Visual Saliency by Keypoints Distribution Analysis

Edoardo Ardizzone, Alessandro Bruno, and Giuseppe Mazzola

Dipartimento di Ingegneria Chimica, Gestionale, Informatica e Meccanica.
Università degli Studi di Palermo, viale delle Scienze ed. 6, 90128, Palermo, Italy
ardizzon@unipa.it,
{bruno,mazzola}@dinfo.unipa.it

Abstract. In this paper we introduce a new method for Visual Saliency detection. The goal of our method is to emphasize regions that show rare visual aspects in comparison with those showing frequent ones. We propose a bottom up approach that performs a new technique based on low level image features (texture) analysis. More precisely, we use SIFT Density Maps (SDM), to study the distribution of keypoints into the image with different scales of observation, and its relationship with real fixation points. The hypothesis is that the image regions that show a larger distance from the mode (most frequent value) of the keypoints distribution over all the image are the same that better capture our visual attention. Results have been compared to two other low-level approaches and a supervised method.

Keywords: saliency, visual attention, texture, SIFT.

1 Introduction

One of the most challenging issues in Computer Vision field is the detection of salient regions in an image. Psychovisual experiments [1] suggest that, in absence of any external guidance, attention is directed to visually salient locations in the image. Visual Saliency or Saliency mainly deal with identifying fixation points that a human viewer would focus on at the first glance. Visual saliency usually refers to a property of a "point" in an image (scene), which makes it likely to be fixated. Most models for visual saliency detection are inspired by human visual system and tend to reproduce the dynamic modifications of cortical connectivity for scene perception. In scientific literature Saliency approaches can be subdivided in three main groups: Bottom-up, Top-down, Hybrid.

In Bottom-up approaches (stimulus driven) human attention is considered a cognitive process that selects most unusual aspects of an environment while ignoring more common aspects. In [2] the method is based on parallel extraction of various feature maps using center-surround differences. In [3] multiscale image features are combined into a single topographical saliency map. A dynamical neural network then selects attended locations in order of decreasing saliency. Harel et al. in [4] proposed graph based activation maps.

In Top-down approaches [5,6] the visual attention process is considered task dependent, and the observer's goal in scene analysis is the reason why a point is fixed

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part I, LNCS 6978, pp. 691–699, 2011.

rather than others. Object and face detection are examples of high level tasks that guide the human visual system in top-down view.

Generally Hybrid systems for saliency use the combination of the two levels, bottom-up and top-down. In hybrid approaches [7,8] Top-down layer usually cleans the noisy map extracted from Bottom-up layer. In [7] top-down component is face detection. Chen et al. [8] used a combination of face and text detection and they found the optimal solutions through branch and bound technique.

A common problem for many of these models is that they often don't match real fixation maps of a scene. A newer kind of approach was proposed by Judd et al. [9] who built a database [10] of eye tracking data from 15 viewers. Low, middle and high-level features of this data have been used to learn a model of saliency. In our work we aimed to further study this problem. We decided to investigate about the relationship between real fixation points and computer generated distinctive points. Our method performs a new measure of visual saliency based on image low level features, particularly through the distribution of keypoints extracted by SIFT algorithm, as descriptor of texture variations into the image. In this work we are not interested in color information. Our method is totally unsupervised and it belongs to bottom-up saliency methods. We measured method effectiveness comparing resulting maps with real fixation maps of the reference database [10] and with two of the most important bottom–up approaches [3][4] and a hybrid method[9].

2 Proposed Saliency Measure

Our method propose a new measure of Visual Saliency, focusing on low level image features such as texture. What's the matter for which we use texture information for detecting visual saliency? The answer is that texture gives us important information about image "behavior". The base for extracting salient regions, according to our method, is to emphasize texture rare event. We decide to study the spatial distribution of keypoints inside an image to describe texture variations all over the image. The levels of roughness of both fine and coarse regions can be very different (in a fine region we will find a larger number of keypoints than in coarse regions), so we use keypoints density, to find various texture events and to identify the most salient regions. In this work we use SIFT algorithm to extract keypoints from an image. Then we introduce the concept of SIFT density maps (SDM) which are used to compute the final saliency map.

2.1 SIFT Feature

SIFT (Scale Invariant Feature Transform) descriptors [11] are generated by finding interesting local keypoints, in a greyscale image, by locating the maxima end the minima of Difference-of-Gaussian in the scale-space pyramid. SIFT algorithm takes different levels (octaves) of Gaussian blur on the input image, and computes the difference between the neighboring octaves. Information about orientation vector is then computed for each keypoint, and for each scale. Briefly, a SIFT descriptor is a 128-dimensional vector, which is computed by combining the orientation histograms of locations closely surrounding the keypoint in scale-space. The most important advantage of SIFT descriptors is that they are invariant to scale and rotation, and relatively robust to perspective changes. SIFT can be very useful for many computer vision application: image registration, mosaicing, object recognition and tracking, etc. Their main drawback is the relatively high dimensionality which make them less suitable for nearest neighbor lookups against a training dataset.

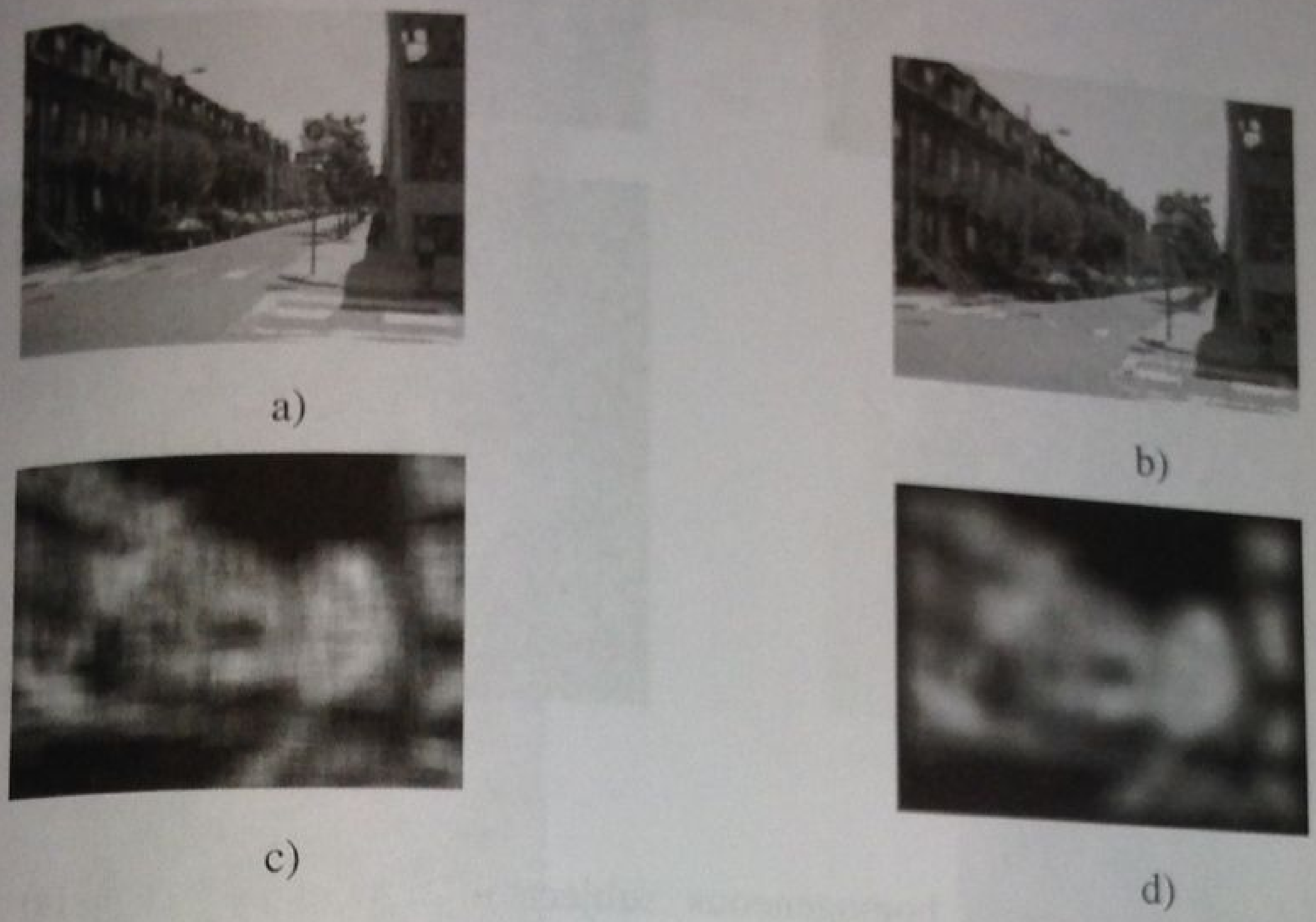


Fig. 1. Original Image (a), SIFT keypoints (b), SIFT Density Map (k=64) (c), final Saliency Map (d)

2.2 SIFT Density Maps

A SIFT Density Map (SDM) is a representation of the density of keypoints in an image, and can give essential information about the regularity of its texture. A SIFT Density Map SDM(k) is built by counting the number of keypoints into a sliding window of size k x k, which represent our scale of observation. Each point in the SDM(k) indicates the number of keyponts into a squared area of size k x k, centered in corresponding point of the image. It is evident that density values are strictly related to the value of k, and are limited by the window size. In fact smaller windows should be sensible to texture variations at a finer level, while larger windows will emphasize coarser deviations. In section 3 we will discuss the sensibility of the results with k.

In real scenes, the simultaneous presence of many elements (the sky, the urban habitations, the urban green spaces) will show many kinds of texture. From a SIFT distribution point of view, the homogeneous surface of the sky has almost null values, the urban green spaces has mean density while urban habitations have high concentration of keypoints. (fig. 1)

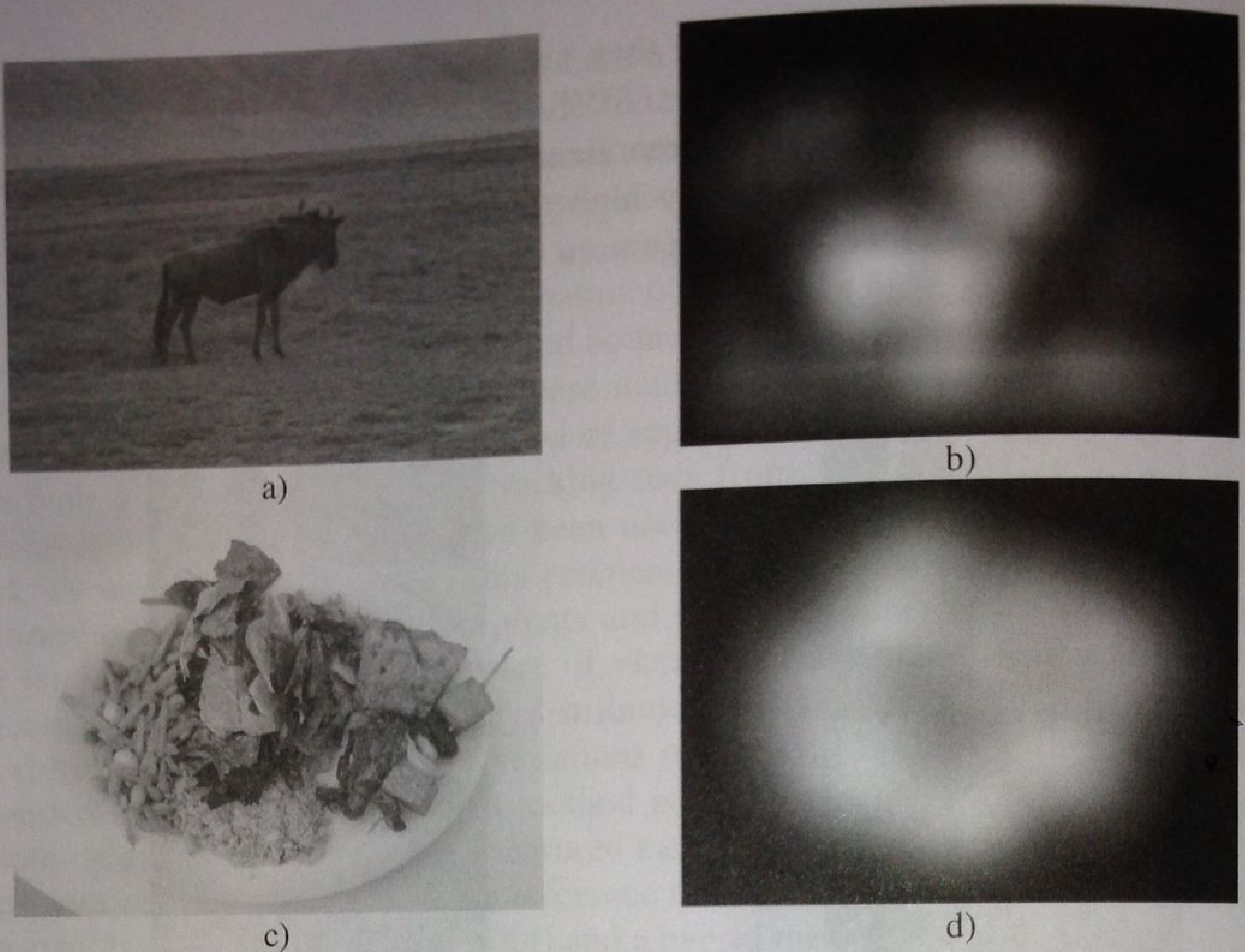


Fig. 2. Two image examples: a homogeneous subject in a textured scene (a) and the corresponding Saliency Map; a textured object in a homogeneous background (c) and the corresponding Saliency Map (d)

2.3 Saliency Map

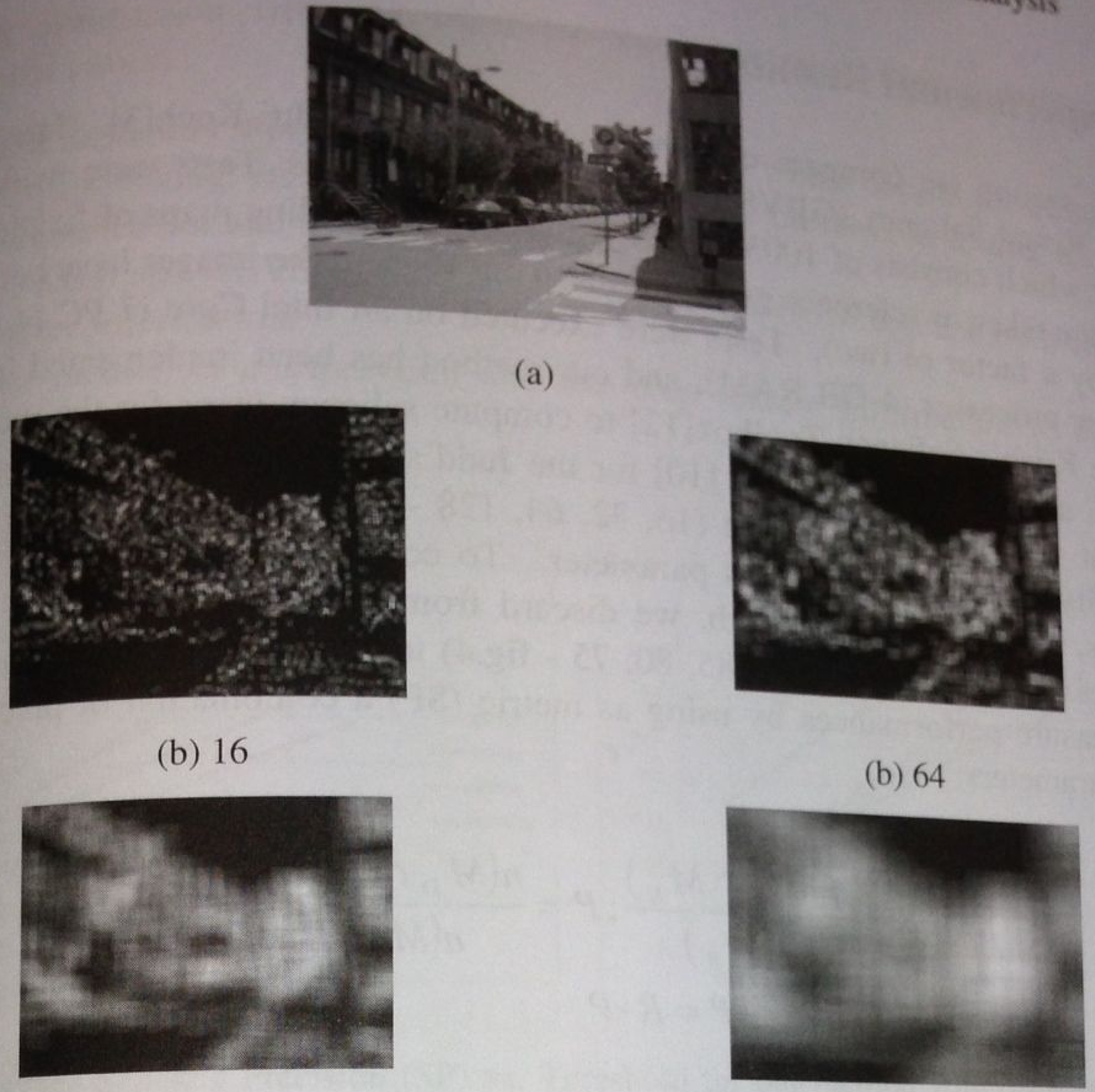
Our saliency map SM, for a given k, is built as the absolute difference between the SDM values and the most frequent value MV of the map:

$$SM(k)=\left|SDM(k)-MV(SDM(k))\right| \tag{1}$$

which is further normalized with respect to the maximum value to restrict SM values to [0,1].

The most salient areas into the image are those related to the SDM values with the maximum deviation from the most frequent value, typically the most rare texture events in the image. This measure emphasizes both the case in which a textured object is the salient region, as it is surrounded by homogeneous areas (the most frequent value near to 0), and the case in which a homogeneous area is surrounded by textured parts (a higher most frequent value). (fig. 2)

In addition, for a smoother representation of the saliency map, we apply to the SM an average filter which has a window size that is a half of that used to build the map (k).



(a)

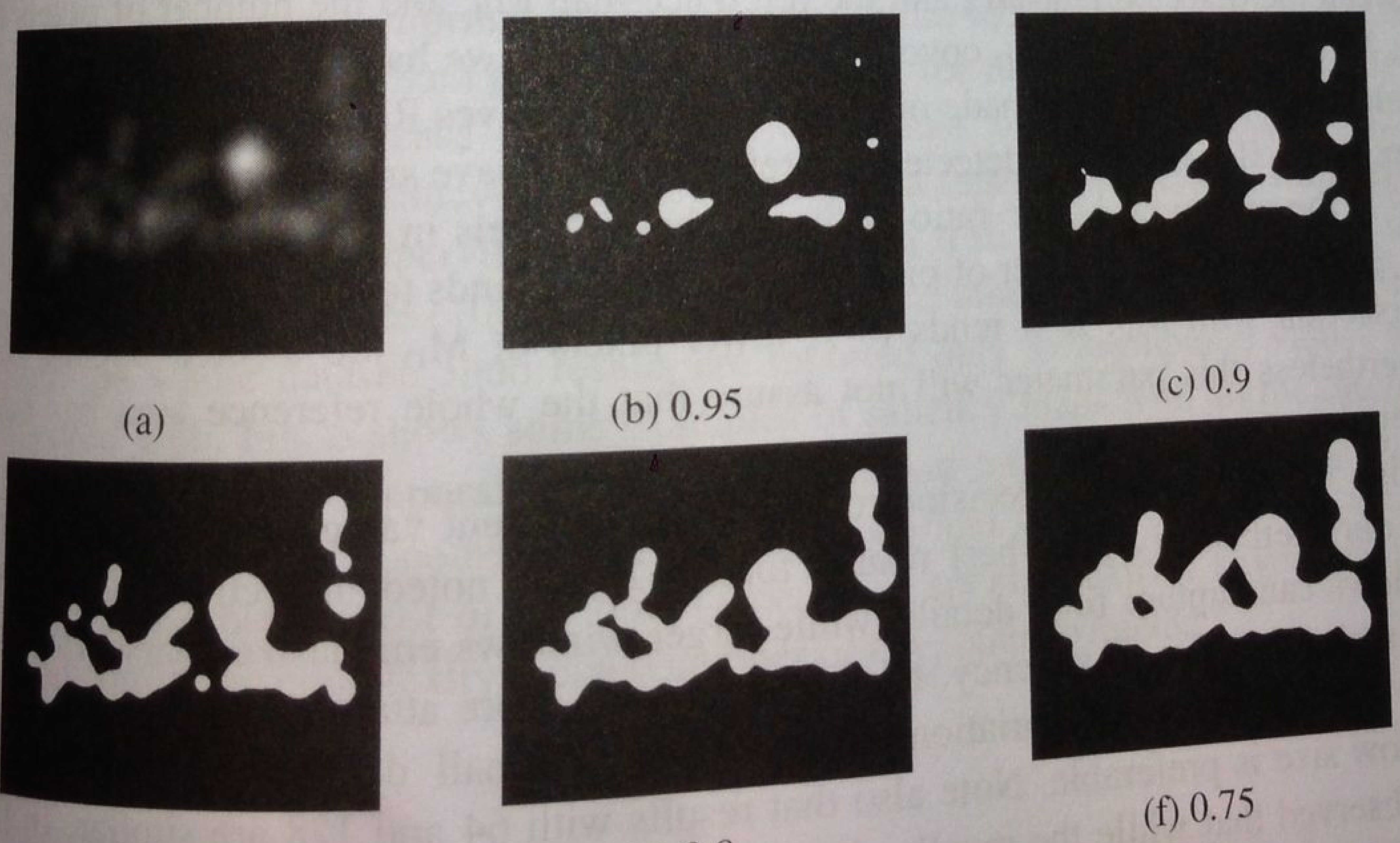
(b) 16

(b) 64

(c) 32

(c) 128

Fig. 3. Original Image (a), SIFT Density Maps with different values of k (16,32,64,128)



(a)

(b) 0.95

(c) 0.9

(d) 0.85

(e) 0.8

(f) 0.75

Fig. 4. Fixation Map (a) of the image in fig. 3.a, Binary maps with different thresholds (0.95, 0.9, 0.85, 0.8, 0.75)

3 Experimental Results

In this section we compare our results with those of Itti-Koch[3], Harel's Graph Based Visual Saliency (GBVS) [4] and Judd [9] methods. Tests were made on [10] dataset which consists of 1003 images and the corresponding maps of fixation points, which are taken as reference groundtruth (in our tests all the images have been resized down by a factor of two). Tests were executed on an Intel Core i7 PC (4 CPU, 1.6 GHz per processor, 4 GB RAM), and our method has been implemented in Matlab. We use Koch's Saliency Toolbox[12] to compute saliency maps for the methods [3] and [4], and the maps given in [10] for the Judd's method. Tests were repeated for different values of window size (16, 32, 64, 128 - fig.3), with the aim to study the sensibility of the results to this parameter. To compare our results with the other methods, and to the groundtruth, we discard from the saliency maps the less N% salient pixels (with N= 95, 90, 85, 80, 75 - fig.4) to create a set of binary maps. We then measure performances by using as metric (SP) a combination of precision and recall parameters:

$$R = \frac{n(M_D \cap M_R)}{n(M_R)}; P = \frac{n(M_D \cap M_R)}{n(M_D)} \qquad (2)$$
$$SP = R \cdot P$$

where M_D is the binary version of the detected saliency map (with our method or the others), while M_R is the binary version of the reference fixation map.

R is the recall, i.e. the ratio between the number of pixels in the intersection between the detected map M_D and the reference map M_R, and the number of pixels in M_R. When it tends to 1, M_D covers the whole M_R, but we have no information about pixels outside M_R (a map made of only salient pixels gives R=1 if compared with any other map). If it tends to 0 detected and reference map have smaller intersection.

P is the precision, i.e the ratio of the number of pixels in the intersection between M_D and M_R, and the number of pixels in M_D. When P tends to 0, the whole M_D has no intersection with M_R. If it tends to 1, fewer pixels of M_D are labeled outside M_R. Nevertheless this parameter will not assure that the whole reference area has been covered.

Fig. 5 shows average precision results versus different values of thresholds. Note that our method gives its best results for k=128. As noted in section 2.2, smaller windows can capture finer details, while larger windows emphasize coarse variation of texture. In terms of saliency, human attention is more attracted by areas in which there are large texture variations, rather than by small deviation. Then a larger window size is preferable. Note also that results with 64 and 128 are similar. In fact we observed that while the recall value increases with the window size, precision, in case of very large window, does not increase as well.

In the comparison with the other methods, we must first underline some fundamental issues:

- Judd method is supervised, and uses 9/10 of the whole dataset for training and 1/10 for testing. Judd results are averaged only on the 100 testing images. It uses both color and texture information.
- Itti-Koch and GBVS method are unsupervised method and use both color and texture information.
- Our method is unsupervised and use only texture information to build the saliency map.

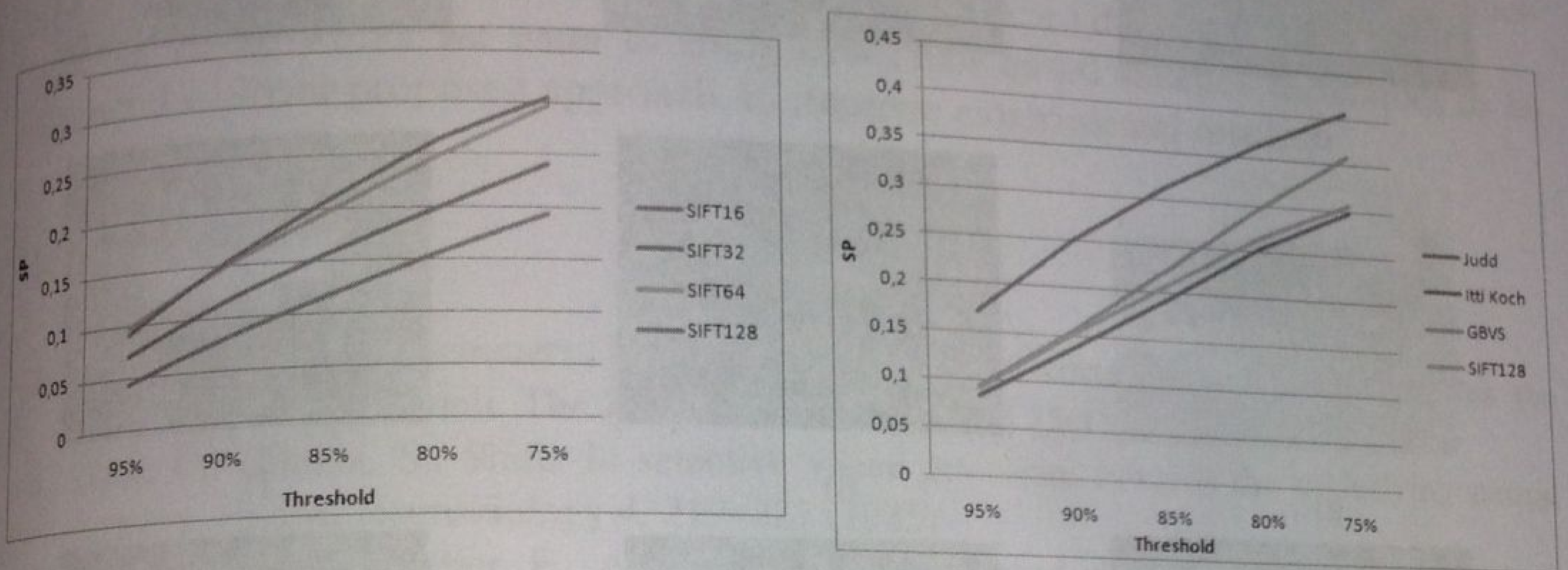


Fig. 5. Average Saliency Precision (SP) vs. Threshold for our method with different window sizes (left), and for our method (k=128) compared to the other methods.

Our saliency map gives better results than Itti-Koch, for all the threshold values, even if we use only texture information. Results are similar to GBVS for higher threshold values (0.95 and 0.9), which give information about the most salient pixels, while our precision does not increase as well for lower values of threshold (0.85, 0.8). As expected Judd method achieves best results, as it is a supervised method, while all the other methods are unsupervised. Furthermore Judd tests refers only within a small selected subset of images (100 testing images), while other methods have been tested within the whole dataset. Judd results are reported only as asymptotic values to be compared with. Fig. 6 shows some examples of saliency maps with all the discussed methods. Regarding temporal efficiency, our method takes less than 10s to build a saliency map, and it is comparable with Itti-Koch and GBVS method for medium images (300 x 600). Most of the time (70% ca) is spent to extract keypoints, but it depends on image complexity, i.e. the number of keypoints extracted.

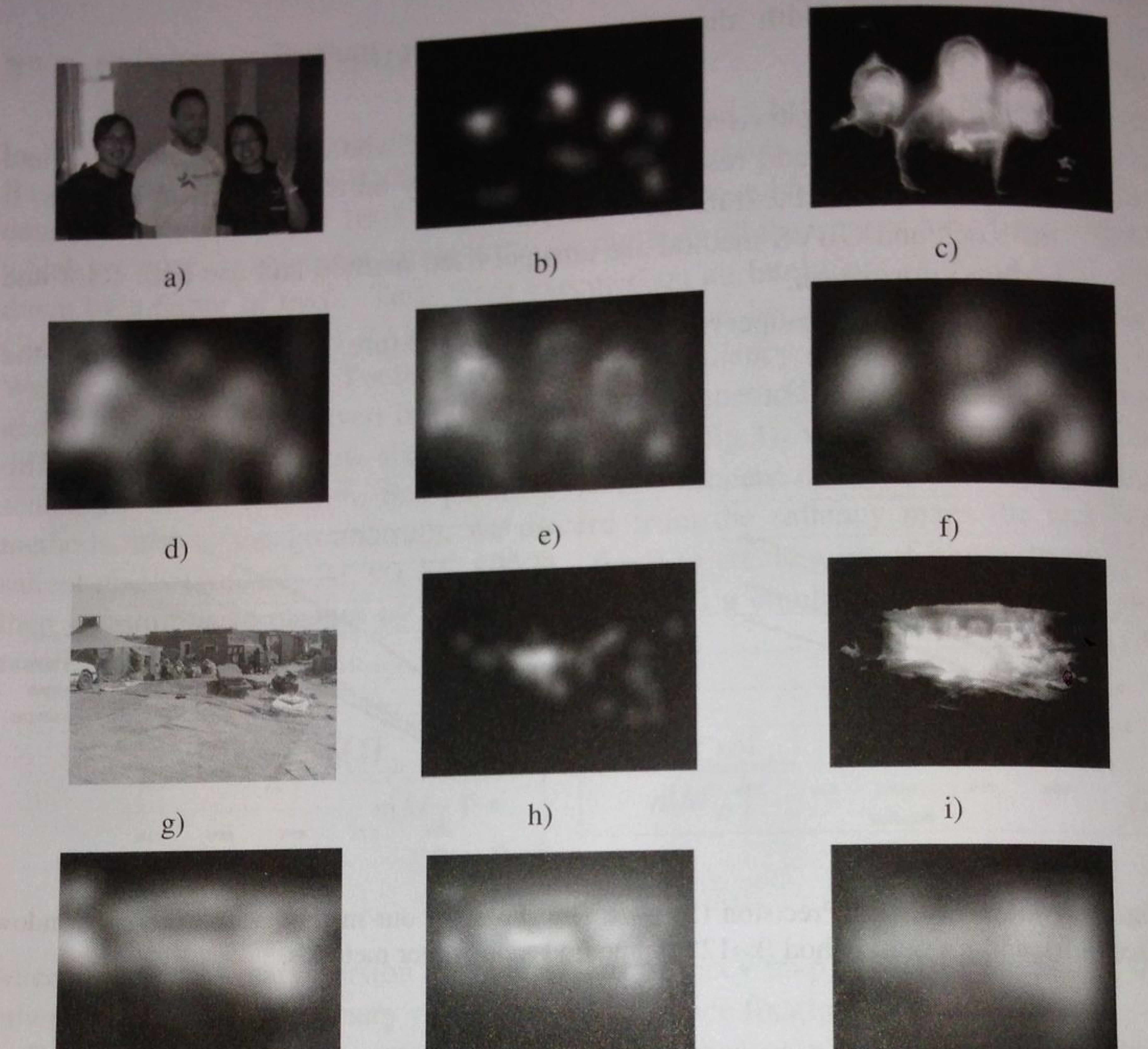


Fig. 6. Some visual results. Original images (a,g,m), fixation maps (b,h,n), Judd maps (c,i,o), Itti-Koch maps (d,j,p), GBVS maps (e,k,q), our method (f,l,r) window size 128.

4 Conclusions

Visual saliency has been investigated for many years but it is still an open problem, especially if the aim is to investigate the relationship between synthetic maps and points, in a real scene, that attract a viewer attention.

The purpose of this paper was to study how computer generated keypoints are related to real fixation points. No color information has been used to build our saliency maps, as keypoints are typically related only to image texture property.

Even if we use only texture information, experimental results show that our method is very competitive with respect of two of the most cited low-level approaches. Judd's method achieves better results as it is a supervised method which has been trained with the fixation maps within the selected dataset.

In our future works we want to study new color based saliency techniques to be integrated with our proposed approach, to improve experimental results.

References

1. Constantinidis, C., Steinmetz, M.A.: Posterior parietal cortex automatically encodes the location of salient stimuli. The Journal of Neuroscience 25(1), 233–238 (2005)
2. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology 4, 219–227 (1985)
3. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)
4. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems 19, pp. 545–552. MIT Press, Cambridge (2007)
5. Luo, J.: Subject content-based intelligent cropping of digital photos. In: IEEE International Conference on Multimedia and Expo (2007)
6. Sundstedt, V., Chalmers, A., Cater, K., Debattista, K.: Topdown visual attention for efficient rendering of task related scenes. In: In Vision, Modeling and Visualization, pp. 209–216 (2004)
7. Itti, L., Koch, C.: Computational modeling of visual attention. Nature Reviews Neuroscience 2(3) (2001)
8. Chen, L.-Q., Xie, X., Fan, X., Ma, W.-Y., Zhang, H.-J., Zhou, H.-Q.: A visual attention model for adapting images on small displays. ACM Multimedia Systems Journal 9(4) (2003)
9. Judd, Y., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: IEEE 12th International Conference on Computer Vision, pp. 2106–2133 (2009)
10. `http://people.csail.mit.edu/tjudd/WherePeopleLook/index.html`
11. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
12. `http://www.saliencytoolbox.net`

Lecture Notes in Computer Science

The LNCS series reports state-of-the-art results in computer science research, development, and education, at a high level and in both printed and electronic form. Enjoying tight cooperation with the R&D community, with numerous individuals, as well as with prestigious organizations and societies, LNCS has grown into the most comprehensive computer science research forum available.

The scope of LNCS, including its subseries LNAI and LNBI, spans the whole range of computer science and information technology including interdisciplinary topics in a variety of application fields. The type of material published traditionally includes

- proceedings (published in time for the respective conference)
- post-proceedings (consisting of thoroughly revised final full papers)
- research monographs (which may be based on outstanding PhD work, research projects, technical reports, etc.)

More recently, several color-cover sublines have been added featuring, beyond a collection of papers, various added-value components; these sublines include

- tutorials (textbook-like monographs or collections of lectures given at advanced courses)
- state-of-the art surveys (offering complete and mediated coverage of a topic)
- hot topics (introducing emergent topics to the broader community)

In parallel to the printed book, each new volume is published electronically in LNCS Online.

Detailed information on LNCS can be found at
www.springer.com/lncs

Proposals for publication should be sent to
LNCS Editorial, Tiergartenstr. 17, 69121 Heidelberg, Germany
E-mail: lncs@springer.com

ISSN 0302-9743

ISBN 978-3-642-24084-3

9 783642 240843

Lecture Notes in Computer Science
LNCS LNAI LNBI

springer.com

Maino • Foresti (Eds.)

LNCS
6978

Image Analysis
and Processing –
ICIAP 2011

1 Part I

IAPR

ICIAP
2011

Giuseppe Maino
Gian Luca Foresti (Eds.)

LNCS 6978

Image Analysis and Processing – ICIAP 2011

16th International Conference
Ravenna, Italy, September 2011
Proceedings, Part I

1 Part I

IAPR

Springer