A Comparison of Feature Detectors with Passive and Task-Based Visual Saliency

Patrick Harding and Neil Robertson
Heriot-Watt University, Edinburgh, UK

This paper investigates the coincidence between six interest point detection methods (SIFT, MSER, Harris-Laplace, SURF, FAST & Kadir-Brady Saliency) with two robust "bottom-up" models of visual saliency (Itti and Harel) as well as "task" salient surfaces derived from observer eye-tracking data. Comprehensive statistics for all detectors vs. saliency models are presented in the presence and absence of a visual search task. It is found that SURF interest-points generate the highest coincidence with saliency and the overlap is superior by 15% for the SURF detector compared to other features. The overlap of image features with task saliency is found to be also distributed towards the salient regions. However the introduction of a specific search task creates high ambiguity in knowing how attention is shifted. It is found that the Kadir-Brady interest point is more resilient to this shift but is the least coincident overall.

1 Introduction and prior work

In Computer Vision there are many methods of obtaining distinctive "features" or "interest points" that stand out in some mathematical way relative to their surroundings. These techniques are very attractive because they are designed to be resistant to image transformations such as affine viewpoint shift, orientation change, scale shift and illumination. However despite their robustness they do not *necessarily* relate in a meaningful way to the human interpretation of what in an image is distinctive. Let us consider a practical example of why this might be important. An image processing operation should only be applied if it aides the observer to perform an interpretation task (enhancement algorithms) or does not destroy the key details within the image (compression algorithms). We may wish to predict the effect of an image processing algorithm on a human's ability to interpret the image. Interest points would be a natural choice to construct a metric given their robustness to transforms. But in order to use these feature points we must first determine (a) how well the interest-point detectors coincide with the human visual system's impression of images i.e. what is visually salient, and (b) how the visual salience changes in the presence of a task such as "find all cars in this image". This paper seeks to address these problems. First let us consider the interest points and then explain in more detail what we mean by feature points and visual salience.

Interest Point detection: The interest points chosen for analysis are: SIFT [1], MSER [2], Harris-Laplace [3], SURF [4], FAST [5,6] and Kadir-Brady Saliency [7].

Fig. 1. An illustration of distribution of the interest-point detectors used in this paper.

These are shown superimposed on one of our test images in Figure 1*. These schemes are well-known detectors of regions that are suitable for transformation into robust regional descriptors that allow for good levels of scene-matching via orientation, affine and scale shifts. This set represents a spread of different working mechanisms for the purposes of this investigation. These algorithms have been assessed in terms of mathematical resilience [8,9]. But what we are interested in is how well they correspond to visually salient features in the image. Therefore we are not investigating descriptor robustness or repeatability (which has been done extensively – see e.g. [8]), nor trying to *select* keypoints based on modelled saliency (such as the efforts in [10]) but rather we want to ascertain how well interest-point locations naturally correspond to *saliency* maps generated under passive and task conditions. This is important because if the interest-points coincide with salient regions at a higher-than coincidence level, they are attractive for two reasons. First, they may be interpreted as primitive saliency detectors *and* secondly can be stored robustly for matching purposes.

Visual Salience: There exist tested models of "bottom-up" saliency, which accurately predict human eye-fixations under passive observation conditions. In this paper, two models were used, those of saliency by Itti Koch and Neibur [11] and the model by Harel, Koch, and Perona [12]. These models are claimed to be based on observed psycho-visual processes in assessing the saliency of the images. They each create a "Saliency Map" highlighting the pixels in order of ranked saliency using intensity shading values. An example of this for Itti and Harel saliency is shown in Figure 2. The Itti model assesses center-surround differences in *Colour*, *Intensity* and *Orientation* across scale and assigns values to feature maps based on outstanding attributes. Cross scale differences are also examined to give a multi-scale representation of the local saliency. The maps for each channel (*Colour*, *Intensity* and

* Note: these algorithms all act on *greyscale* images. In this paper, colour images are converted to grey values by forming a weighted sum of the RGB components (0.2989 R + 0.5870 G + 0.1140 B)

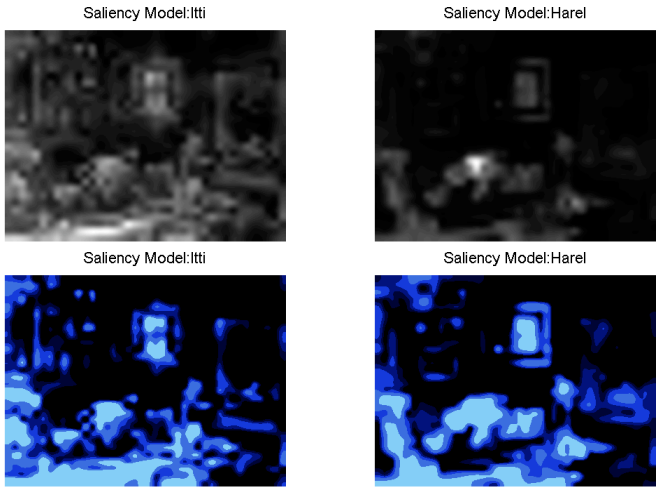

Fig. 2. An illustration of the passive saliency maps on one of the images in the test set. (Top left) Itti Saliency Map, (Top right) Harel Saliency map (Bottom left) thresholded Itti, (Bottom right) thrsholded Harel. Threshold levels are 10, 20, 30, 40 & 50% of image pixels ranked in saliency, represented at descending levels of brightness.

Orientation) are then combined by normalizing and weighting each map according to the local values. Homogenous areas are ignored and "interesting" areas are highlighted. The maps from each channel are then combined into "conspicuity maps" via cross-scale addition. These are combined into a final saliency map by normalization and summed with an equal weighting of 1/3 importance. The model is widely known and is therefore included in this study. However, the combination weightings of the map are arbitrary at 1/3 and it is not the most accurate model at predicting passive eye-scan patterns [12]. The Harel *et al*. method uses a similar basic feature extraction method but then forms *activation* maps in which "unusual" locations in a feature map are assigned high values of activation. Harel uses a Markovian graph-based approach based on a ratio-based definition of dissimilarity. The output of this method is an activation measure derived from pairwise contrast. Finally, the activation maps are normalized using another Markovian algorithm which acts as a mass concentration algorithm, prior to additive combination of the activation maps. Testing of these models in [12] found that the Itti and Harel models achieved, respectively, 84% and 96-98% of the ROC area of a human-based control experiment based on eye-fixation data under passive observation conditions. Harel *et al*. explain that their model is apparently more robust at predicting human performance than Itti because it (a) acts in a center-bias manner, which corresponds to a natural human tendency, and (b) it has superior robustness to differences in the size of salient regions in their model compared to the scale differences in Itti's.

Both models offer high coincidence with eye-fixation from passive viewing observed under strict conditions. The use of both models therefore provides a pessimistic (Itti) and optimistic (Harel) estimation of saliency for passive attentional guidance for each image.

The impact of tasking on visual salience: There is at present no corresponding model of task performance on the saliency map of an image but there has been much work performed in this field, often using eye-tracker data and object learning [13,14,15,16]. It is known that an observer acting under the influence of a specific task will perceive the bottom-up effects mentioned earlier *but will impose constraints* on his observation in an effort to priority-filter information. These impositions will result from experience and therefore are partially composed of memory of likely target positions under similar scenarios. (In Figure 4 the green regions show those areas which *became* salient due to a task constraint being imposed.)

2. Experimental Setup

Given that modeling the effect of tasking on visual salience is not readily quantifiable, in this paper eye-tracker data is used to construct a "task probability surface". This is shown (along with eye-tracker points) in Figure 3, where the higher values represent the more salient locations, as shown in Figure 2. The eye-tracker data generated by Henderson and Torralba [16] is used to generate the "saliency under task" of each test image. This can then be used to gauge the resilience of the interest-points to top down factors based on real task data. The eye tracker data gives the coordinates of the fixation points attended to by the participants. This data, collected under a search-task condition, is the "total task saliency", which is composed of both the bottom-up factors as well as the top down factors.

Task Probability Surface Construction: The three tasks used to generate the eye-tracker data were: (a) "count people", (b) "count cups" and (c) "count paintings". There are 36 street scene images, used for the people search, and 36 indoor scene images, used for both the cup and painting search. The search target was not always present in the images. A group of eight observers was used to gather the eye-tracker data for each image with an accuracy of fixation of +/- 4 pixels. (Full details in [17].)

To construct the task surfaces for all 108 search scenarios over the 72 images, the eye tracker data from all eight participants was combined into a single data vector. Then for each pixel in a mask of the same size as the image, the Euclidean distance to each eye-point was calculated and placed into ranked order. This ordered distance vector was then transformed into a value to be assigned to the pixel in the mask using the formula $P = \sum_{i=1}^{N} d_i / i^2$ in which, d is the distance to eye point, i and N is the number of fixations from all participants. The closer the pixel to an eye-point cluster, the lower the P value is assigned. When the pixel of the mask coincides with an eye-point there is a notable dip compared to all other neighbours because d_1 in the above P-formula is 0. To avoid this problem, pixels at coordinates coinciding with the eye-tracker data are exchanged for the mean value of the eight nearest neighbours, or the mean of valid neighbours at image boundary regions. The mask is then inverted and normalised to give a probabilistic task saliency map in which high intensity represents high task saliency, as shown in Figure 3. This task map is based on the ground truth of the eye-tracker data collected from the whole observer set focussing their priority on a

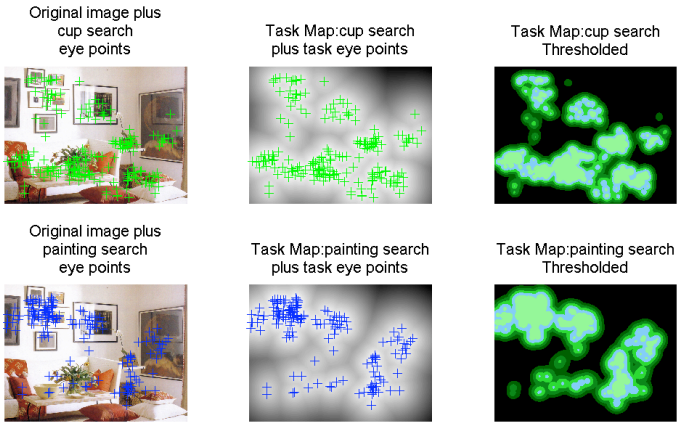

Fig. 3. Original image with two sets of eye tracking data superimposed representing two different search tasks. *Green points = cup search, Blue points =painting search*. (Centre top) Task Map derived from cup search eye-tracker data, (Centre bottom) Task Map generated from painting search eye-tracker data. (Top right) Thresholded cup search. (Bottom right) Thresholded painting search.

particular search task. It should be noted that the constructed maps are derived from a mathematically plausible probability construction (the closer the eye-point to a cluster, the higher the likelihood of attention). However, the formula does not explicitly model biological attentional tail off away from eye-point concentrations, which is a potential source of error in subsequent counts.

Interest-points vs. Saliency: The test image data set for this paper comprises 72 images and 108 search scenarios (3x36 tasks) performed by 8 observers. In doing so, the bottom-up and task maps can be directly compared. The Itti and Harel saliency models were used to generate bottom-up saliency maps for all 72 images. These are interpreted as the likely *passive* viewing eye-fixation locations. Using the method described previously, the corresponding task saliency maps were then generated for all 108 search scenarios. Finally, the interest-point detectors were applied to the 72 images (an example in Figure 1). The investigation was to determine how well the interest-points match up with each viewing scenario surface – passive viewing and search task in order to assess interest-point coincidence with visual salience. We perform a count of the inlying and out lying points of the different interest-points in both the bottom-up and task salincy maps. Each of these saliency maps are thresholded at different levels i.e. the X% *most salient* pixels of each map for each image is counted as being above threshold X and the interest-points lying within threshold are counted. This method of thresholding allows for comparison between the bottom-up and the task probability maps even though they have different underlying construction mechanisms. X = 10, 20, 30, 40 and 50% were chosen since these levels clearly represent the “more salient” half of the image to different degrees. This quantising of the saliency maps into contour-layers of equal-weighted saliency is another possible source of error in our experimental setup, although it is plausible. An

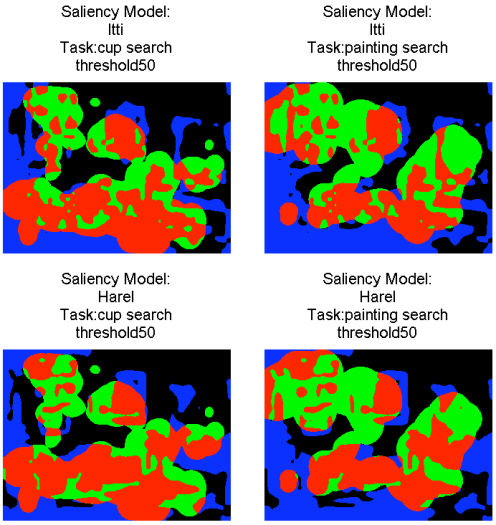

Fig. 4. An illustration of the overlap of the thresholded passive and task-directed saliency maps. Regions in neither map are in *Black*. Regions in the passive saliency map exclusively are in *Blue*. Regions exclusively in the task map *Green*. Regions in both passive and task-derived maps are in *Red*. The first row shows Itti saliency for cup search (left) and painting search (right) task data. The second row shows the same for the Harel saliency model. For Harel vs. "All Tasks" the average % coverages is Black – 30%, Blue – 20%, Green – 20%, Red – 30%, (+/- 5%). For Harel (at 50%), there is a 20% attention shift away from the bottom-up-only case due to the influence of a visual search task.

example of thresholding is shown in Figure 2. In summary, the following steps were performed:

1. The interest-points were collected for the whole image set of 72 images.
2. The Itti and Harel saliency maps were collected for the entire image set.
3. The task saliency map surfaces were calculated across the image set (36 x people search and 2 x 36 for cup and painting task on the same image set).
4. The saliency maps were thresholded to 10, 20, 30, 40 and 50% of the map areas.
5. The number of each of the interest-points lying within the thresholded saliency maps was counted.

It can be seen in Figure 1 that the interest points are generally clustered around visually "interesting" objects i.e. those which stand out from their immediate surroundings. *This paper analyses whether they coincide with measurable visual saliency*. For each image, the number of points generated by each interest point detector was limited to be equal or slightly above the total number of eye-tracker data points from all observers attending the image under task. For the 36 images with two tasks applied, the number of "cup search" task eye-points was used for this purpose.

The bottom-up models of visual saliency are illustrated in Figure 2, both in their raw map form and at the different chosen levels of thresholding. In Figure 3 the eye-tracker patterns from all eight observers are shown superimposed upon the image for

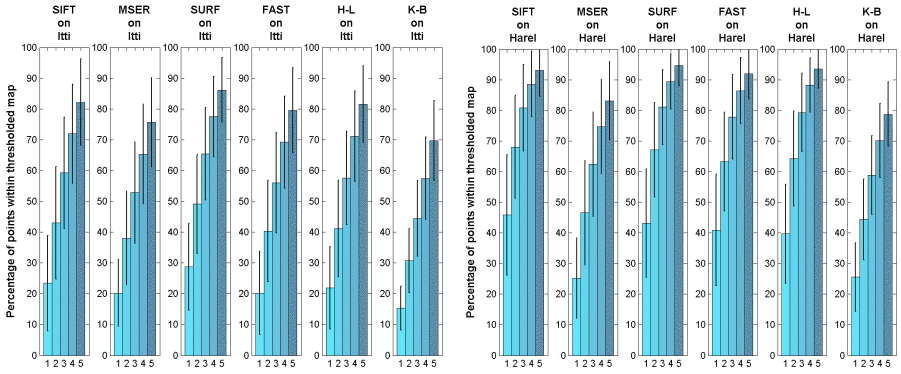

Fig. 5. The results of the *bottom up saliency map by Itti* (left) *and Harel* (right) models computed using the entire data set in comparison to the interest-point detectors. The bar indices 1 to 5 correspond to the 10 to 50 surface percentage coverage of the masks. The main axis is the percentage of interest points over the whole image set that lie within the saliency maps at the different threshold levels. The bars indicate average overlap at each threshold. Errors are gathered across the 72 image set: *standard deviation* is plotted in *black*.

two different tasks. The derived task-saliency maps are also shown, as are the task maps at different levels of thresholding. Note how changing the top down information (in this case varying the search task) alters the visual search pattern considerably. Figure 4 shows the different overlaps of the search task maps and the bottom-up saliency maps at 50% thresholding. There is a noticeable difference between the bottom-up models of passive viewing and the task maps. Note that the green-shaded pixels in these maps show where the task constraint is diverting overt attention away from the naturally/contextually/passively salient regions of the image.

3. Results and Discussion

Coincidence of Interest Points with Passive Saliency: The full count of interest-point overlap with the two models of bottom-up saliency at different surface area thresholds across the entire image set is shown in Figure 5. In comparing the interest-point overlap at the different threshold levels it is important to consider what the numbers mean in context. In this case, the chance level would correspond to a set of randomly distributed data points across the image, which would tend to the threshold level over an infinite number of images. Therefore at the thresholds in this investigation the chance levels are 10, 20, 30, 40, and 50% overlap corresponding to the threshold levels. If the distribution of interest-points is notably above (or below) chance levels, the interest-point detectors are concentrated in regions of saliency (or anti-saliency/background) and they can be considered statistical saliency detectors. Considering first the Itti model, it is clear that in general the mean percentages of data points are distributed in favour of lying within salient regions. For example, the SURF model (best performer) has 29% of SURF interest-points lying within the top 10% of ranked saliency points, 49% of SURF points distributed towards the top 20% of

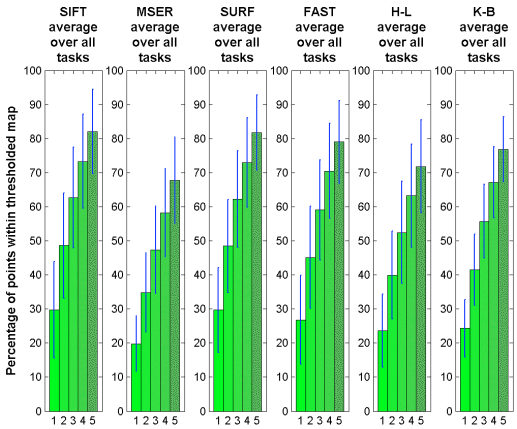

Fig. 6. The overlap of the interest-points with the *task probability surfaces* across the all 108 search scenarios. The bar indices 1 to 5 correspond to the 10 to 50 surface percentage coverage of the masks. The main axis is the percentage of interest points over the whole image set that lie within the task maps at the different threshold levels. The bars indicate average overlap at each threshold. Errors are gathered across all 108 tasks: *standard deviation* is plotted in *blue*.

saliency points and 86% of the SURF points lie within the top 50% of saliency points. Overlap with the Harel model is better than for the Itti map. This is interesting because the Harel model was found to be more robust than Itti's model in predicting eye-fixation points under passive viewing conditions. The overlap levels of the SIFT and SURF are almost identical for Harel, with 46%, 68% and 93% of SIFT points overlapping the 10%, 20% and 50% saliency thresholds, respectively. All of the values are well above mere coincidence with very strong distribution towards the salient parts of the image. They are therefore a statistical indicator of saliency. For each saliency surface class, the overlaps of SIFT, SURF, FAST and Harris-Laplace are similar while the MSER and Kadir-Brady detectors have lower overlap.

Coincidence of Interest Points with Task-Based Saliency: The interest-point overlap with levels of the thresholded task maps is illustrated in Figure 6: bottom up *and* task data is combined in Figure 7. As illustrated in Figure 4, the imposition of a task can promote some regions that are "medium" or even "marginally" salient under passive conditions to being "highly" salient under task. The interest-points remain fixed for all of the images. This section therefore needs to consider the chance overlap levels as before, but also how the attention-shift due to task-imposition impacts upon the count relative to the passive condition. The detectors are again well above chance level in all cases, with both SIFT and SURF the strongest performers, with 30%, 48% and 83% of SIFT points overlapping the 10%, 20% and 50% thresholds respectively. In the task overlap test, the Kadir-Brady detector performs at a similar level of overlap to the others - in contrast to the passive case, where it has the poorest overlap. The Kadir-Brady "information saliency" detector clearly does highlight regions that might be of interest under task, while not picking out points that are the best overlap with bottom-up models. K-B saliency is not the best performer under task and there is not enough information in this test to draw strong inference as to why this favourable shift should take place.

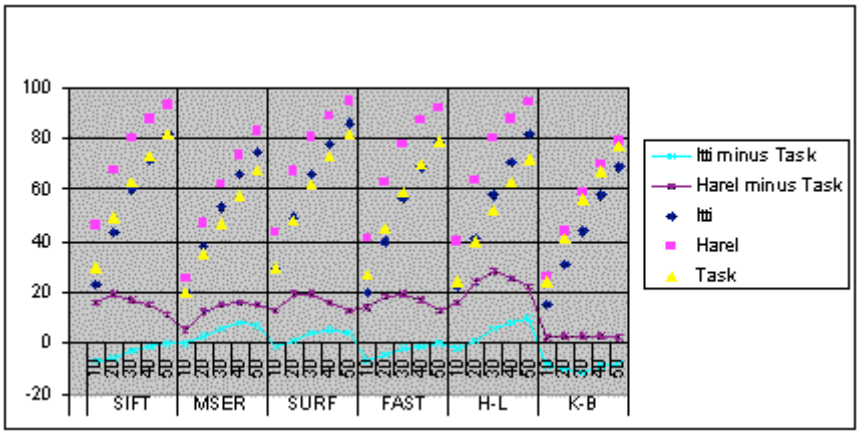

Fig 7. The *average* percentage overlaps of the interest-points at different threshold levels of the two bottom-up and the task saliency surfaces. The difference between the passive and task cases is plotted to emphasise the *overlap difference* resulting from the application of "task".

Looking at Figure 4 this should not be surprising since there exist conditions where the bottom-up and task surface overlap changes significantly: between 8% and 20% shift (*Green*, "only task" case in Figure 4) for coverage of 10% and 50% of surface area. Figure 7 reveals that the average Itti vs. interest-points overlap is overall very similar to the aggregate average task vs. interest-points overlap (between approx. +/- 7% *at most* for SIFT and SURF) implying that any attention shift due to task is directed towards other interest-points that do not overlap with the thresholded bottom-up saliency. Considering the Harel vs. task data, the task factors do reduce the surface overlap compared to the Harel surfaces by around 12% to 20% for the best performers, but very low for the Kadir-Brady. The initial high coincidence with the Harel surfaces (Figure 5) may cause this drop-off, especially since there is a task-induced shift of around 20% in some cases by the addition of a task (Figure 4).

4. Conclusion

In this paper the overlap between six well-known interest point detection schemes, two parametric models of bottom up saliency and task information derived from observer eye-tracking search experiments under task were compared. It was found that for both saliency models the SURF interest-point detector generated the highest coincidence with saliency. The SURF algorithm is based on similar techniques to the SIFT algorithm, but seeks to optimize the detection and descriptor parts using the best of available techniques. SIFT's Gaussian filters for scale representation are approximated using box filters and a fast Hessian detector is used in the case of SURF. Interestingly, the overlap performance was *superior* for the supposedly more robust saliency model for passive viewing, Graph Based Visual Saliency by Harel *et al*. Interest-points coinciding with bottom-up visually-salient information are valuable because of the robust description that can be applied to them for scene matching.

However, under task the attentional guidance surface is shifted in an unpredictable way. Even though statistical coincidence between Interest-points and the task surface remain well above chance levels, there is still no way of knowing what is being shifted where. The comparison of Kadir-Brady information-theoretic saliency with verified passive visual saliency models shows that Kadir-Brady is not in fact imitating the mechnisms of the human visual system, although it does pick out task-relevant pieces of information at the same level as other detectors.

References

1. Lowe, D. G.: Distinctive Image Features from Scale-Invariant Interest points. In: International Journal of Computer Vision, Vol. 60, pp. 91-110 (2004).
2. Matas, J., Chum, O., Urban, M. and Pajdla, T.: Robust wide baseline sterio from maximally stable extremal regions. In: Proc. of British Machine Vision Conference, pp. 384-393 (2002).
3. Mikolajczk, K. and Schmid, C.: An Affine Invariant Interest Point Detector. In: Proc. European Conference on Computer Vision, pp. 257-263. Springer (2003).
4. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded Up Robust Features. In: Proc. 9th European Conference on Computer Vision, LCNS, Volume 3951/2006, pp. 404-417 (2006).
5. Rosten, E and Drummond, T.: Fusing points and lines for high performance tracking. In: 10th IEEE International Conference on Computer Vision, Vol. 2, pp. 1508-1511, (2005).
6. Rosten, E and Drummond, T.: Machine learning for high-speed corner detection. In: Proc 9th European Conference on Computer Vision, LNCS, Part I, pp. 430-443 (2006).
7. Kadir, T and Brady, M.: Saliency, Scale and Image Description, In: Int Journ. Comp. Vision. 45 (2), pp. 83-105 (2001).
8. K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir and L. Van Gool, "A comparison of affine region detectors." In: Int. Journ. Comp. Vision 65(1/2), pp. 43-72 (2005).
9. Mikolajczyk, K. and Schmid, C.: A performance evaluation of local descriptors. In: IEEE Transactions on Pattern Analysis & Machine Intelligence, Volume 27, Number 10, pp. 1615-1630 (2005).
10. Gao, K., Lin, S., Zhang, Y., Tang, S., Ren, H.: Attention Model Based SIFT Keypoints Filtration for Image Retrieval. In: Proc. ICIS 2008, Vol 00, pp. 191-196 (2008)
11. Itti, L., Koch, C., Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 11, pp. 1254-1259 (1998).
12. Harel, J., Koch, C. and Perona, P.: Graph-Based Visual Saliency. In: Advances in Neural Information Processing Systems 19, pp. 545--552 (2006).
13. Navalpakkam, V., Itti, L.: Search goal tunes visual features optimally. In: Neuron, Vol. 53, No. 4, pp. 605-617 (2007).
14. Navalpakkam, V., Itti, L.: Modeling the influence of task on attention. In: Vision Research, Vol. 45, No. 2, pp. 205-231 (2005).
15. Peters, R. J., Itti, L.: Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8 (2007).
16. Torralba, A., Oliva, A., Castelhano, M. and Henderson, J. M.: Contextual Guidance of Attention in Natural scenes: The role of Global features on object search. In: Psychological Review. Vol. 113(4), pp. 766-786 (2006).