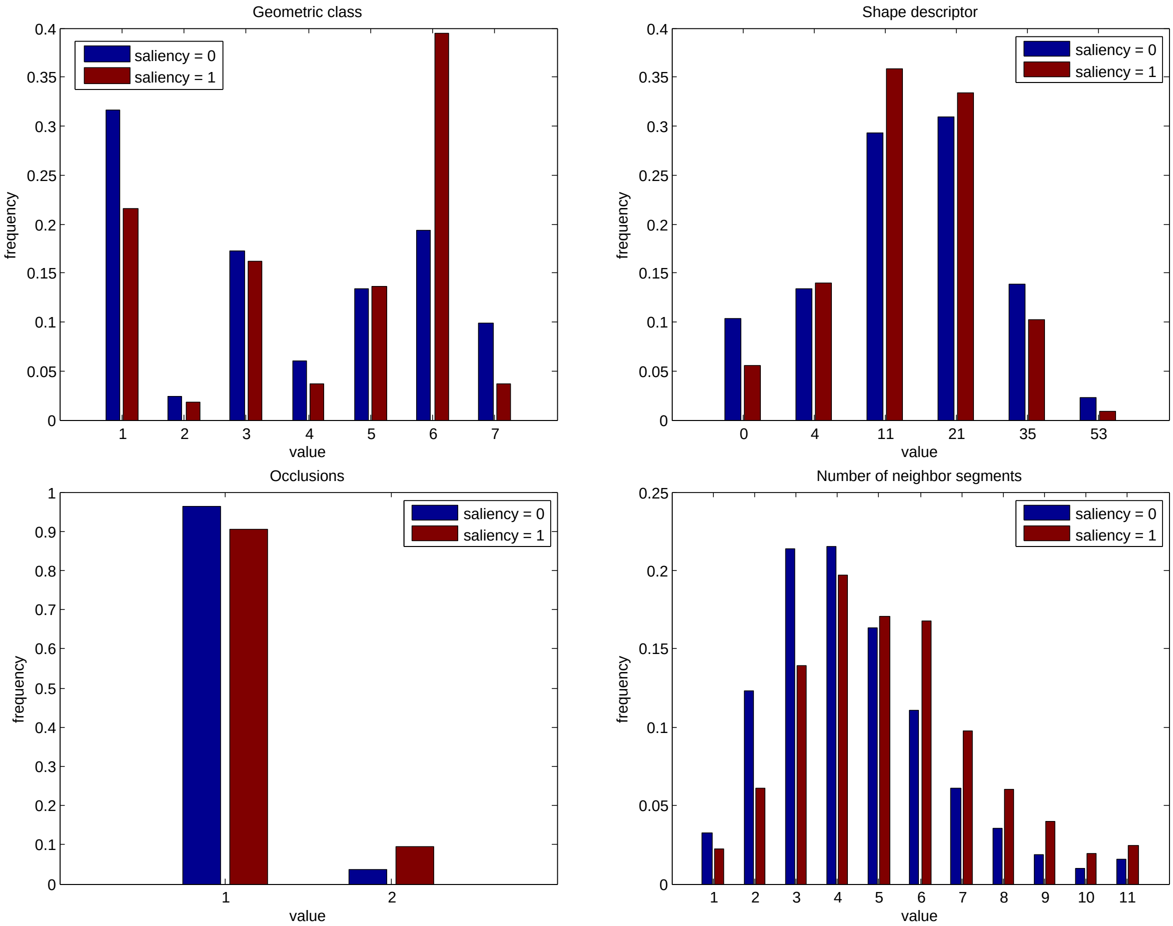


Figure 2.18: Features analysis continue. Here the dependencies of geometric class, shape descriptor, occlusions and neighbors features on saliency are shown.

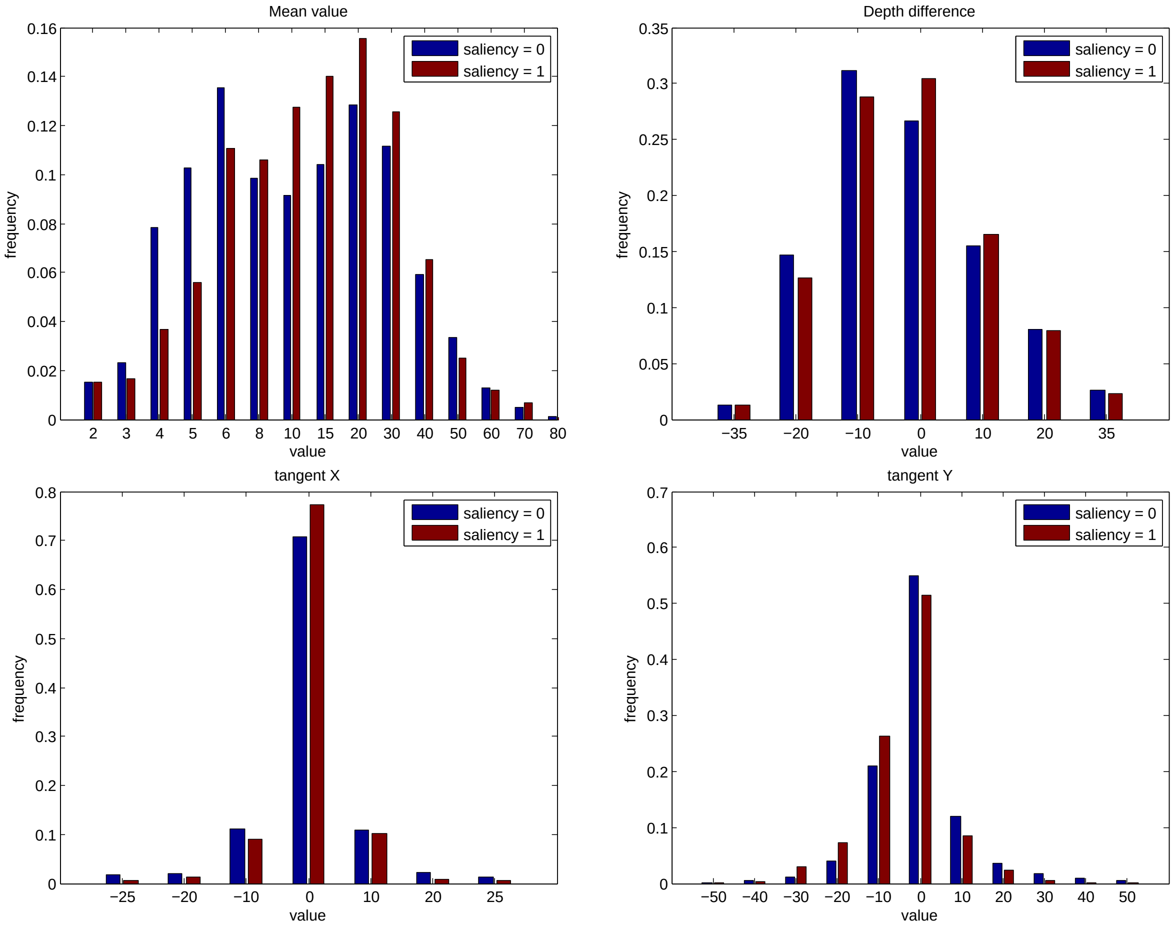


Figure 2.19: Features analysis depth-based method. Here the dependencies of depth mean value, depth contrast, depth difference, tangent X and tangent Y features on saliency are shown.

is possible to substitute this data by estimated depth maps, sample tests have shown dramatic decrease in performance. For this reason no depth methods were included into the comparison. The comparison is shown in Figure 2.20. This qualitative comparison shows that the proposed methods outperform the majority of the concurrent methods. Close performance is demonstrated only by methods [10] and [38] (E and F images in Figure 2.20). The former method is utilizing segmentation. The authors used mean-shift-like segmentation approach that is similar to that used in the methods proposed in this work. For this reason output maps are very close in shape. Unlike the proposed segment-based method the authors of [10] provide binary saliency map. Another method having close performance is described in [38]. Unlike the above mention method it is a pixel-based method. There is no numerical comparison to the concurrent methods due to the problem of matching segment-wise data to pixel-wise data. More comparison image can be found in Appendix A.

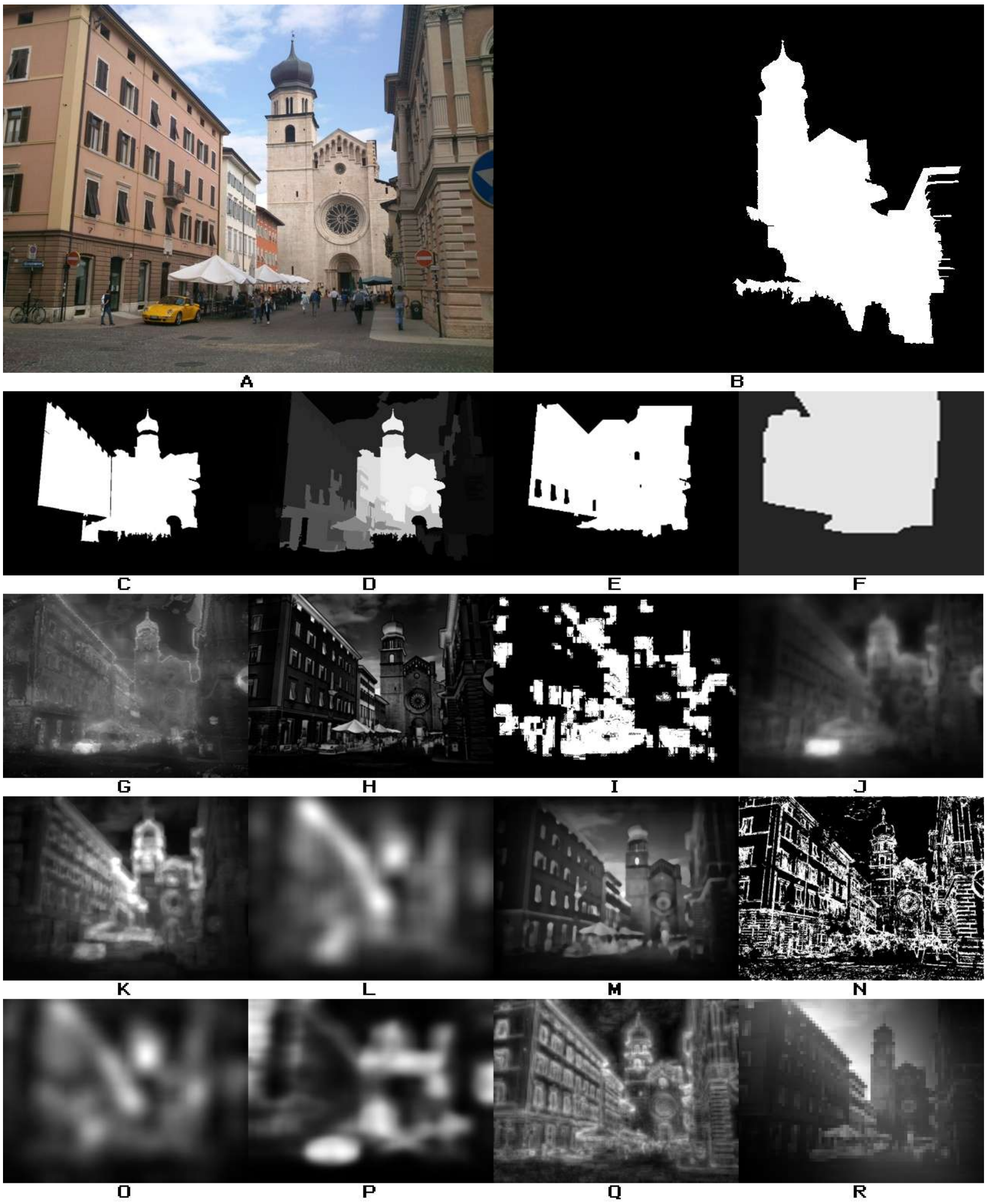


Figure 2.20: Evaluation. A: original image, B: ground-truth map, C: depth-based method, D: segment-based method, E: Cheng et al [10], F: Liu et al [38], G: Judd et al [34], H: Achanta et al [3], I: Harel et al [29], J: Goferman et al [25], K: Margolin et al [43], L: Schauerte et al [51], M: Vikram et al [60], N: Ma et al [41], O: Hou et al [31], P: Seo et al [52], Q: Torralba et al [56], R: Fang et al [18]

Chapter 3

Diversification of visual content using saliency maps

3.1 Introduction

Diversity of contents is an important feature and an added value in the Internet, and in general in all applications characterized by a large amount of information coming from different sources. It is the result of the large variety of situations, contexts, cultural backgrounds, religious and political beliefs, ideologies and time. Thus, to fully exploit the huge and ever increasing amount of information available on the Web, diversity has to be appropriately taken into account as a key instrument to achieve deeper understanding and reliable interpretation of the information and knowledge available.

In the specific domain of media search, diversity is usually associated to a problem of visual diversification. Being based on textual tags associated to images, search engines on the web typically could not offer this kind of diversification, thus retrieving many near duplicates. Instead, users would better appreciate a set of results able to show different aspects of that query; this is especially important when the query is poorly specified or ambiguous [53].

Diversification of results in media search engines is a relatively new area of research. In the last years, several techniques have been proposed to achieve this goal, mainly using textual information or algorithms imported from natural language processing domain. Although image annotation could be an important source of information, quite often it turns out to be quite unreliable. For instance, user generated contents are often unannotated or sparsely annotated, thus making text-based approaches hardly applicable. Additionally, annotations may contain noisy or irrelevant data that in turn could produce irrelevant outputs. In the same way, the degree of results diversification depends on how annotations grasp the content of the image both from visual and semantic points of view.

On the other hand, images contain much more information than their textual descriptions and the use of visual features deserves special attention in this context. In terms of image search, a simple yet effective way to increase diversity is to ensure that duplicates or near-duplicates in the retrieved set are hidden from the user [64]. This approach however works as a posteriori filter on the result, while a mechanism to enforce diversification in the retrieval process would be more impactful. An insight on the most significant approaches so far proposed will be presented in Section 3.2.

When dealing with visual perception of a media object, the concept of saliency is of paramount importance. Visual saliency provides information about the areas of an image perceived as most important and instinctively targeted by humans when shooting a photo or looking at a picture. Intuitively, saliency can play an important role in the framework of diversification, by providing information on what the user perceives as the key subject of an image, thus making it possible to focus the diversification on the most relevant contents. Stated another way, since visual saliency is as an additional dimension of the data implicitly embedded in a picture by its creator, it looks natural to use this information for defining a higher dimensional feature space that allows more accurate description of images, emphasizing both semantic and visual diversity.

In this chapter the usefulness of visual saliency to increase diversity in image retrieval is investigated. I propose a method to re-rank the results of a query based on visual content, to achieve better diversity in top results. Then, it is shown how the introduction of a saliency-based modification of the re-ranking strategy can achieve significantly better performance as compared to the baseline approach. In particular, I propose to use saliency information to stress the importance of certain parts of an image. This will be achieved by using different sets of features to describe important (foreground) and less important (background) parts of an image, and applying different weights in the similarity function. I will demonstrate that this allows achieving better diversification of the main subject of the picture (e.g., different viewpoints, different models of the same object, etc.), or vice versa providing different views of similar objects in different contexts (e.g., different backgrounds).

3.2 State of the art on diversification in image retrieval

The idea of diversification of image retrieval results has been studied recently by many researchers [12]. A good comparison of different methods can be found in [55]. The authors compare 8 diversification methods submitted to ImageCLEF retrieval contest[1] including text-only, hybrid and pure content-based methods. Their study shows that with current technologies hybrid methods outperform text-only and content-based methods. A notable example of a hybrid method was presented in [14]. In their approach, unlike the many methods performing diversification as a

[1] http://www.imageclef.org/

post-processing step, the authors proposed a dynamic-programming-like ranking that jointly optimizes relevance and diversity measures. To this purpose, they use a broad variety of input features that include colour histograms, texture descriptors, bag of visual words, and text data. Another approach with similar characteristics can be found in [63]: here, unlike [14], the authors used visual and textual features separately. Text features are responsible for the relevance by estimating the distance of tags, while visual features are used for diversification by maximizing the distance among candidate images. A pure text-based method is described in [65]. The authors proposed probabilistic model of image tags, with respect to the query that models both relevance and diversity. The main disadvantages of the above methods is in that they rely on the semantic relationship of textual annotation, thus making them hardly applicable in to unreliably annotated data.

An interesting approach dealing with unannotated data has been presented in [54]. The authors addressed the problem of diversification through automatic annotation of images based on their visual features. This text information is then used for creation of a topic graph of the set. Finally, the results are diversified using topic reachness score, so that images with higher score appear at the top of the ranking. In addition, a topic coverage score is proposed, which is a measure defining the diversity of the image set and is based on the number of text-topics presented in the results. Although this method is independent of image annotation, its performance is highly dependent on the performance of annotation prediction method.

Use of clustering techniques as a post retrieval processing step for topic coverage enhancement has been proposed in the work by Van Leuken et al. [59]. The authors performed comparison of several clustering strategies and analysed their effect on relevance and topic coverage. They also proposed a dynamic feature weighting technique that allows better fusion of features.

Clustering is performed using a visual similarity measure based on low-level features and descriptors. Like in most content-based methods, all the content is treated uniformly, without differentiating between important areas and background.

Apart from pure content-based approaches there exist numerous hybrid and text-only approaches for diversification of image retrieval results. A good comparison of different methods for results diversification can be found in [55]. Among other methods they also compared the method they proposed in(bimodel text and image retrieval with diversity enhancement), that can serve both as text-only or content-based only method depending on the availability of text annotation with an image. When working with text annotations diversity is obtained by clustering location information data.where they propose utilization of visual similarity for diversification when no text information is available, using maxmin approach that is maximization of minimum distance of a candidate image with respect to the selected images.

3.3 The proposed approach

In this section I provide a detailed description of the proposed method for image search results diversification based on visual saliency. I start with describing the visual features used for diversification and how they are related to the relevant application. After that, the proposed re-ranking approach is presented.

3.3.1 Visual Features

In order to quantitatively measure the visual dissimilarity among images, it is necessary to define a set of features that efficiently encode the perceptual appearance of visual data. In this work the model relies on low level features that are correlated with human vision system (HVS) characteristics. In particular, each image is described by 6 features, namely: foreground and background color histograms, foreground and background orientation histograms, foreground size and foreground location. In the following details about this description are provided. Foreground and background parts of images in this case are salient and non-salient areas detected by segment-based saliency extraction method described in Section 2.4.

Colors are recognized to be one of the most important perceptual features of images. In particular, color histograms provide a meaningful and convenient representation, accounting for relatively fast processing and easy comparison. Furthermore, the use of color histograms in previous works has shown their good applicability for the task of diversification. Color histograms can be applied to different color spaces and with different chromatic resolutions. Some works propose the use of entire full-color RGB color space, others use alternative color representations such as L*a*b* or HSV, with different numbers of bins. In this work a 9-bin color space based on HSV color representation is used. Three bins stand for different monotone color luminosity values (black, white and grey), while other bins count the occurrences of basic color tones (red, yellow, green, cyan, violet and pink). Input colors are transformed into HSV color space, followed by grey tone classification. This is done by analysing the S and V color components. Pixel's color is considered to be grey if:

$$V < 0.1 \ \vee \ S < 0.1 + \frac{0.01}{V^2}$$

Here, three levels of monotone illumination are defined: black ($V \leq 0.23$), grey ($0.23 < V < 0.85$) and white ($V \geq 0.85$). After that, color classification is performed on pixels that at previous step were not classified as greyscale. The color tone is determined by splitting H color component into 6 equally spaced regions with centers at [0.083 0.25 0.417 0.583 0.75

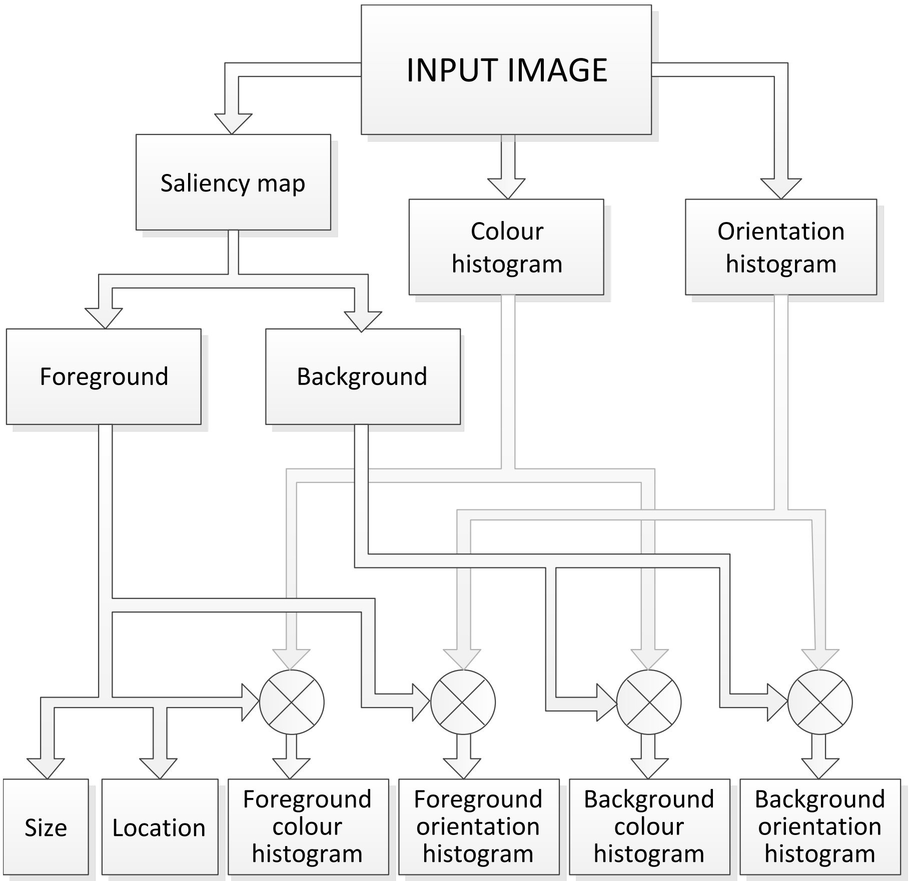

INPUT IMAGE
Saliency map
Colour histogram
Orientation histogram
Foreground
Background
Size
Location
Foreground colour histogram
Foreground orientation histogram
Background colour histogram
Background orientation histogram

Figure 3.1: Feature extraction scheme

0.917], and mapping pixel's color to the closest color region. The use of a limited color description of this type accounts for the fact that slight variations in half tones are hardly detectable by the HVS in the absence of a reference image, and this information is useful only when visually very close images (near duplicates) are compared. On the other hand, the absence or presence of some basic color tone have a great impact on perception. In addition, tests have shown that histograms with large number of bins suffer from a drawback that causes small spectrum shifts to result in relatively high distance values.

Another feature employed in the proposed method are orientation histograms. There are several reasons for using such descriptors. Firstly, they allows a simple yet effective analysis of texture contents. Furthermore, they allow to some extent estimating observation viewpoint for objects that have dominant orientation of straight edges on their bodies. This is the typical case for man-made objects like cars, building, etc. Orientations are detected by applying directional filters at different scales. In this work, the following directions are used for orientation filters: 0°, 30°, 60°, 90°, 120° and 150°. These filters are applied at six scales, and responses at different scales are summed up per each orientation.

Finally, saliency information allows for extraction of object-specific features. For this reason, in addition to color and orientation histograms, foreground size and location features are used. The size is computed by normalizing the area of the foreground by the image size. Location is defined by the vector of the coordinates of the foreground region centroid, normalized over image dimensions. The overall feature extraction scheme is shown in Figure 3.1.

3.3.2 Search Results Diversification

Ranking is the key component of the system. Given a query, in order to find relevant and diversified results, it is necessary to find a suitable trade-off between similarity and diversity of images, which are controversial constrains. Since pure content-based search is still a tough problem [47], and the set of features used is insufficient to achieve a sufficient accuracy of visual search results, it is assumed to have in input a set of images returned by text-based search, providing a set of relevant images (precision=1), thus the task is limited to re-ranking of results in order to achieve a higher diversity on top N results. In principle, the proposed system acts as a post-retrieval filter that sorts the results to increase the diversity.

This given, the major contribution of the proposed work is in the use of saliency to perform this task in a more effective way. There are several possibilities to make use of saliency information. For instance, one can force foreground similarity while differentiating the background, thus resulting in the same object appearing in different contexts. On the contrary, one may differentiate the foreground independently of the background, thus achieving a larger variety of subjects. This way of proceeding however would neglect the strong correlation between foreground and background, which appears evident when analysing the data (see Figure 3.2). Another problem is that frequently occurring images should be promoted to the top places as very rare images are likely to be irrelevant.

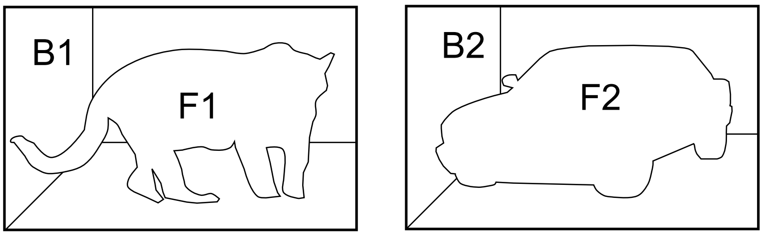


Figure 3.2: Foreground and background correlation example. Consider the left image to be a picture of a mammal in its natural environment and the right one is a picture of a man-made object in an industrial environment. Then $P(F = F1|B = B1) \gg P(F = F1|B = B2)$, likewise $P(F = F2|B = B2) \gg P(F = F2|B = B1)$.

According to the above consideration, I propose a weighting method that jointly considers background and foreground diversity, while at the same time putting frequently occurring images at the top places. Given the feature vectors associated to a pair of images im_1 and im_2, their dissimilarity is computed according to the following Equation:

$$D(im_1, im_2) = \sum_{i=0}^{n} Dist(f_1^i, f_2^i)_i \bullet w_i, \tag{3.1}$$

where $Dist(f_1^i, f_2^i)_i$ represents the distance between image im_1 and im_2 with respect to feature f^i, and w_i is the corresponding weight. Dissimilarity of histogram features is computed using cosine distance. As mentioned above, I assume that the input set consists of relevant images only, their initial order is not used by the ranking. Candidate images are selected by minimizing the score term CM. For the sake of simplicity a linear ranking method was selected.

$$CM(im) = w_{res} \bullet D(im)_{res} - w_{nran} \bullet D(im)_{nran}, \qquad (3.2)$$

where $D(im)_{nran}$ is the overall normalized distance of image im to images in unranked results list, while $D(im)_{res}$ is the overall normalized distance of image im to images in results list. Relevant weights are w_{nran} and w_{res}. As a result, the optimization is done by means of minimization of similarity with images in results list and maximization of similarity with unranked images. Thereby diversity is achieved through promotion of representative images from the unranked list and penalty of similar images in the results list. The re-ranking algorithm is shown in Listing 1.

Algorithm 1 Re-ranking algorithm

```
procedure RE-RANKING($RES, NRAN$)
    while $\|NRAN\| > 0$ do
        for all $im_{cand} \in NRAN$ do
            $D_{NRAN} = \sum_{im_t \in NRAN \ni im_{cand}} \sum_{i=0}^{N} Dist(f_t^i, f_{cand}^i) \bullet w_i$
            $D_{RES} = \sum_{im_t \in RES} \sum_{i=0}^{N} Dist(f_t^i, f_{cand}^i) \bullet w_i$
            $CM(im_{cand}) = w_{RES} \bullet D_{RES} - w_{NRAN} \bullet D_{NRAN}$
        end for
        $im_{max} = im_{cand} \rightarrow max(CM(im_{cand}))$
        $ADD\ im_{max}\ to\ RES$
        $REMOVE\ im_{max}\ from\ NRAN$
    end while
end procedure
```

3.4 Evaluation

Assessment of image retrieval results diversity is an unsolved problem yet. Recently several ways of measuring diversity have been proposed. However, they mostly tackle only the concept aspect of diversity by analysing text-annotations [59] or by counting number of clusters [4] in retrieval results. This results in diversity towards variety of concepts leaving visual diversity out of scope. Since saliency allows for discrimination between foreground and background parts, it allows to achieve object representation diversification. Thereby it is of interest to be able to measure both concept and representation parts of diversity. Thus resulting in a most effective visual diversification. Here, I propose a measure which to some extend allows for evaluation of both semantic and visual diversity. Although recent works proposed several plausible diversity measures for tagged images they are not fully applicable in the scope of this work. For example, for the case of commonly used approach of data clustering, the number of clusters for a category consisting of 100 images can be as high as 70. As a results number of clusters in first 20 ranked images for almost all cases is 20 that makes comparison of different rankings intractable. Other measures that require semantic understanding are possible but require a natural language processing framework. The measure I propose is based on text-based representation of visual content by annotations. Such an annotation consists of list of properties, which are related both to visual and semantic variations of the main object within a given set. Each property consists of a list of tags that define its possible values. To each tag I assign its weight that is computed as follows:

$$w_t = \frac{t_i}{i \bullet p}, \tag{3.3}$$

where t_i is the number of images this tag was assigned to, p is the number of properties for a set of images and i is the total number of images in

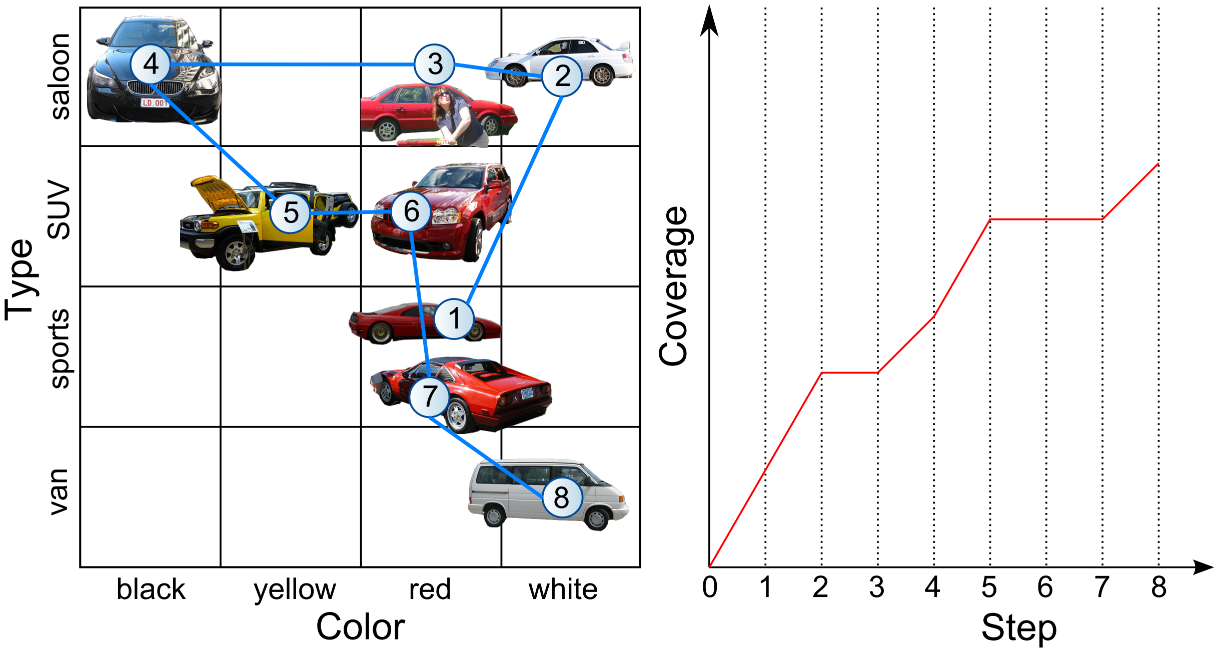


Figure 3.3: Example of coverage measure computation for a category with two properties. It can be noticed that coverage measure increases only when a new value for a property is included to the ranking. For example, at step 4 an image with value black for color property is added resulting in coverage increase. Also note, when at one step several new values appear the increase is higher (e.g., steps 1-2 and 4-5 compare to 2-3 and 7-8).

the initial image set. Then diversity is measured as coverage over tags. Coverage of a set is defined as the sum of weights of unique tags assigned to images in this set. Thereby only weights of newly introduces tags are counted, and the maximum possible value of coverage doesn't exceed 1. In Figure 3.3 I give an example of how coverage is computed. The measure that I proposed allows capturing both of diversity and relevance. It is done by giving higher weight to tags that are assigned to more images. Then the overall score increases when an image, that represents a larger group of images is added compared to an image with a very rare tag.

The method proposed is designed in such a way that it is a post processing step of retrieval and it is applied for results re-ranking. Thereby, for the evaluation of the approach it is necessary to set up a retrieval system in order to get an initial ranking. However, for the sake of simplicity this step can be eliminated if we use a set of images grouped in a some way, e.g. by categories. Indeed, these grouped images can serve as an output of an image retrieval image. From section 3.3 it is clear that the proposed approach does not depend on initial positions of images in the original ranking, thereby from the point of view of the this method the order of images does not play any role.

For the evaluation of the approach a dataset was created based on two image datasets: Caltech 256 [27] and Pascal VOC 2008 [17]. Caltech 256 provides a category-based image sets. There is no any grouping in Pascal VOC 2008 dataset, but all images are provided with text description containing information about the type, bounding box and pose of objects of interest presented on images. Image sets for several categories were created by querying the type field. However, for some images annotated objects were occupying very insignificant area of an image making these objects less visual important and relevant with respect to other objects on the same image. To eliminate this effect images with area of objects of interest less than 10 percent of image's area were excluded. The proposed measure is aimed at capturing visual properties of an image. Since there were created separate sets of annotations for each category, then the dataset should capture variations of the main category object and its surroundings. After analysis of the dataset I came up with an annotation guide that encodes the following properties: colors of the main object, quantity of objects belonging to the main object class, location and size of the object, subtype of the object, viewing angle, distance, etc. These annotations briefly cover visual properties of an image. Although there is plenty of other possible properties one can add, there is always a problem of subjectivity. Thereby properties included were limited to these that are to less extend subjective. In addition, applicable properties are very dependant on the content, thus only relevant properties to a category were included into annotations. It is possible to come up with a fixed set of properties but it would result in few properties that hardly grasp semantic and visual content of categories. It turns out that omitting irrelevant properties is equal to using a large fixed set of properties for all category with tag null for these properties that are irrelevant for a specific category.

An example annotation for category bear is reported in Table 3.1.

Table 3.1: Example annotation for category bear

type	spectacled bear
location	zoo
pose	sitting
quantity	alone
viewpoint	side
zoom	tele

Firstly, for the purpose of testing the idea of using saliency for improving diversity I decided to perform evaluation using ground-truth data. Although there are no ground truth saliency data neither for Caltech 256 nor for Pascal VOC 2008 datasets, the later includes object segmentation ground-truth data, that to some extend is close to maps that the saliency detection method provides. So at this step there was made an assumption that these labellings correspond to main object of an image that for most images were true. During comparison only first 20 re-ranked images were counted. Mostly this number was selected due to the fact that usually a close number of images is shown to a user per page. In addition, this number is perfect for illustrating difference in diversity - with more images coverage difference tends to zero, while a fewer images are practically out of interest. Finally, in many other research papers on diversity first 20 images are used for evaluation. In Table 3.2 comparison of re-ranking method using automatically extracted saliency maps, ground-truth data and entire image area is reported. As it can be seen, inclusion of saliency data (no matter if it is extracted automatically or provided as a ground-truth data) improves diversity. In addition, close performance of automatically extracted maps with ground-truth data shows that detection tool gives reasonable maps and its accuracy is sufficient for such a task.

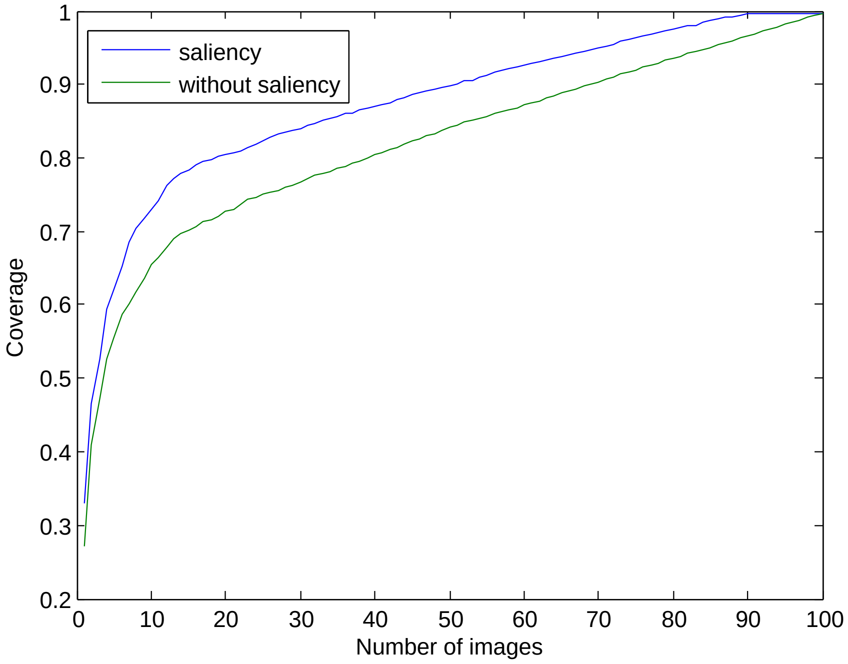


Figure 3.4: Coverage measure pace. As it can be seen ranking with saliency allows for faster coverage of a dataset. Also one can notice that after 15 images coverage grows much slower.

While the first test proves the approach to be working the amount of numerical data is insufficient since it is based on comparison of only five categories. In addition, due to the fact that ground-truth labelled data provided only for a limited number of images initial image sets for categories consists of only few dozens images making it a more easy task to cover all variations of initial set. For this purpose another test was performed. It is different in that it covers more categories and gives more information about generalization of our approach. At this time I compared performance of the

Table 3.2: Experiment 1 results. Coverage measure considering first 20 images. As it can be noticed ranking using labelled data out performs the one without using saliency. This proves that the idea of using visual saliency information results in higher diversity. Comparison of labelled data performance with saliency shows that their performance is close to each other and slight difference can be explained by the fact that not always labelling belongs to visually salient object, also saliency maps estimation may have some errors due to accuracy is not 100%.

category	without saliency	labelled data	saliency
bird	0.972	0.970	0.980
car	0.939	0.944	0.959
cow	0.970	0.985	0.983
boat	0.846	0.893	0.893
sheep	0.937	0.950	0.939

re-ranking method with saliency and without it. Results of this comparison are reported in Table 3.3. The comparison was performed on twenty categories taken both from Caltech 256 and Pascal VOC 2008 datasets. As it can be seen in Table 3.3 use of saliency information results in increase in diversity of 11%. Figure 3.6 shows visual comparison of rankings for category bear. Additional comparison images can be found in B.

The evaluation results of the first and seconds experiments depend on human-labellings. To show the advantages of using saliency without allowing for subjectivity of the results an objective way to create annotations was found. Although there exist a number of methods for automatic estimation of annotation informations for images, such methods does not provide enough details for judging the diversity of a set of images. However, there are special cases. For example, there are tools that are able to perform face recognition and provide information such as sex, edge, mood, viewpoint, etc. Although this is a very specific case, it allows for human-less annotation of a set of images. The Caltech database includes face category which was used for the evaluation. The annotation was done using public API to

Table 3.3: Experiment 2 results. Coverage measure considering first 20 images. (P) and (C) denotes categories taken from Pascal VOC 2008 and Caltech 256 datasets respectfully.

category	saliency	without saliency
bird (P)	0.912	0.823
car (P)	0.814	0.739
cow (P)	0.916	0.801
table (P)	0.840	0.753
bicycle (P)	0.861	0.842
boat (P)	0.683	0.635
sheep (P)	0.817	0.775
pyramid (C)	0.668	0.579
bridge (C)	0.712	0.521
bear (C)	0.707	0.663
blimp (C)	0.863	0.729
butterfly (C)	0.713	0.685
gas-pump (C)	0.477	0.454
teapot (C)	0.615	0.586
wolf (C)	0.924	0.798
sea animal (C)	0.943	0.869
fox (C)	0.915	0.832
military vehicle (C)	0.835	0.668
train (C)	0.995	0.942
airplane (C)	0.895	0.846

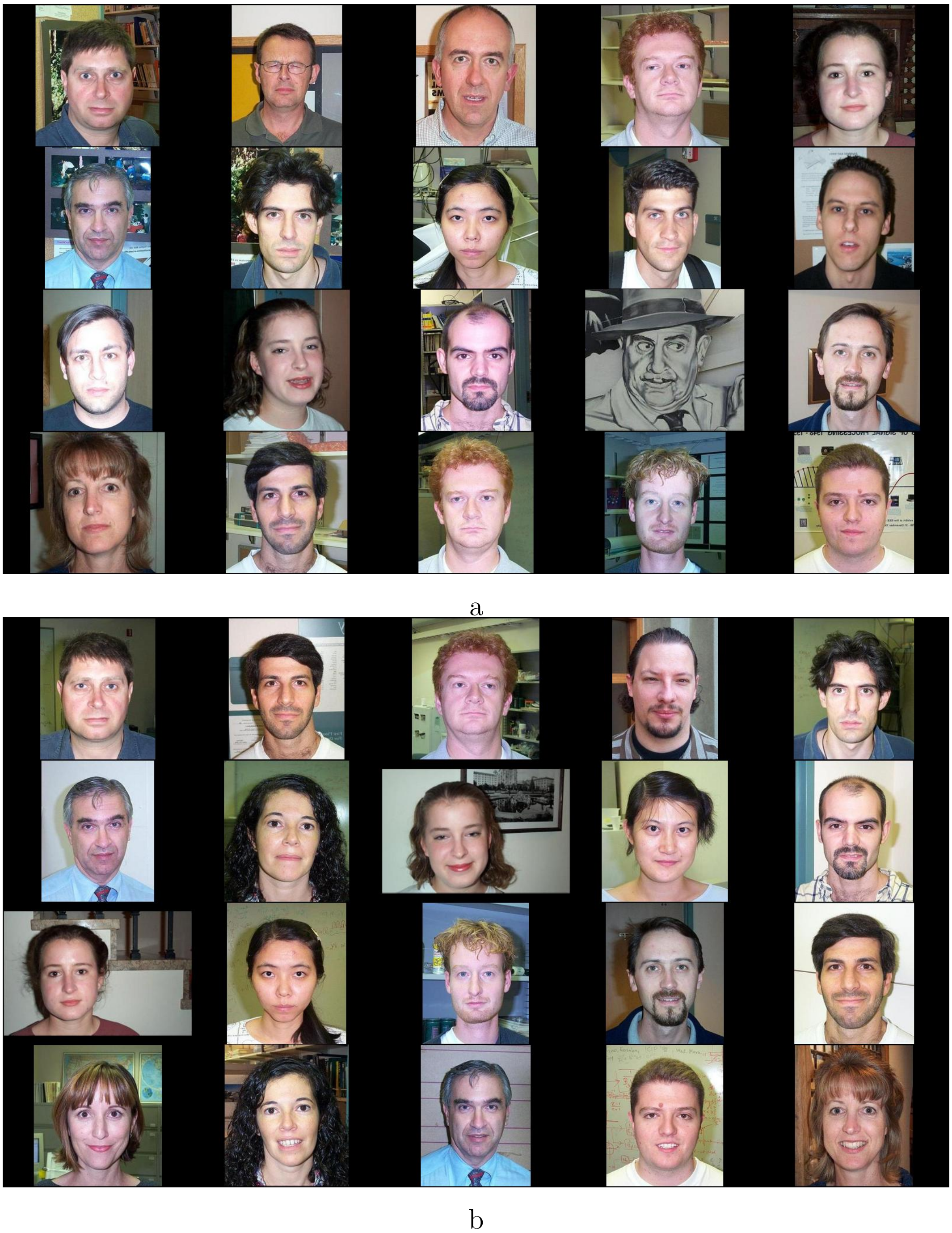
Table 3.4: Example of face annotation

person Id	27
gender	female
age	20
glasses	false
smiling	false
mood	surprised
lips	sealed
yaw	-1
pitch	-13
roll	-2
face width	50
face height	44
face x position	51
face y position	56

face.com service[2]. An example of an annotation produced by this service is reported in Table 3.4. Same as in the aforesaid experiments diversity is determined via coverage measure. The obtained results are: 0.56, 0.39 for ranking with and without saliency respectfully.

The proposed approach depends on 8 weights for the method with the use of saliency (6 weights for features and 2 for ranking) and 4 weights for the method without the use of saliency (2 weights for features and 2 for ranking). In the experiments these weights were obtained empirically by running genetic algorithm optimization on the entire dataset. During experiments the weights obtained were: [7.6 1.8 2.0 0.6 4.0 0.7] [1 1] for the method with the use of saliency information and [3.7 5.7] and [1 0.7] for the method without saliency.

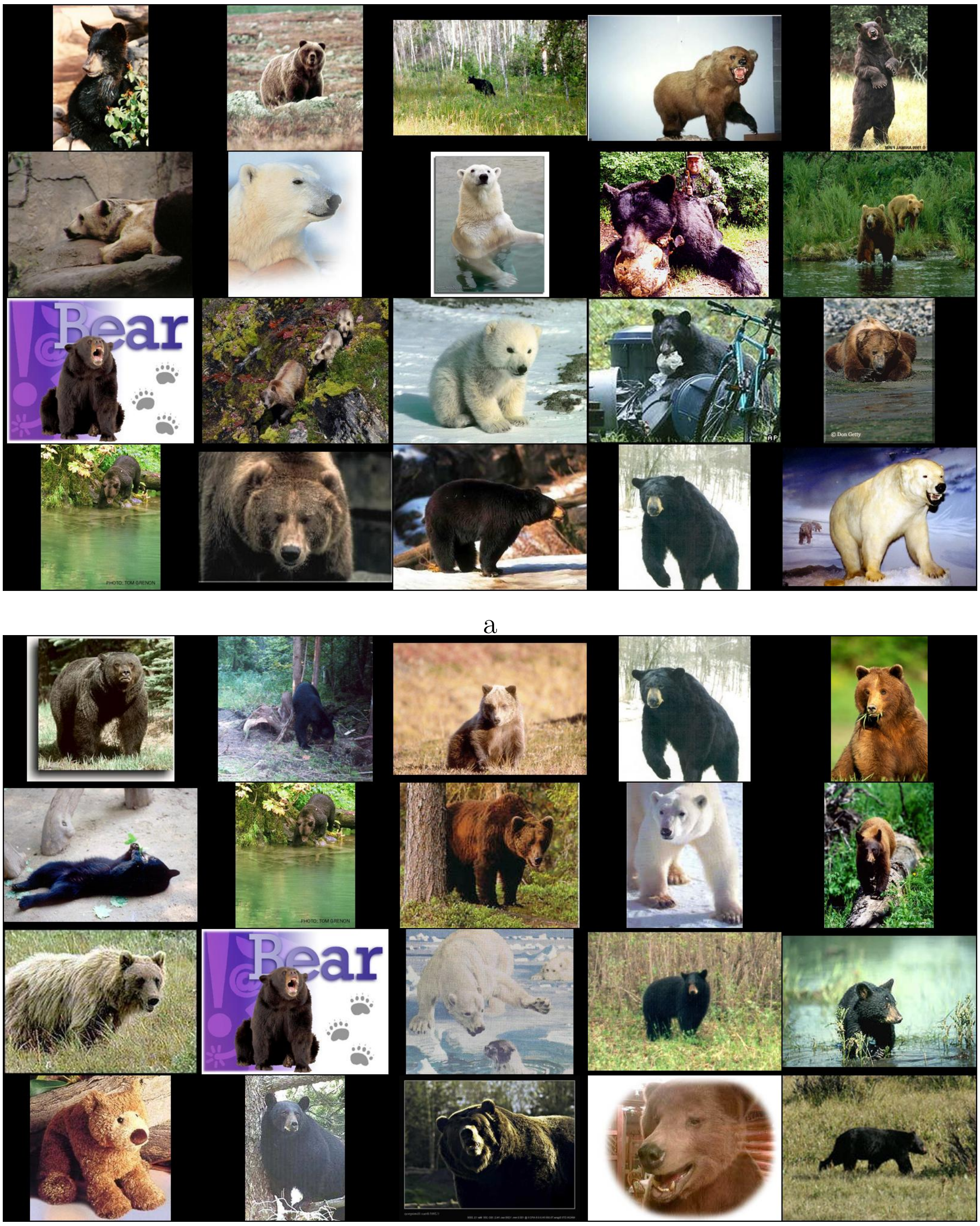
[2] http://www.face.com/ public API is discontinued since August 2012



a

b

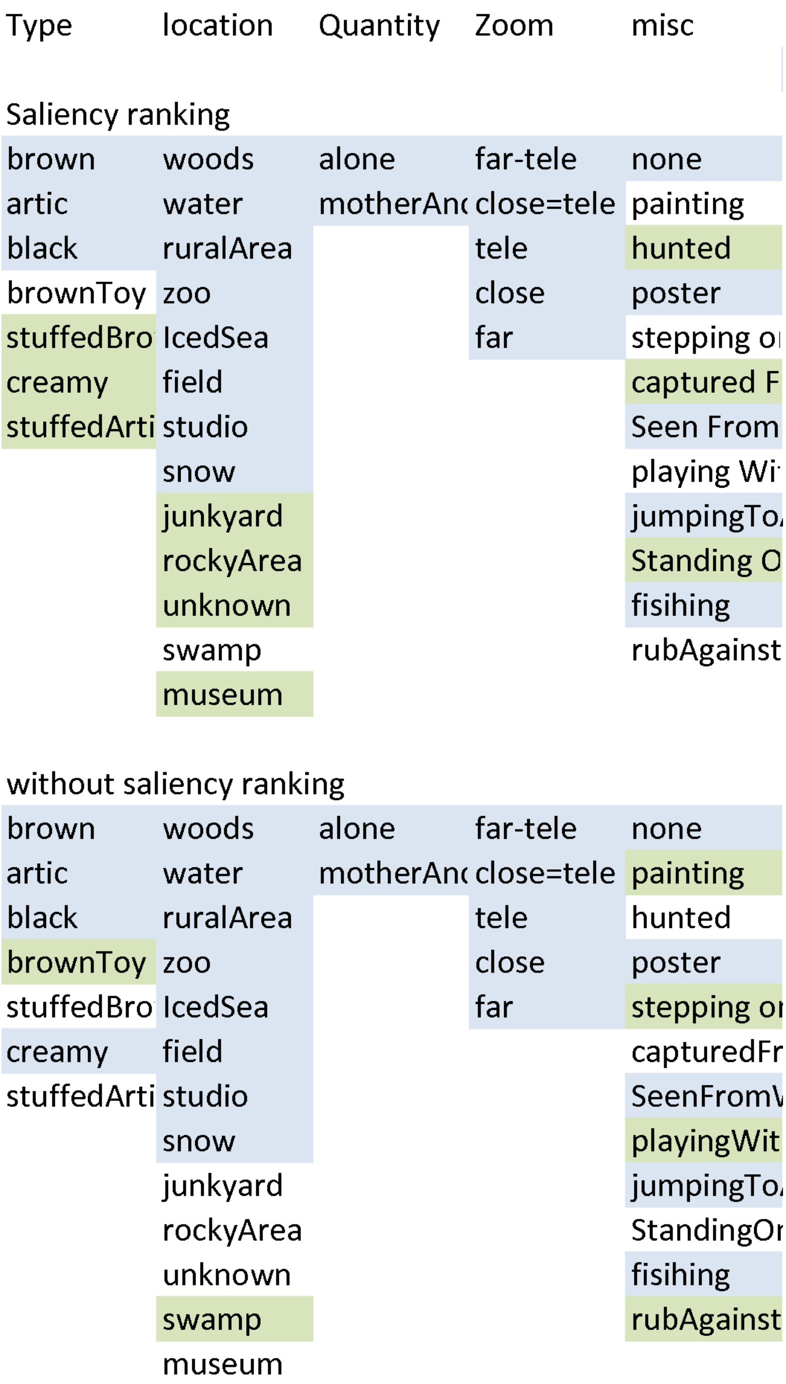
Figure 3.5: Results of re-ranking for faces category from Caltech database. (a) – diversification using saliency, (b) – diversification without saliency. Notice (a) contains only one pair of pictures of the same person compared to two pairs for (b). Also (a) contains: a portrait of a person wearing spectacles and drawing.



a

b

Figure 3.6: Example of re-ranking for category bear using the method with saliency information (a) compared to the method without saliency information (b). Ranking (a) provides more instances of white bears, age variation is higher (notice images of offspring), more locations are captured and there are pictures of a group of bears.



Type	location	Quantity	Zoom	misc
Saliency ranking				
brown	woods	alone	far-tele	none
artic	water	motherAn	close=tele	painting
black	ruralArea		tele	hunted
brownToy	zoo		close	poster
stuffedBro	IcedSea		far	stepping o
creamy	field			captured F
stuffedArti	studio			Seen From
	snow			playing Wi
	junkyard			jumpingTo
	rockyArea			Standing O
	unknown			fisihing
	swamp			rubAgainst
	museum			
without saliency ranking				
brown	woods	alone	far-tele	none
artic	water	motherAn	close=tele	painting
black	ruralArea		tele	hunted
brownToy	zoo		close	poster
stuffedBro	IcedSea		far	stepping o
creamy	field			capturedFr
stuffedArti	studio			SeenFrom
	snow			playingWit
	junkyard			jumpingTo
	rockyArea			StandingO
	unknown			fisihing
	swamp			rubAgainst
	museum			

Figure 3.7: Example of re-ranking for category bear: annotations. Here only annotation values present in first 20 images are shown. Blue color denotes that a particular value of property is present in both rankings. Green color shows values unique for a ranking.

Chapter 4

Conclusion

In this work saliency detection and its use in multimedia applications were studied. Two novel methods of salient region detection were proposed. The initial idea of region-based processing has been proven to be working. Another novel idea of exploiting depth-spatial relation of objects has also shown to be a possible solution to saliency detection. The evaluation and comparison to state-of-the-art methods has shown high performance of the proposed approaches. In addition, an application-based evaluation has been done through finding a new niche for saliency detection. A novel approach to diversification of visual content based on saliency detection has been presented. Its evaluation has shown the rightfulness of this idea and became a good test-bench for proposed saliency detection algorithm.

Although the initial goals of the work were reached there is a room for further development. The proposed models of saliency detection are to large extend based on low-level properties of objects found in images. However, studies on human visual attention has shown that there are two stimulus driving our attention bottom-up and top-down. Addressing top-down driven attention via exploiting semantic properties and relations of objects in images may lead to much higher performance. Though this research is possible only through solving tough tasks from natural language processing, semantics understanding and object categorization domains.

Bibliography

[1] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," *Computer Vision Systems*, pp. 66–75, 2008.

[2] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.* IEEE, 2009, pp. 1597–1604.

[3] R. Achanta and S. Susstrunk, "Saliency Detection using Maximum Symmetric Surround," in *Proceedings of IEEE International Conference on Image Processing*, ser. IEEE International Conference on Image Processing ICIP. Ieee Service Center, 445 Hoes Lane, Po Box 1331, Piscataway, Nj 08855-1331 Usa, 2010.

[4] T. Arni, J. Tang, M. Sanderson, and P. Clough, "Creating a test collection to evaluation diversity in image retrieval," in *The Proceedings of the Workshop on Beyond Binary Relevance: Preferences, Diversity and Set-Level Judgments*, 2008.

[5] T. Avraham and M. Lindenbaum, "Esaliency (extended saliency): Meaningful attention using stochastic image modeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 4, pp. 693–708, 2010.

[6] J. Blanc-Talon, D. Bone, W. Philips, D. C. Popescu, and P. Scheunders, Eds., *Advanced Concepts for Intelligent Vision Systems - 12th International Conference, ACIVS 2010, Sydney, Australia, December 13-16, 2010, Proceedings, Part I*, ser. Lecture Notes in Computer Science, vol. 6474. Springer, 2010.

[7] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2004.60

[8] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 1. IEEE, 2001, pp. 105–112.

[9] N. Bruce and J. Tsotsos, "Saliency based on information maximization," *Advances in neural information processing systems*, vol. 18, p. 155, 2006.

[10] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *IEEE CVPR*, 2011, pp. 409–416.

[11] C. M. Christoudias, B. Georgescu, and P. Meer, "Synergism in low level vision," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 4. IEEE, 2002, pp. 150–155.

[12] C. L. Clarke, N. Craswell, I. Soboroff, and A. Ashkan, "A comparative analysis of cascade measures for novelty and diversity," in *Proceedings of the fourth ACM international conference on Web search and data mining.* ACM, 2011, pp. 75–84.

[13] D. DeCarlo and A. Santella, “Stylization and abstraction of photographs,” in *ACM Transactions on Graphics (TOG)*, vol. 21, no. 3. ACM, 2002, pp. 769–776.

[14] T. Deselaers, T. Gass, P. Dreuw, and H. Ney, “Jointly optimising relevance and diversity in image retrieval,” in *Proceedings of the ACM International Conference on Image and Video Retrieval*, ser. CIVR ’09. New York, NY, USA: ACM, 2009, pp. 39:1–39:8. [Online]. Available: http://doi.acm.org/10.1145/1646396.1646443

[15] B. Deville, G. Bologna, M. Vinckenbosch, T. Pun *et al.*, “Guiding the focus of attention of blind people with visual saliency,” in *Workshop on Computer Vision Applications for the Visually Impaired*, 2008.

[16] N. Dhavale and L. Itti, “Saliency-based multifoveated mpeg compression,” in *Signal processing and its applications, 2003. Proceedings. Seventh international symposium on*, vol. 1. IEEE, 2003, pp. 229–232.

[17] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” pp. 303–338, 2010.

[18] Y. Fang, W. Lin, B.-S. Lee, C.-T. Lau, Z. Chen, and C.-W. Lin, “Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum,” *Trans. Multi.*, vol. 14, no. 1, pp. 187–198, Feb. 2012. [Online]. Available: http://dx.doi.org/10.1109/TMM.2011.2169775

[19] S. Feng, C. Lang, and D. Xu, “Localized content-based image retrieval using saliency-based graph learning framework,” in *Signal Processing (ICSP), 2010 IEEE 10th International Conference on.* IEEE, 2010, pp. 1029–1032.

[20] S. Frintrop, E. Rome, A. Nchter, and H. Surmann, "A bimodal laser-based attention system," *Computer Vision and Image Understanding*, vol. 100, p. 124151, 2005.

[21] Y. Fu, J. Cheng, Z. Li, and H. Lu, "Saliency cuts: An automatic approach to object segmentation," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on.* IEEE, 2008, pp. 1–4.

[22] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 6, pp. 989–1005, 2009.

[23] K. Gao, S. Lin, Y. Zhang, S. Tang, and H. Ren, "Attention model based sift keypoints filtration for image retrieval," in *Computer and Information Science, 2008. ICIS 08. Seventh IEEE/ACIS International Conference on.* IEEE, 2008, pp. 191–196.

[24] E. D. Gelasca, D. Tomasic, and T. Ebrahimi, "Which colors best catch your eyes: a subjective study of color saliency," in *First International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, Arizona, USA*, 2005.

[25] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-Aware Saliency Detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. [Online]. Available: http://www.ee.technion.ac.il/˜ayellet/Ps/10-Saliency.pdf

[26] V. Gopalakrishnan, Y. Hu, and D. Rajan, "Salient region detection by modeling distributions of color and orientation," *Multimedia, IEEE Transactions on*, vol. 11, no. 5, pp. 892–905, 2009.

[27] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," Tech. Rep., 2007.

[28] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *Image Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 185–198, 2010.

[29] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, B. Schlkopf, J. C. Platt, and T. Hoffman, Eds. MIT Press, 2006, pp. 545–552.

[30] D. Hoiem, A. A. Efros, and M. Hebert, "Geometric context from a single image," vol. 1, pp. 654–661, 2005.

[31] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR07). IEEE Computer Society*, 2007, pp. 1–8.

[32] ——, "Thumbnail generation based on global saliency," *Advances in Cognitive Neurodynamics ICCN 2007*, pp. 999–1003, 2008.

[33] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 11, pp. 1254–1259, 1998.

[34] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[35] C. Koch and S. Ullman, “Shifts in selective visual attention: towards the underlying neural circuitry.” *Hum Neurobiol*, vol. 4, no. 4, pp. 219–27, 1985.

[36] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, “A coherent computational approach to model bottom-up visual attention,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 5, pp. 802–817, 2006.

[37] H. Liu, X. Qiu, Q. Huang, S. Jiang, and C. Xu, “Advertise gently-in-image advertising with low intrusiveness,” in *Image Processing (ICIP), 2009 16th IEEE International Conference on.* IEEE, 2009, pp. 3105–3108.

[38] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, “Learning to detect a salient object,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 2, pp. 353–367, 2011.

[39] W. Liu, W. Xu, and L. Li, “A tentative study of visual attention-based salient features for image retrieval,” in *Intelligent Control and Automation, 2008. WCICA 2008. 7th World Congress on.* IEEE, 2008, pp. 7635–7639.

[40] Z. Liu, H. Yan, L. Shen, K. N. Ngan, and Z. Zhang, “Adaptive image retargeting using saliency-based continuous seam carving,” *Optical Engineering*, vol. 49, no. 1, pp. 017 002–017 002, 2010.

[41] L. Ma, J. Tian, and W. Yu, “Visual saliency detection in image using ant colony optimisation and local phase coherence,” *Electronics Letters*, vol. 46, no. 15, pp. 1066 –1068, 22 2010.

[42] L. Marchesotti, C. Cifarelli, and G. Csurka, “A framework for visual saliency detection with applications to image thumbnailing,” in *Com-*

puter Vision, 2009 IEEE 12th International Conference on. IEEE, 2009, pp. 2232–2239.

[43] R. Margolin, L. Zelnik-Manor, and A. Tal, "Saliency for image manipulation," *The Visual Computer*, pp. 1–12, 2012. [Online]. Available: http://dx.doi.org/10.1007/s00371-012-0740-x

[44] P. Mehrani and O. Veksler, "Saliency segmentation based on learning and graph cut refinement," in *Proc. British Machine Vision Conference*, 2010.

[45] O. Muratov, D.-T. Dang-Nguyen, G. Boato, and F. G. De Natale, "Saliency detection as a support for image forensics," in *Communications Control and Signal Processing (ISCCSP), 2012 5th International Symposium on.* IEEE, 2012, pp. 1–5.

[46] N. Ouerhani and H. Hugli, "Computing visual attention from scene depth," in *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, vol. 1. IEEE, 2000, pp. 375–378.

[47] M. L. Paramita, M. Sanderson, and P. Clough, "Diversity in photo retrieval: Overview of the imageclefphoto task 2009," in *CLEF (2)*, 2009, pp. 45–59.

[48] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2. IEEE, 2004, pp. II–37.

[49] C. Sawyer, "Evaluating context-aware saliency detection method," Santa Barbara City College, Tech. Rep., 2012.

[50] A. Saxena, S. H. Chung, and A. Y. Ng, "3-d depth reconstruction from a single still image," *International Journal of Computer Vision (IJCV*, vol. 76, p. 2007, 2007.

[51] B. Schauerte and R. Stiefelhagen, "Predicting human gaze using quaternion dct image signature saliency and face detection," in *Applications of Computer Vision (WACV), 2012 IEEE Workshop on.* IEEE, 2012, pp. 137–144.

[52] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of Vision*, vol. 9, no. 12, 2009. [Online]. Available: http://www.journalofvision.org/content/9/12/15.abstract

[53] K. Song, Y. Tian, W. Gao, and T. Huang, "Diversifying the image retrieval results," in *Proceedings of the 14th annual ACM international conference on Multimedia.* ACM, 2006, pp. 707–710.

[54] K. Song, Y. Tian, and T. Huang, "Improving the image retrieval results via topic coverage graph," *Advances in Multimedia Information Processing*, 2006.

[55] S. Tollari, P. Mulhem, M. Ferecatu, H. Glotin, M. Detyniecki, P. Gallinari, H. Sahbi, and Z.-Q. Zhao, "A comparative study of diversity methods for hybrid text and image retrieval approaches," in *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, ser. LNCS, C. Peters, D. Giampiccolo, N. Ferro, V. Petras, J. Gonzalo, A. Penas, T. Deselaers, T. Mandl, G. J. F. Jones, and M. Kurimo, Eds., vol. 5706. Springer Berlin / Heidelberg, 2009, pp. 585–592.

[56] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes:

the role of global features in object search." *Psychological review*, vol. 113, no. 4, p. 766, 2006.

[57] R. Valenti, N. Sebe, and T. Gevers, "Image saliency by isocentric curvedness and color," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 2185–2192.

[58] I. Van Der Linde, U. Rajashekar, A. C. Bovik, and L. K. Cormack, "Doves: a database of visual eye movements," *Spatial vision*, vol. 22, no. 2, pp. 161–177, 2009.

[59] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol, "Visual diversification of image search results," in *Proceedings of the 18th international conference on World wide web*, ser. WWW '09. New York, NY, USA: ACM, 2009, pp. 341–350.

[60] T. N. Vikram, M. Tscherepanow, and B. Wrede, "A saliency map based on sampling an image into random rectangular regions of interest," *Pattern Recognition*, vol. 45, no. 9, pp. 3114 – 3124, 2012, ¡ce:title¿Best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011)¡/ce:title¿. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320312000714

[61] S. Wan, P. Jin, and L. Yue, "An approach for image retrieval based on visual saliency," in *Image Analysis and Signal Processing, 2009. IASP 2009. International Conference on*. IEEE, 2009, pp. 172–175.

[62] B. Wang, X. Zhang, M. Wang, and P. Zhao, "Saliency distinguishing and applications to semantics extraction and retrieval of natural image," in *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on*, vol. 2. IEEE, 2010, pp. 802–807.

[63] M. Wang, K. Yang, X.-S. Hua, and H. Zhang, "Towards a relevant and diverse search of social images," *IEEE Transactions on Multimedia*, vol. 12, no. 8, pp. 829–842, 2010.

[64] P. Zontone, G. Boato, F. De Natale, A. De Rosa, M. Barni, A. Piva, J. Hare, D. Dupplaw, and P. Lewis, "Image diversity analysis: Context, opinion and bias." CEUR-WS, 2009.

[65] R. V. Zwol, V. Murdock, L. G. Pueyo, and G. Ramirez, "G.: Diversifying image search with user generated content," in *In: Proc. MIR 08: Proceeding of the 1st ACM international conference on Multimedia information retrieval*, 2008, pp. 67–74.

List of publications

O. Muratov, P. Zontone, G. Boato, F. G. B. De Natale. A segment-based image saliency detection. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011.
O. Muratov, D. T. Dang-Nguyen, G. Boato, F. G. De Natale. Saliency detection as a support for image forensics. In 5th International Symposium on Communications Control and Signal Processing (ISCCSP), 2012.
O. Muratov, G. Boato, F.G. De Natale. Diversification of visual media retrieval results using saliency detection. In Electronic Imaging Science and Technology (IS&T SPIE), 2013.

Appendix A

Saliency detection evaluation

Here additional comparison images are presented. Each evaluation figure consists of 18 images composed as: .

A: original image

B: ground-truth map

C: depth-based method (see Section 2.5)

D: segment-based method (see Section 2.4)

E: Cheng et al [10]

F: Liu et al [38]

G: Judd et al [34]

H: Achanta et al [3]

I: Harel et al [29]

J: Goferman et al [25]

K: Margolin et al [43]

L: Schauerte et al [51]

M: Vikram et al [60]

N: Ma et al [41]

O: Hou et al [31]

P: Seo et al [52]

Q: Torralba et al [56]

R: Fang et al [18]

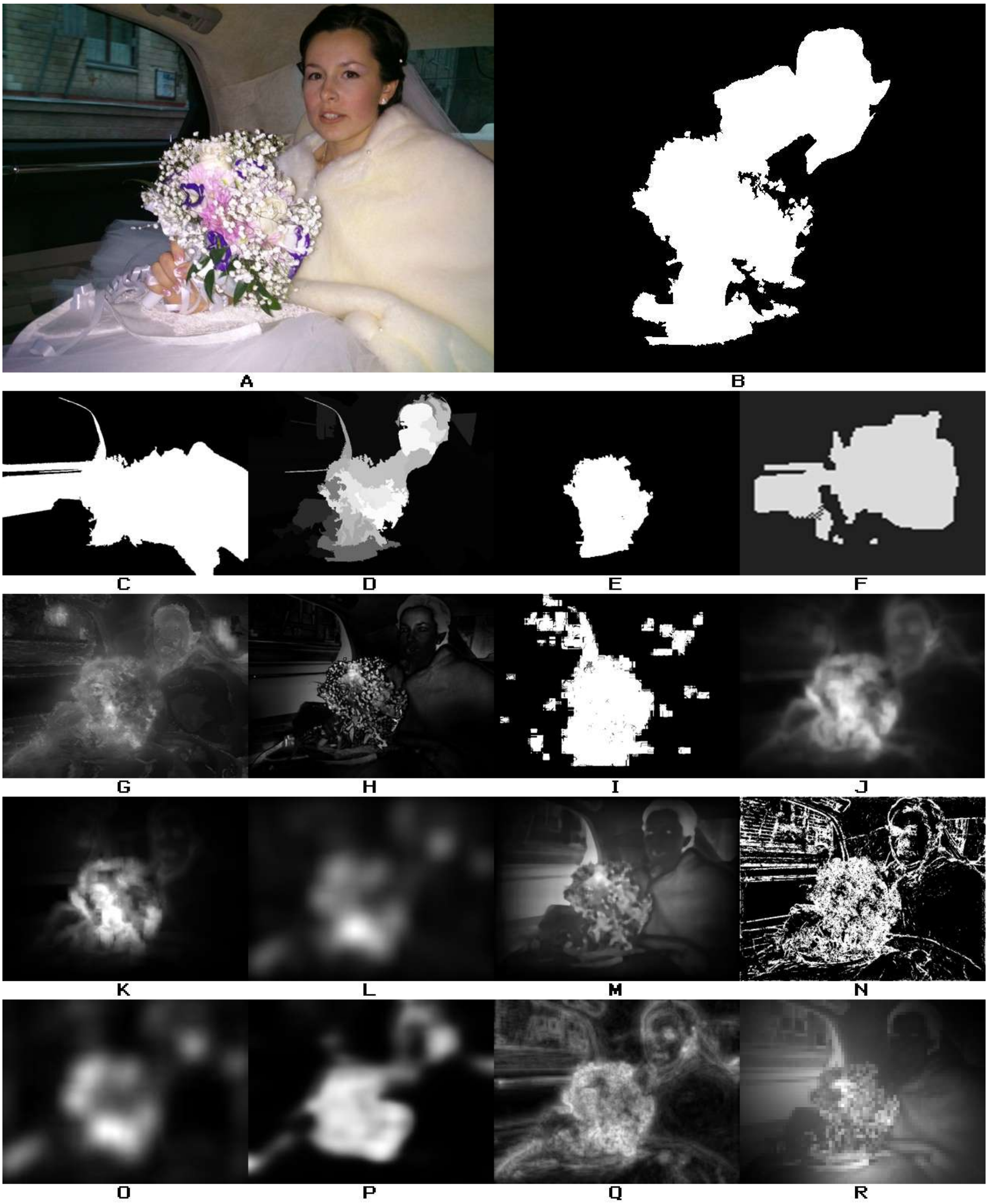


Figure A.1: Test image: marriage

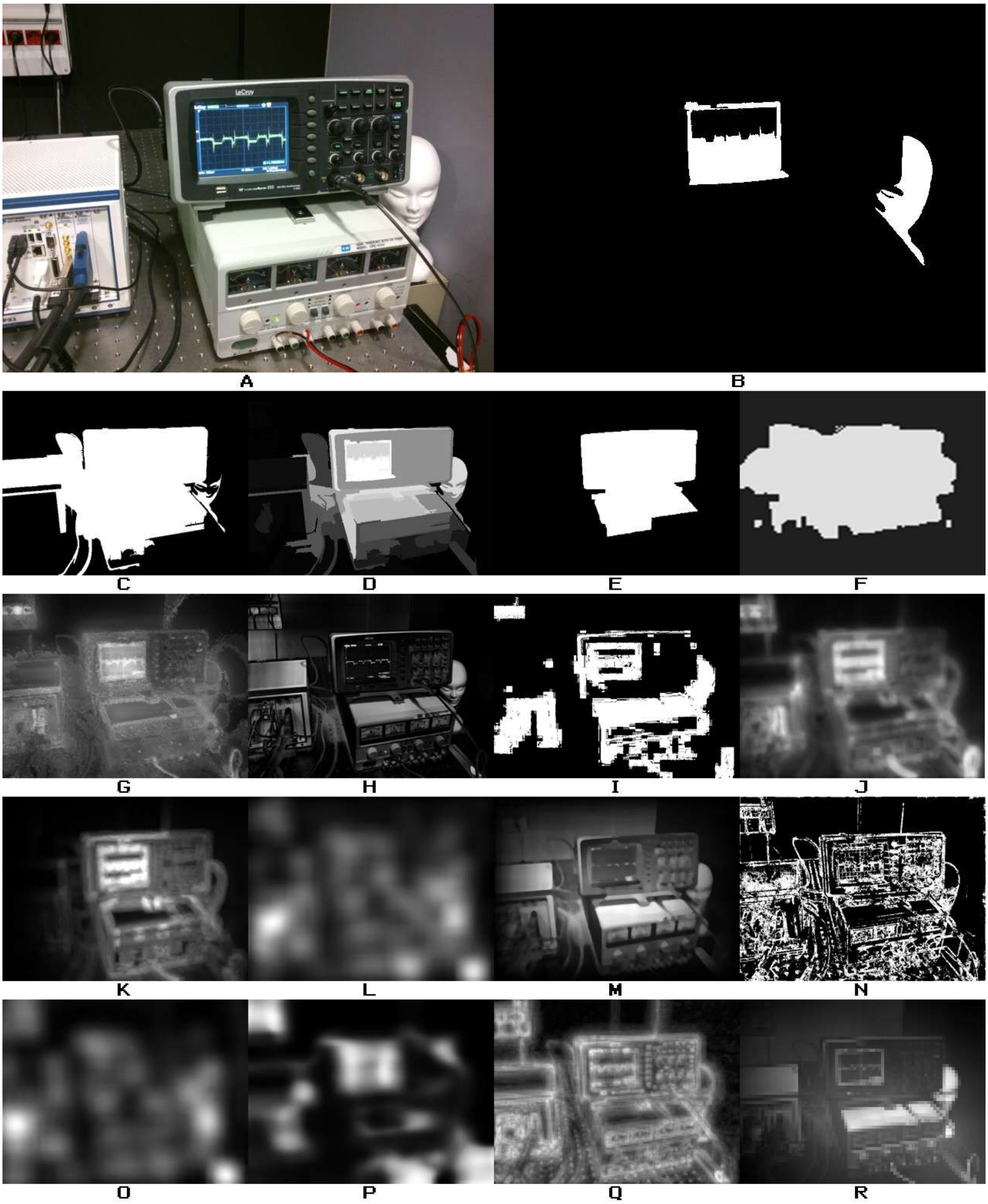


Figure A.2: Test image: laboratory

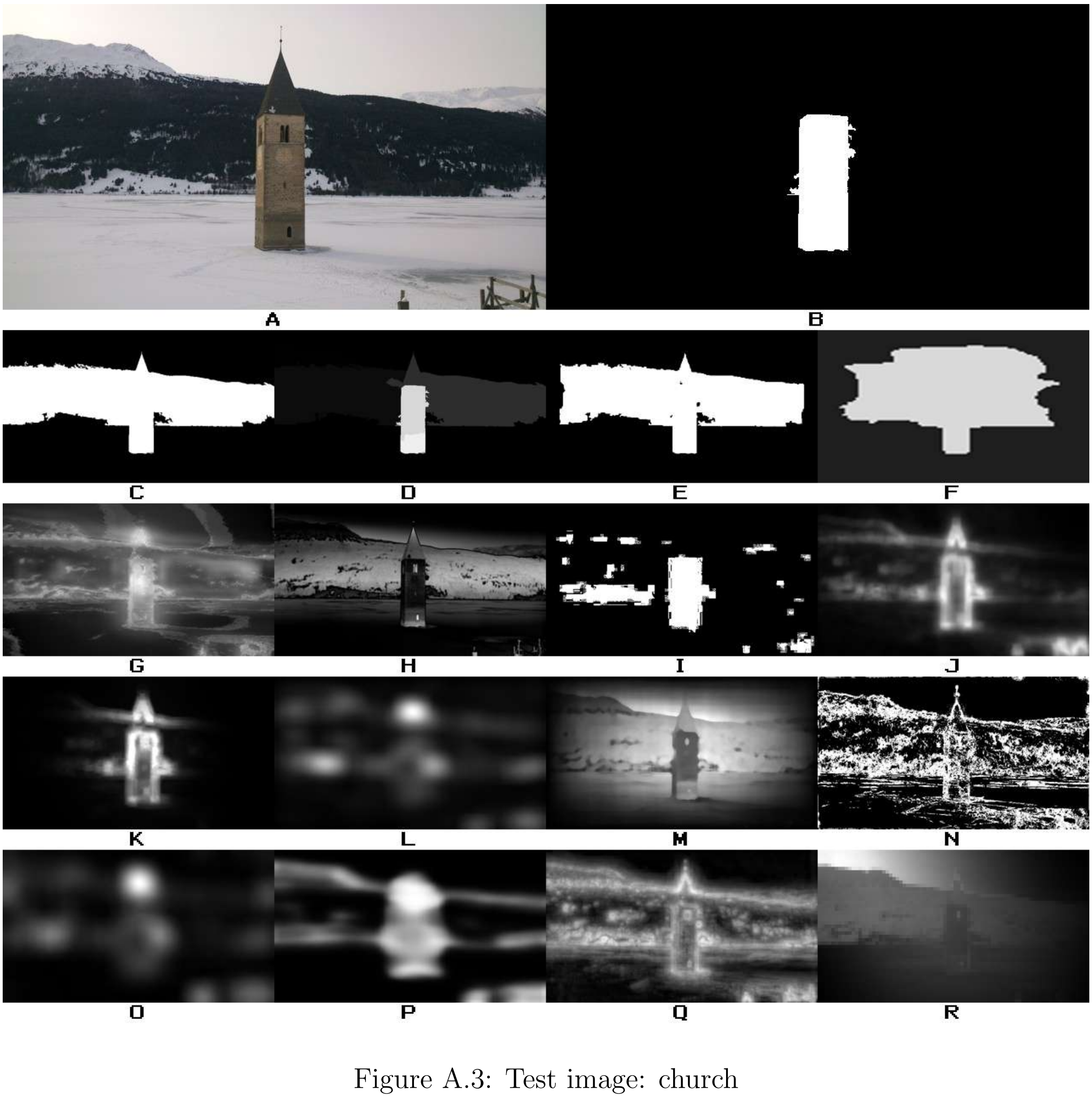


Figure A.3: Test image: church

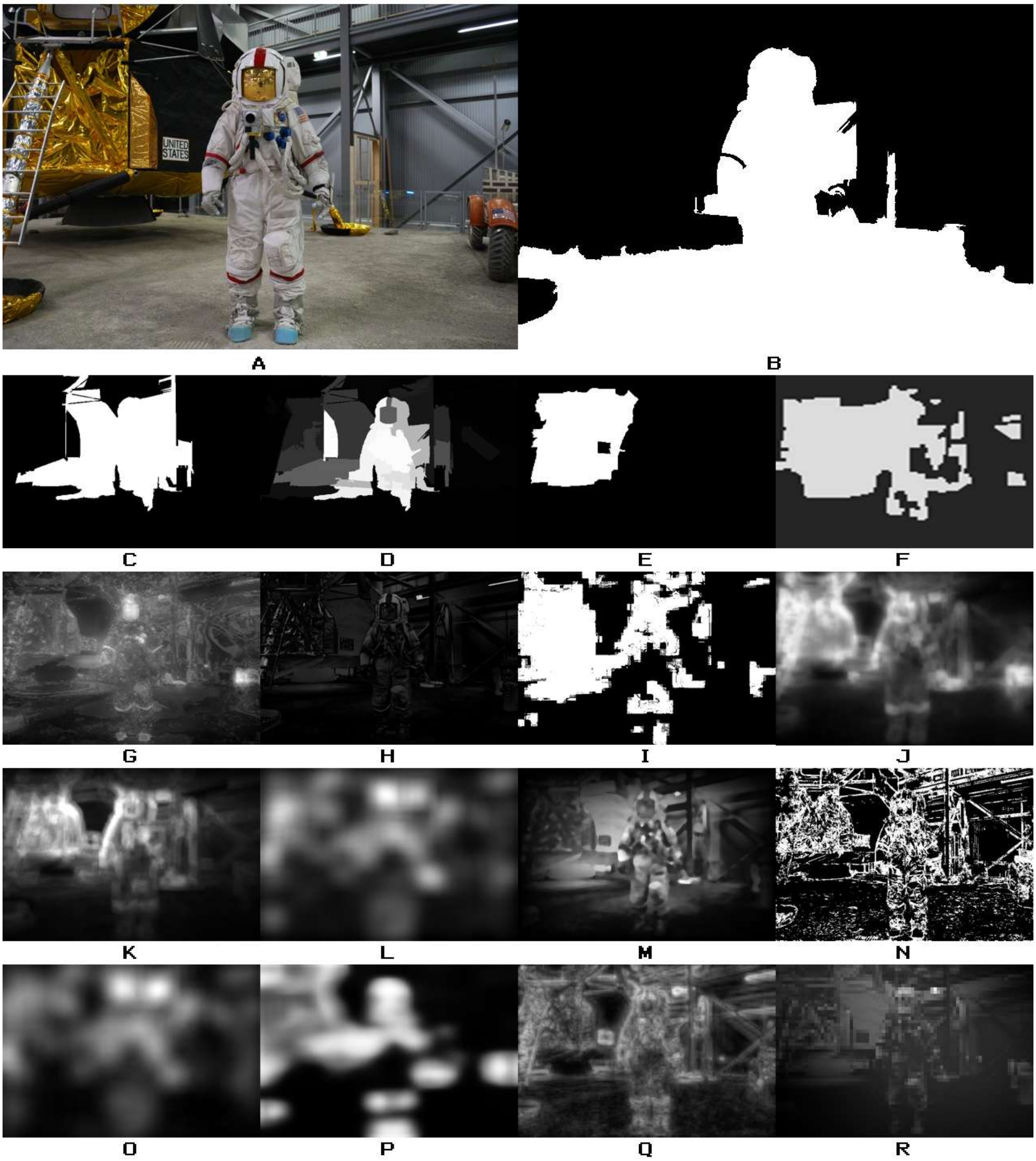


Figure A.4: Test image: astronaut

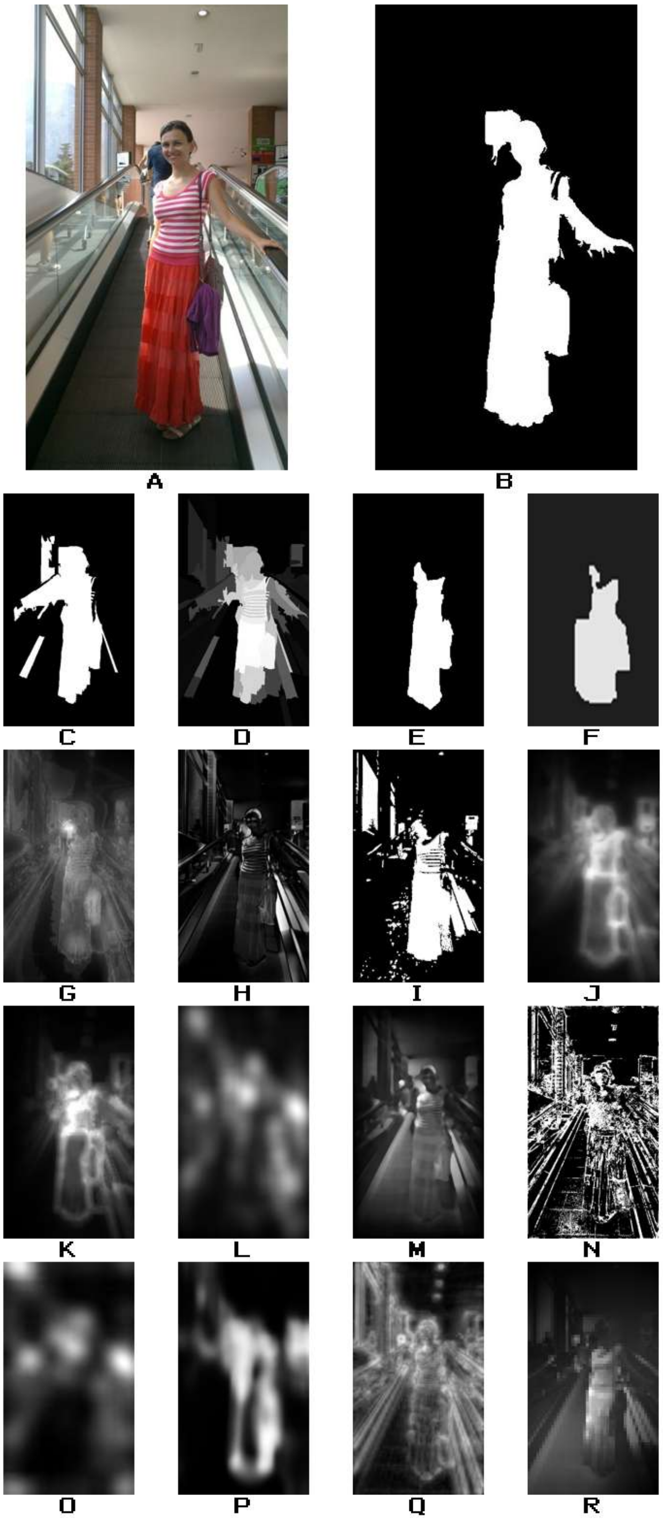


Figure A.5: Test image: lady

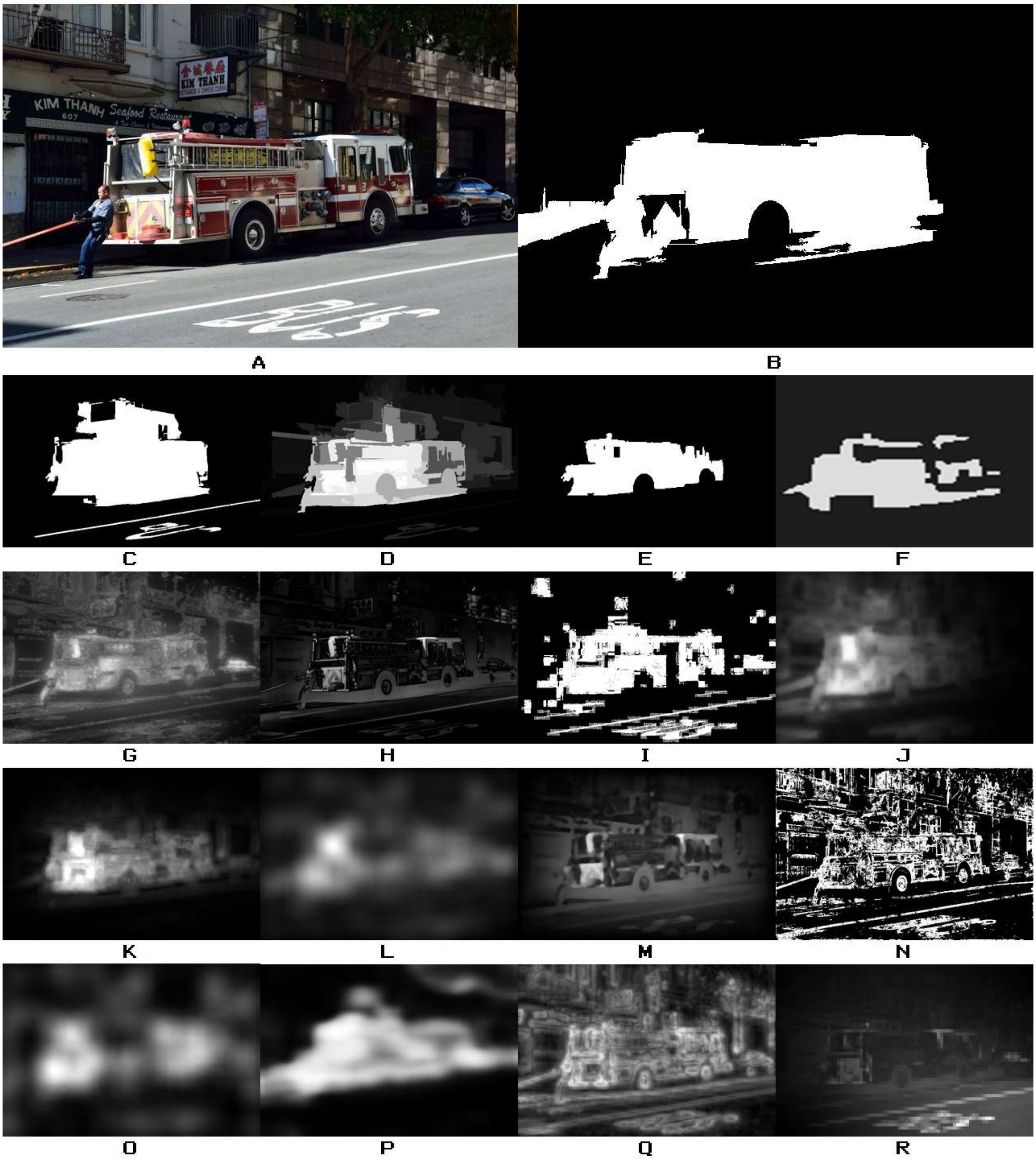


Figure A.6: Test image: fire-car

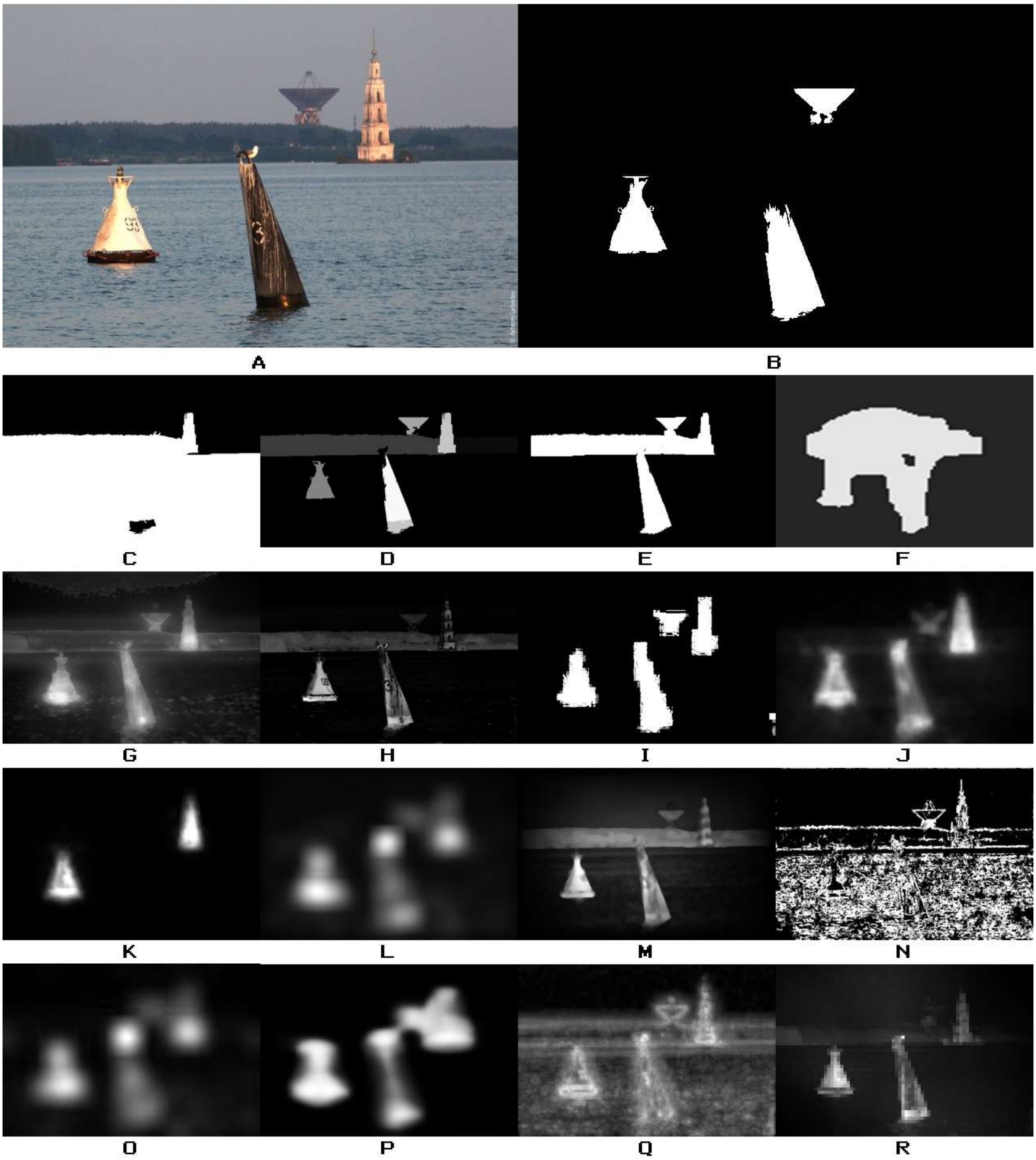


Figure A.7: Test image: aerial

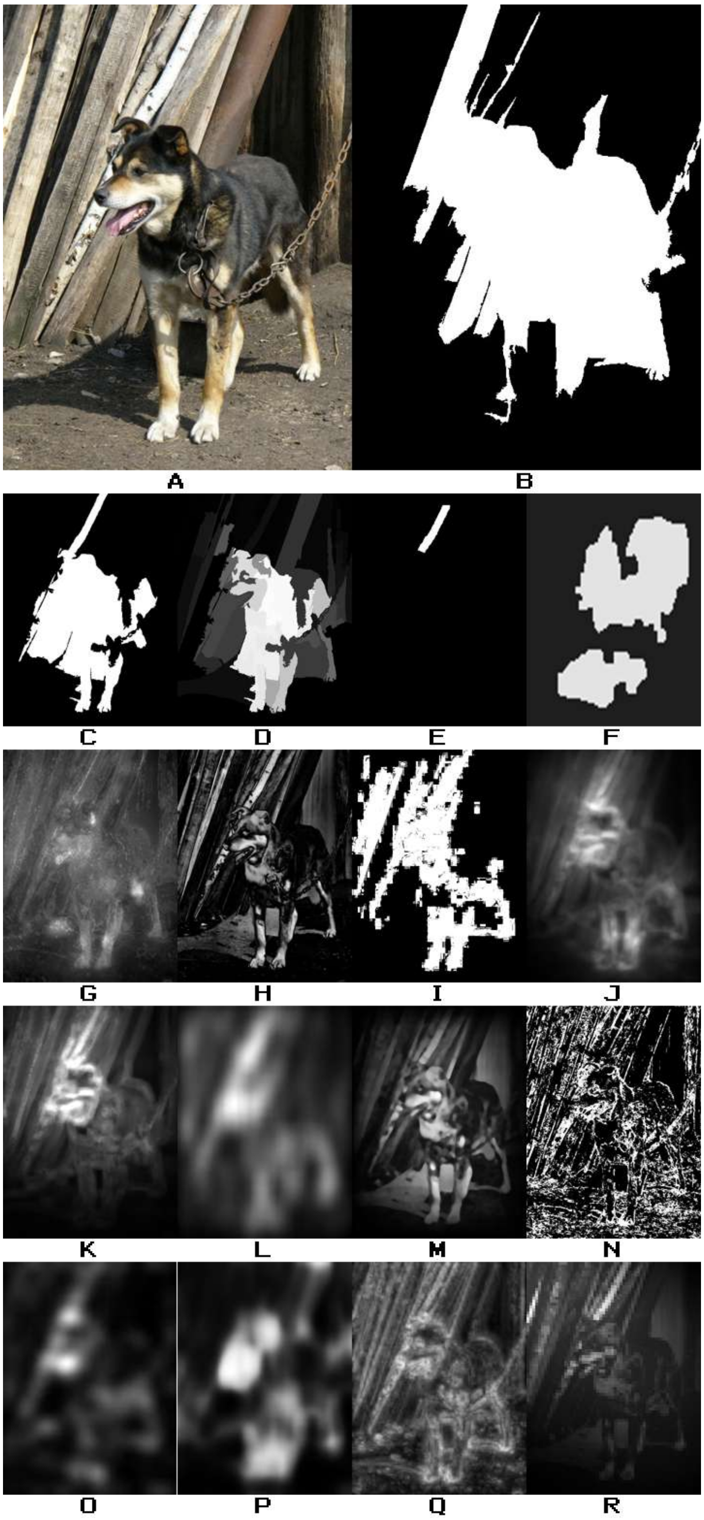


Figure A.8: Test image: dog

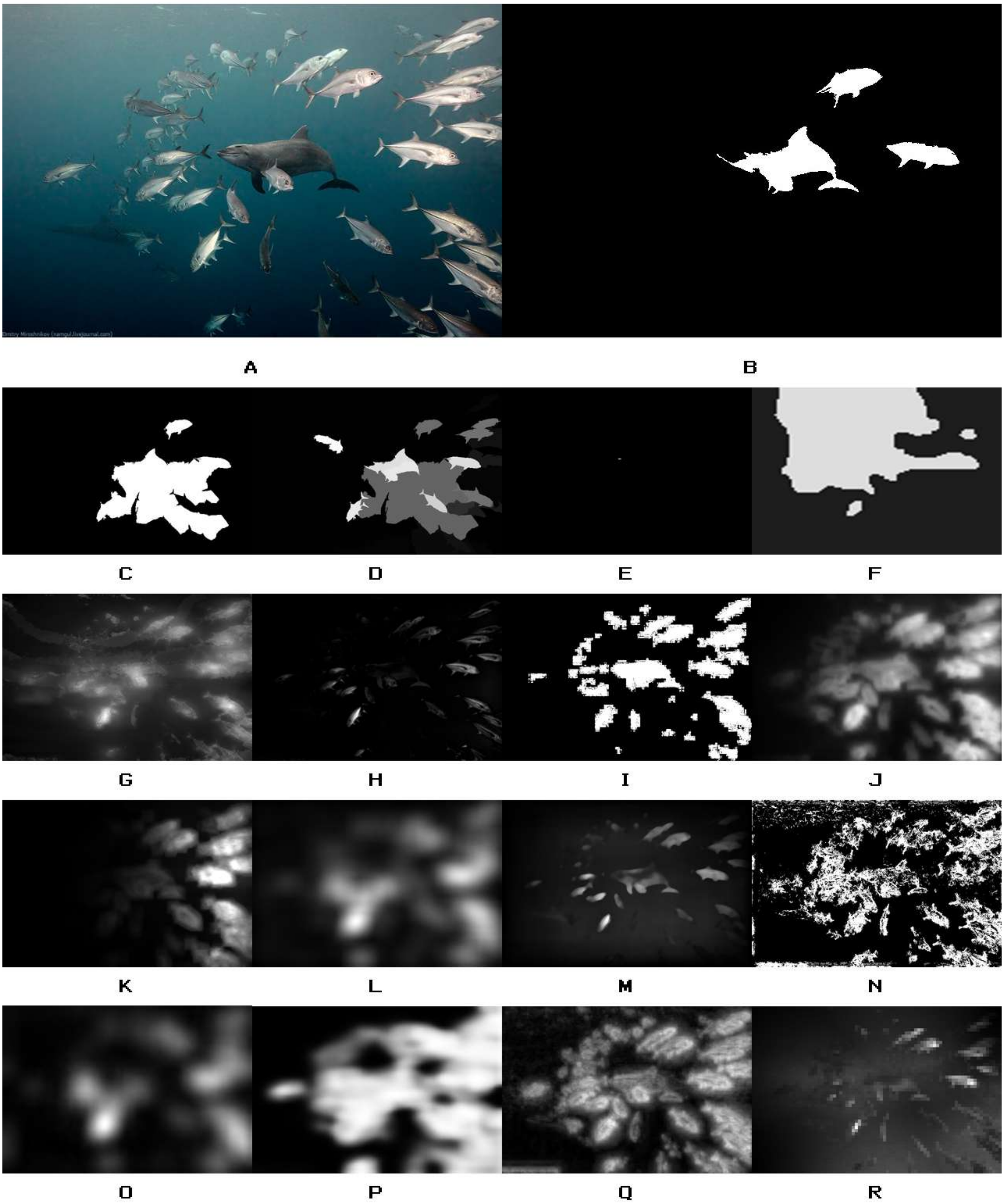


Figure A.9: Test image: dolphin

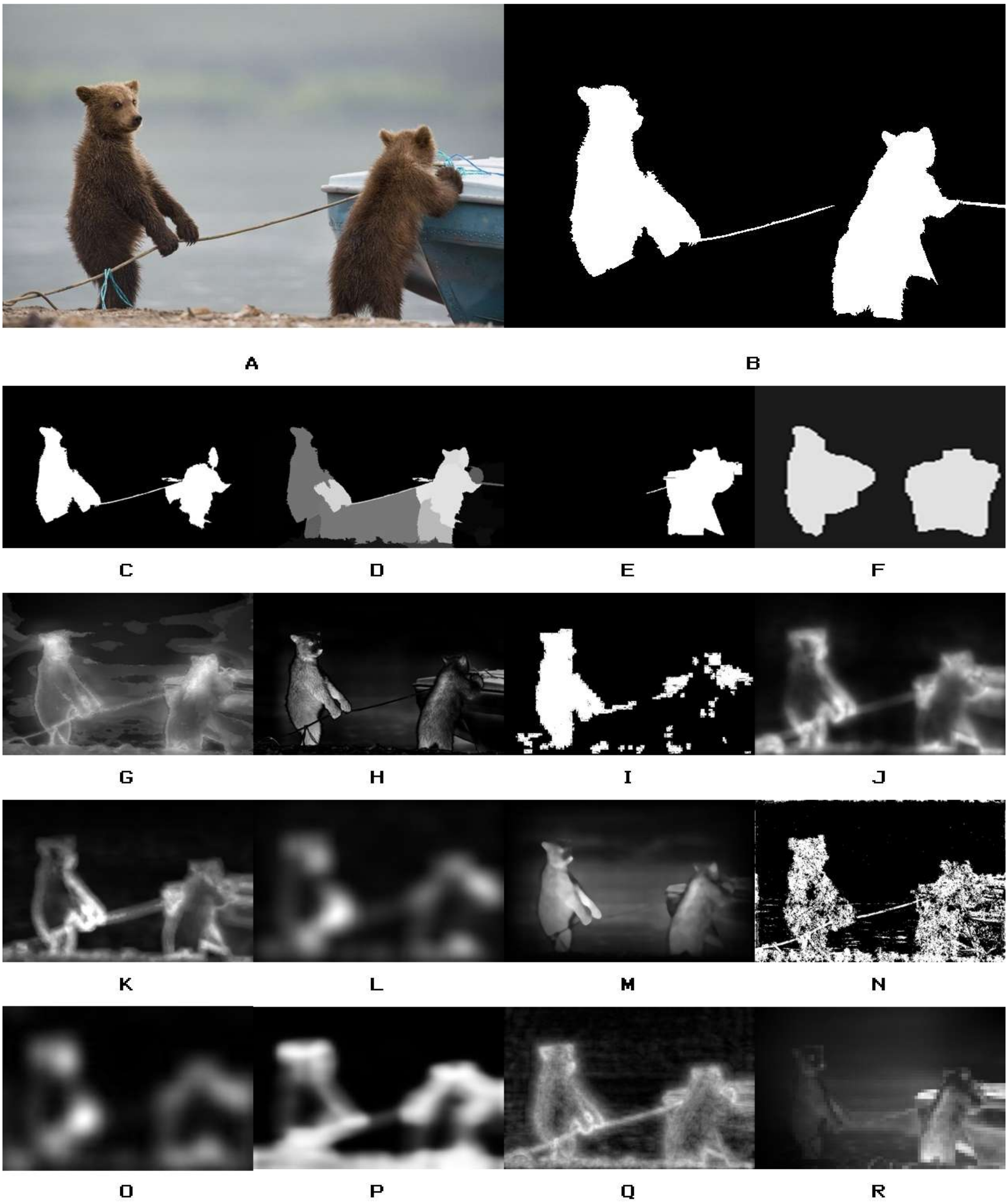
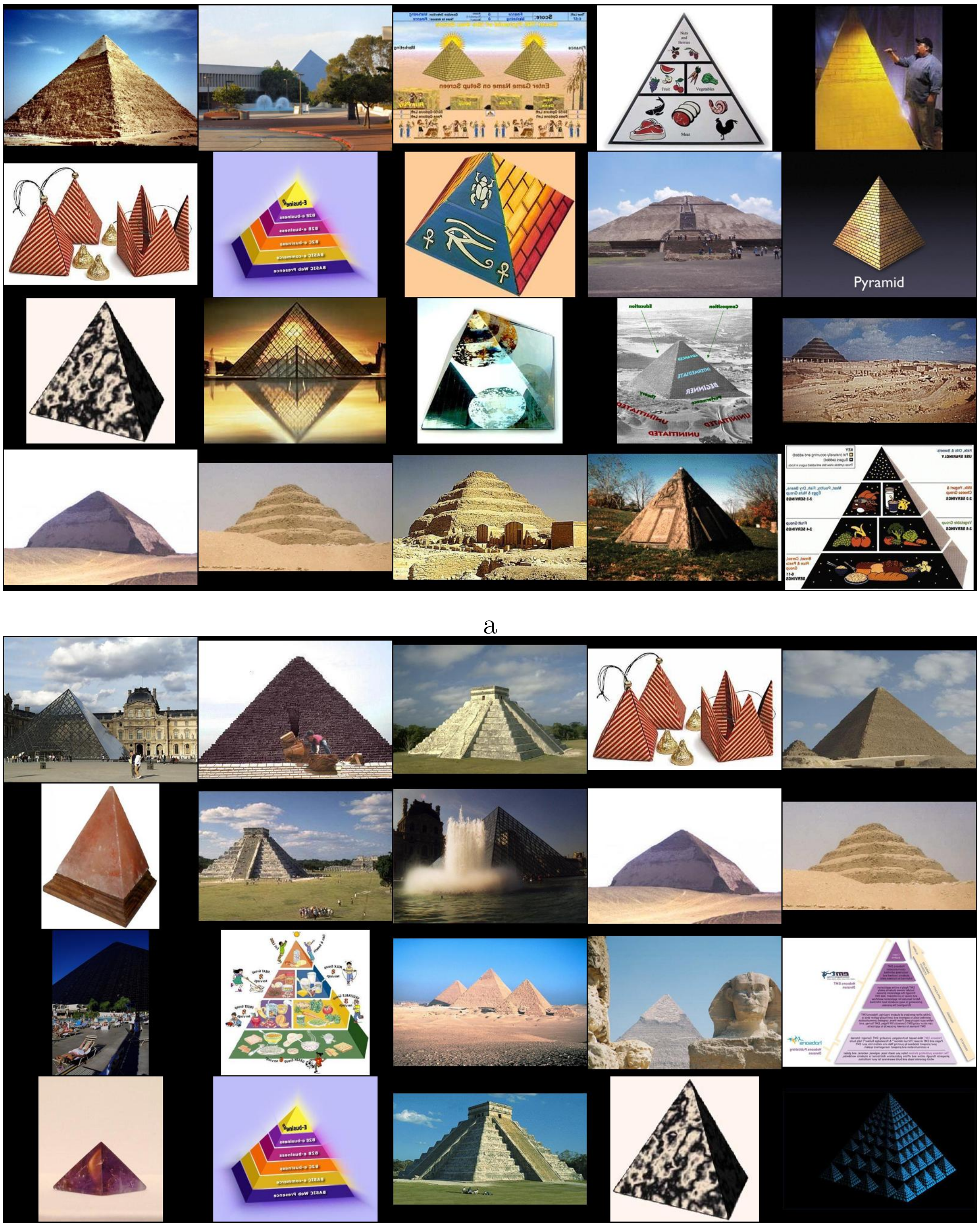


Figure A.10: Test image: bears

Appendix B

Diversity evaluation

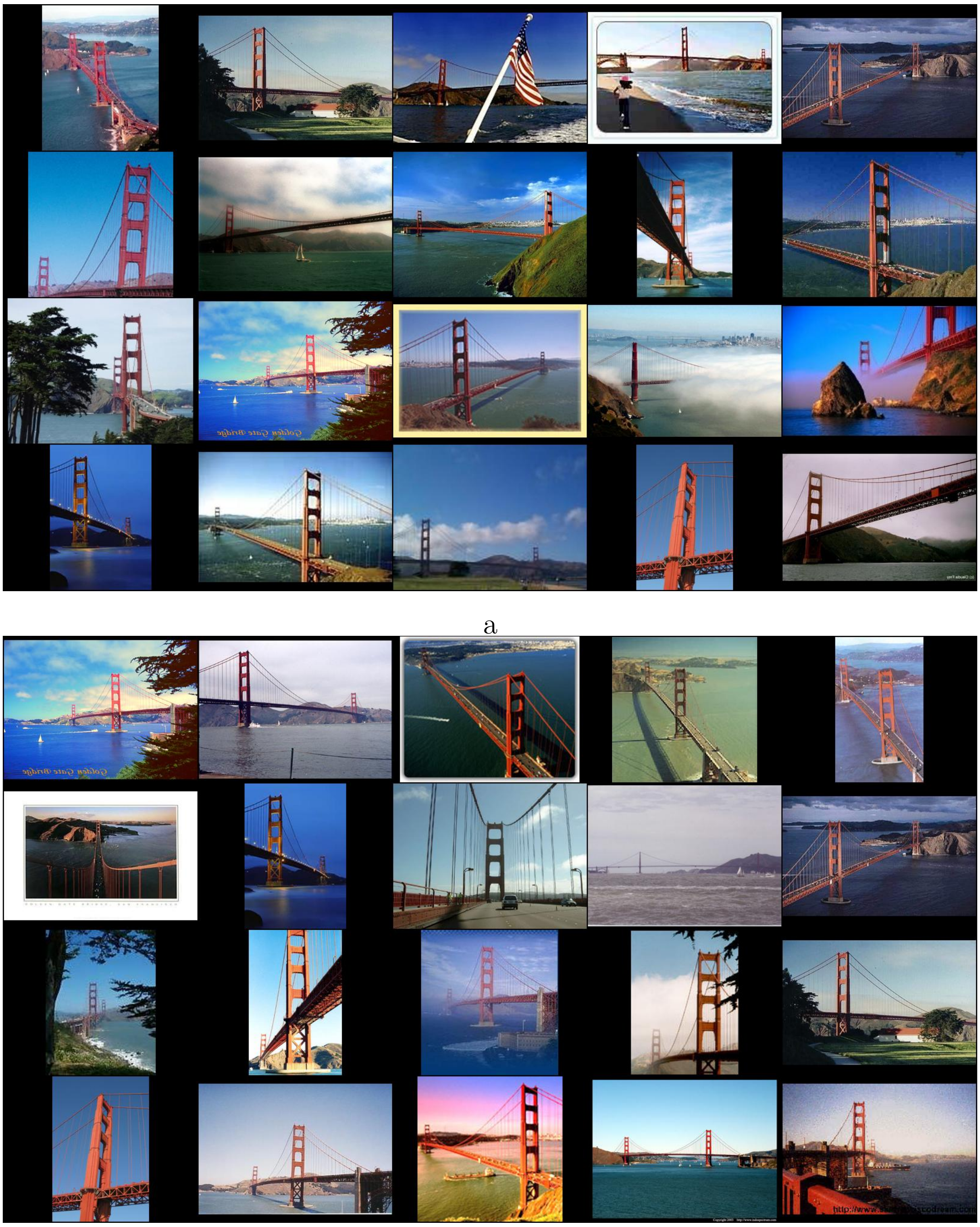
On the following pages additional examples on image set diversification are given. Namely, the examples include pyramid, bridge, blimp and gas pump categories from Caltech 256 dataset. Below each image set comments on main differences in concept representation are provided.



a

b

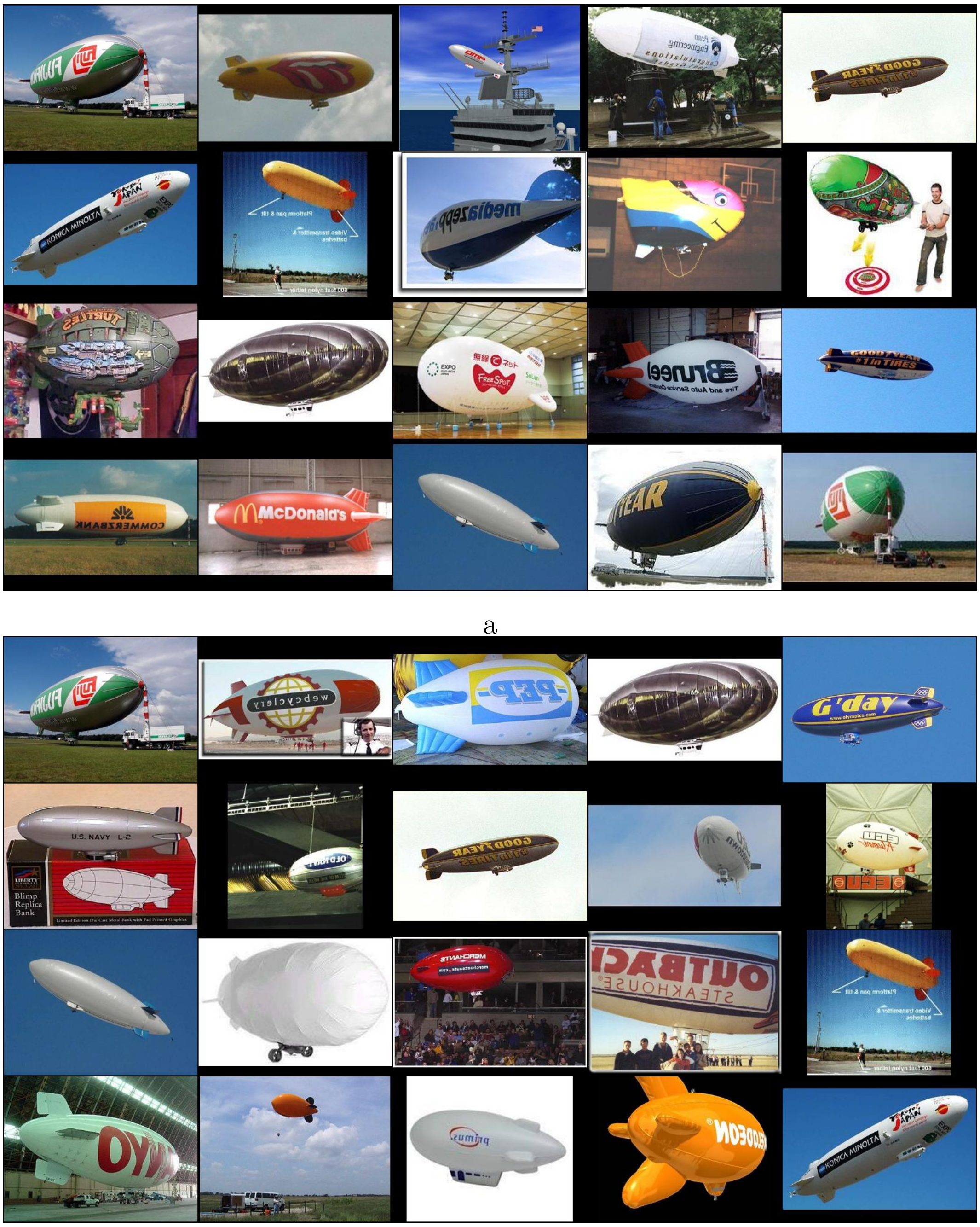
Figure B.1: Example of re-ranking for category pyramid using the method with saliency information (a) compared to the method without saliency information (b). Notice the difference in number of pyramid instances in upper and lower images.



a

b

Figure B.2: Example of re-ranking for category bridge using the method with saliency information (a) compared to the method without saliency information (b). Notice presence of images with a person running, flag and stone in the upper image set.



a

b

Figure B.3: Example of re-ranking for category blimp using the method with saliency information (a) compared to the method without saliency information (b). Notice presence of a compute generation, fish-shaped and accompanied with a human blimps in the upper set.

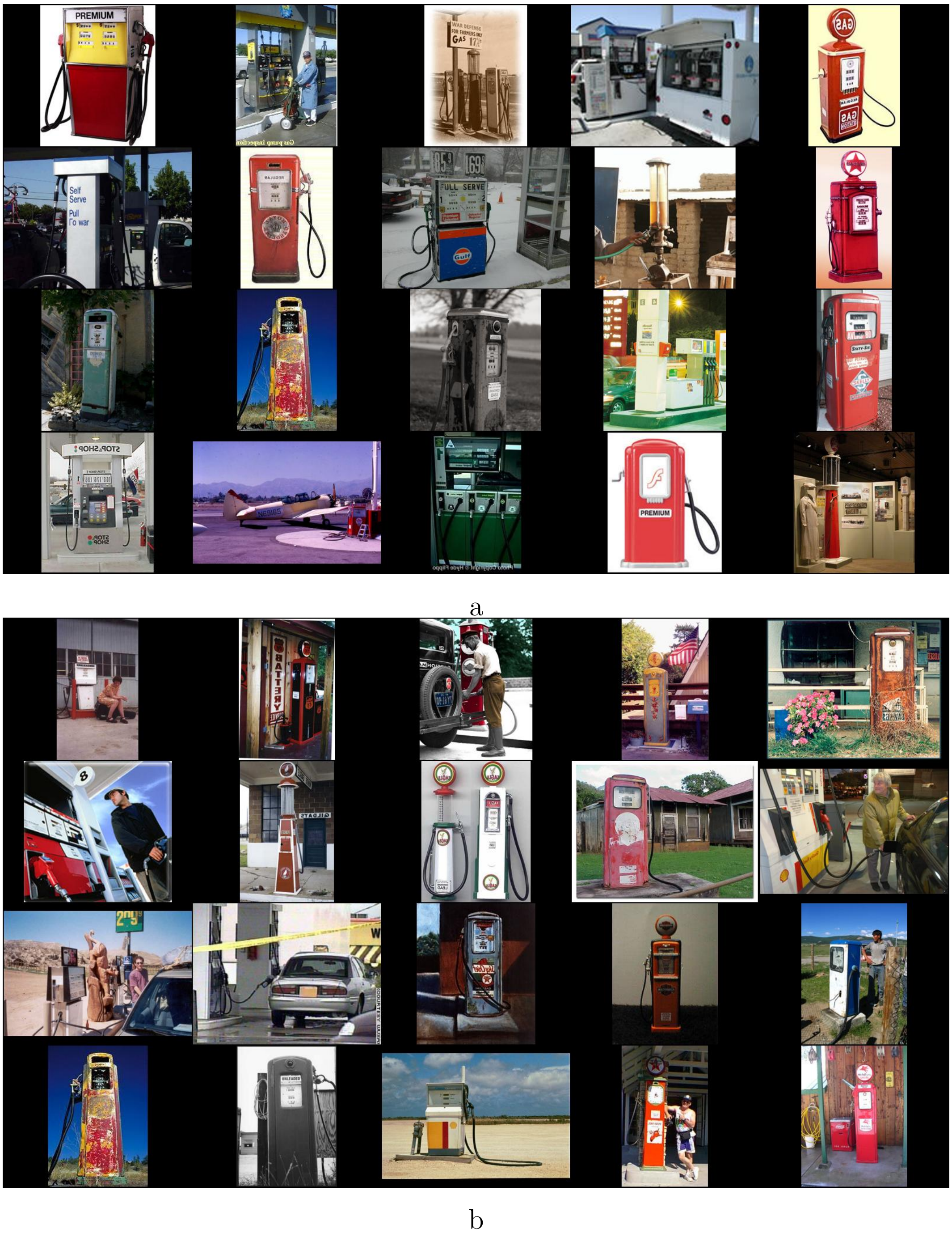


Figure B.4: Example of re-ranking for category gas-pump using the method with saliency information (a) compared to the method without saliency information (b). Notice broader variety of locations and types of gas-pumps in the upper image set.