Evaluating local feature detectors in salient region detection

Teemu Kinnunen[1], Mari Laine-Hernandez[1], and Pirkko Oittinen[1]

Department of Media Technology
Aalto University
Espoo, Finland

Abstract. In this work, we study local feature extraction methods and evaluate their performance in detecting local features from the salient regions of images. In order to measure the detectors' performance, we compared the detected regions to gaze fixations obtained from the eye movement recordings of human participants viewing two types of images: natural images (photographs) and abstract/surreal images. The results indicate that all of the six evaluated local feature detectors perform clearly above chance level. The Hessian-Affine detector performs the best and almost reaches the performance level of state-of-the-art saliency detection methods.

1 Introduction

People can recognise thousands of objects quickly and accurately [3]. The Human Visual System (HVS) is capable of processing tremendous amounts of information. However, only a fraction of the information is important. Thus, the HVS functions so that the focus of visual attention can be moved quickly to detect important things. Visual attention is controlled by both top-down and bottom-up processes. Top-down processing is determined by high-level context factors, such as the task that the observer is performing, as well as the semantic contents of the scene. In bottom-up processing, visual attention shifts from one location to another based on low-level features which "pop up" from the scene. This "pop up" effect is called saliency [10] and it is based on the distinctiveness of an object from its surround regarding intensity, color and orientation. In other words, salient regions attract more attention because of their dissimilarity with their surroundings [4].

In addition to being able to shift attention quickly, people are also far superior to computers in Visual Object Categorisation (VOC). VOC has attracted a significant amount of interest during the last decade. New methods have been introduced and the performance of the state-of-the-art methods has been improved. Most of the best performing VOC methods are based on local features [6]. Local feature detectors should be able to detect the same locations from a scene regardless to various transformations and errors introduced to the images. In this study, local feature detectors were used (see Fig. 1) in order to see if there

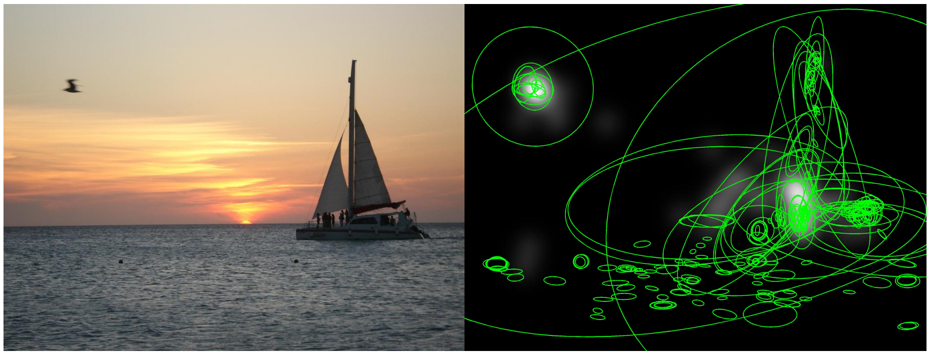

Fig. 1: Local feature detection and saliency. Left: Original image from Judd et al. [11] image set; Right: Ground truth saliency map and 10% of the detected local features (randomly chosen).

is a connection between the performance of local feature detection methods in a VOC task and their capability for saliency detection.

The contributions of this work are the following:

- A new method for comparing the regions of local features with fixations from human participants.
- Comparison of the performance of local feature detectors in saliency detection with two dissimilar image sets (natural images and abstract/surreal images).
- Experimental results which indicate that the local feature detectors detect local features from salient regions clearly more accurately than random chance and almost achieve the state-of-the-art.

1.1 Background

A local feature is a combination of a spatial location and a description of the visual appearance of a local patch in an image. There are two steps in local feature extraction: i) detection, and ii) description. These descriptions of image patches are typically invariant to transforms such as rotation, scale, location and illumination. In this work, we study local feature detection and compare detection outputs with fixations obtained from human participants. We benchmark some of the most popular and best performing local feature detectors such as DoG (SIFT) [14], MSER [15], Harris-Laplace (HarLap) [17], Harris-Affine (HarAff) [18], Hessian-Laplace (HesLap) [19] and Hessian-Affine (HesAff) [18].

Local features are used in many computer vision applications such as specific object detection [14], content based image retrieval (CBIR) [5], wide-baseline matching for stereo images [15], texture recognition and object recognition [21]. There are many studies where different features of the local feature detectors are compared and their performances in various tasks are evaluated [16, 21]. In [21], a scale invariant version of the Harris detector performed the best in the VOC experiment with natural images. However, the Hessian-Laplace detector and its affine invariant version were missing from the comparison. In [16], MSER and Hessian-Affine detectors perform the best in repeatability experiments. We compare the results of the experiments conducted in this work with the results in visual object categorisation accuracy achieved in earlier studies [16, 21].

Even though some of the local feature detectors are claimed to work using the same principle as the HVS [14], local feature detectors and their ability to detect features from salient regions have not been studied until very recently.

The first study where local feature detectors were compared with saliency information was published by Harding and Robertson [8]. These authors compared six interest point detectors with two bottom-up saliency models and eye-tracking data. In their experiment, they compared local feature detectors and saliency detectors directly by computing overlaps of "predicted" saliency maps. In saliency detection research, it is more common to try to predict where humans fixate on an image, instead of computing overlaps of predicted and ground truth saliency maps. We performed a similar experiment, but instead of computing overlaps, we computed how well the method can predict ground truth human fixations as has been proposed earlier e.g. in [11].

Harding and Robertson used three different methods to obtain ground truth saliency maps to evaluate local feature detectors. The first method was to combine fixations from human participants who were performing tasks such as counting the number of people in the images, finding cups in the images etc. In this case, the top-down processes of visual attention are influenced by the given task. On the other hand, local feature detectors use only low level information, and thus, the local feature detectors are not capable of detecting features from these saliency maps as was found by Harding and Robertson. Therefore, we consider human fixation data from free viewing experiments, such as [11, 13] to be more suitable. The second method was based on the standard saliency detection method by Itti and Koch [10] and the third method was based on a more sophisticated saliency detection method by Harel [9]. However, Judd et al. [11] found out that the Itti and Koch and Harel detectors do not detect salient regions accurately enough to be considered as ground truth.

The second study where local feature detectors were compared with human fixation data was published by Akshat et al. [1]. They studied if fixation locations obtained from the eye movements of human observers and interest points captured using local feature detectors match spatially. Akshat et al. used two methods to compare local feature detectors and human fixations. The first method globally compared distributions of spatial locations of detected features and fixations. At first, spatial locations were smoothed by using a kernel density estimation method. Then the difference between the two distributions was computed using the Bray-Curtis similarity method. The second method compared the detected local features and fixations from human participants locally. At first, they randomly selected a set of local features and then used the locations of these features to predict fixations. They chose different numbers of local features in order to compute ROC curves. Akshat et al. used the MSER, SIFT and SURF [2] detectors in their experiments. These detectors did not, however, perform the best in the visual object categorisation task where the goal is to predict which objects exist in a given image [6, 21]. Therefore, we want to repeat the experiment with local feature detectors which have been more successful in the VOC task such as Harris-Laplace [17], Hessian-Laplace [19] and affine-invariant versions of them.

According to Akshat et al.'s experiment results from the global evaluation (Bray-Curtis similarity) method, randomly chosen edge points and SURF detectors performed better than the randomly selected points and the rest of the local features performed equally or worse. According to the results from the local evaluation (ROC) method, all of the detectors performed worse than randomly selected points. In our work, we repeat the experiment with the same set of images (Natural image set by Judd et al. [11]) and use the state-of-the-art local feature detectors, but we use our own approach to evaluate performance. In addition to the spatial locations, our approach also takes into account scale (and possible affine parameters). We also use the same method to evaluate predicted saliency maps which has been used in [11] to evaluate the performance of saliency detectors.

In addition to the comparison of local feature detectors in saliency detection, we compare these "saliency detectors" to the state-of-the-art detector by Judd et al. [11]. Judd et al.'s model of saliency was developed by training a classifier directly from human eye movement data. It is based on low, middle and high-level image features. The low-level features include color, intensity and orientation features; mid-level consists of a horizon line detector, and as high-level features they use the Viola&Jones face detector [20] and the Felzenszwalb car and person detectors [7]. Because of the well-known central bias in image viewing (objects of interest are typically located near the centre of the image), Judd et al. also added a central bias to their model.

2 Saliency map generation from local feature detections

We compared the local feature detectors by calculating how large of a portion of the salient ground truth region of each of them captures. Detected regions were converted into saliency maps as presented in Algorithm 1.

The algorithm takes as inputs the regions of the detected local features $\mathbf{L}$ and an input image $\mathbf{I}$. At first, the saliency map $\mathbf{A}$ value of each pixel is set to zero. Then, the saliency values of the pixels, $\boldsymbol{x} = (x, y)$, belonging to the region i of detected local features, are increased by one. This is repeated for every detected local feature. Finally, the saliency map $\mathbf{A}$ is normalised by dividing the saliency values by their sum.

Fig. 2 shows how the saliency maps generated using different local feature detectors differ from each other. The Harris-Laplace and the Harris-Affine detect the highest number of local features, and thus, the saliency maps are covered with salient pixels. Additionally, Hessian-Laplace and Hessian-Affine detectors detect

Algorithm 1 Generate saliency maps from regions of local features

```
Require: $\mathbf{L}, \mathbf{I}$
  {Initialise saliency map with zeroes}
  $\mathbf{A}_{1,\ldots,width(\mathbf{I}),\ 1,\ldots,height(\mathbf{I})} \leftarrow 0$
  for $i = 1, \ldots, numberOfFeatures(\mathbf{L})$ do
    {Select all the indexes of pixels belonging to region $\mathbf{L}_i$ and store them in $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ where $\boldsymbol{x}_j = (x, y)$}
    $\mathbf{x} \leftarrow getPixelsOfRegion(\mathbf{L}_i)$
    for $j = 1, \ldots, length(\mathbf{x})$ do
      $\mathbf{A}_{x_j,y_j} \leftarrow \mathbf{A}_{x_j,y_j} + 1$
    end for
  end for
  $\mathbf{A} \leftarrow \mathbf{A}/sum(\mathbf{A})$
  return $\mathbf{A}$
```

a large number of features. However, saliency maps generated from Hessian-Laplace and Hessian-Affine local features seem to also cover more non-salient areas (i.e. black pixels).

3 Experiments

In salient region detection, the goal is to predict the saliency of each pixel, i.e. define how much attention each pixel attracts from human observers. To evaluate local feature detection methods in saliency detection, we followed the procedure introduced in [11], where predicted saliency maps are compared with fixations gathered from human participants and the performance is reported as a ROC curve. The ROC curve is computed by choosing $p_x\%$ (*Percent Salient*) of the most salient pixels from the predicted saliency map. True positive rate can then be computed by dividing the number of fixation points inside the thresholded area by the total number of fixation points. This is repeated with p_x = 1%, 3%, 5%, 10%, 15%, 20%, 25%, and 30% in order to obtain the ROC curve.

We compared saliency maps predicted using local feature detectors with inter-subject, central bias and the current state-of-the-art method by Judd et al. [11, 12]. The central bias saliency defines the saliency of a pixel as the inverse of the distance from the centre of the image, i.e. pixels near the centre of the image are more salient than pixels near the borders of the image. "Inter-subject" is a measure of inter-observer agreement, i.e. the congruency of the human attention maps generated from gaze fixations. It was calculated by forming an attention map for each individual participant and image, and comparing this to the attention map derived from the fixation locations of all other participants for the same image. An ROC area was computed for each participant, and the average of the values was taken as the inter-subject ROC value. As the final result, we computed an average (over the images) ROC value for each method.

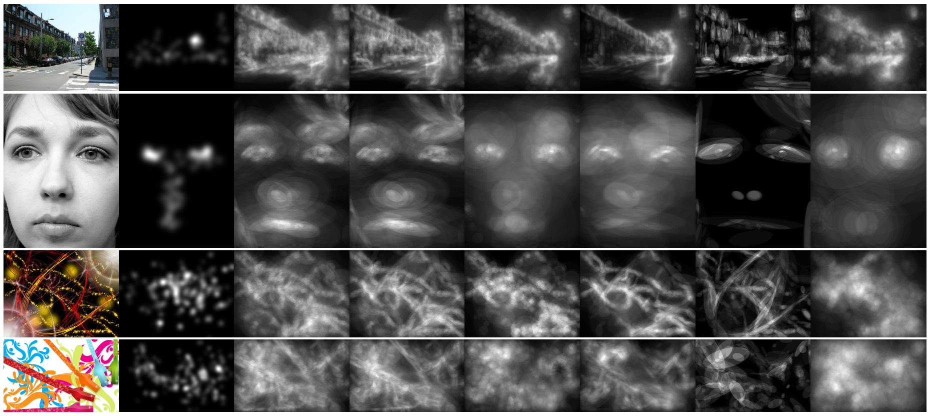

Fig. 2: Outputs of the saliency maps using local feature detectors. From the left: Original image, ground truth, Harris-Laplace, Harris-Affine, Hessian-Laplace, Hessian-Affine, MSER and SIFT (DoG). The top two images are from the data set introduced by Judd et al. [11] and the bottom two images are from the data set by Laine-Hernandez et al. [13].

3.1 Experiment 1: Natural image set

In the first experiment, we used the dataset gathered by Judd et al. [11] which contains 100 natural images and eye movement data from 15 participants for the images. The participants' task in the eye-tracking experiment was to free-view the images for 3 seconds each. Fig. 3 shows that saliency detectors based on Hessian-Laplace and Hessian-Affine local feature detectors perform the best. The DoG-detector used in SIFT performs slightly worse than the Hessian-Laplace and Hessian-Affine detectors. The Harris-Affine and Harris-Laplace local feature detectors perform slightly worse than the Hessian-Laplace and Hessian-Affine and DoG detectors. MSER does not perform as well as the Hessian-Laplace and Hessian-Affine, DoG and Harris-Laplace and Harris-Affine detectors. However, all the detectors are far from the inter-subject performance and even behind the centre biased saliency.

The state-of-the-art detector by Judd et al. [11] performs the best and is better than any of the methods based on local feature detectors. However, the difference is not as large as one might assume based on the previous study by Akshat et al. [1]. All the local feature detectors used in this study performed clearly above chance level.

3.2 Experiment 2: Abstract image set

In the second experiment we used a set of 100 abstract/surreal images with eye-tracking data from 12 participants per image. The participants' task in the eye-tracking experiment was to free-view the images for 5 seconds each. [13]. We used the same set of methods to predict salient regions from the given images as in Experiment 1. The results of the experiment are shown in Fig. 4.

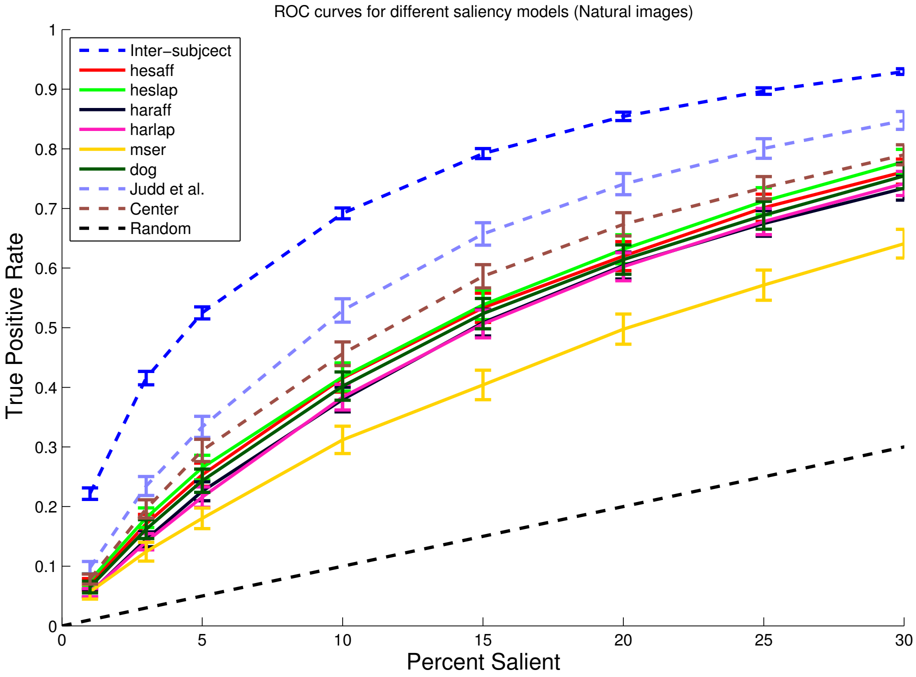

Fig. 3: ROC curves with standard error bars for saliency detection for Natural image set.

According to the results presented in Fig. 4, we can see that the saliency detector based on the Hessian-Laplace local feature detector performs the best. Hessian-Affine, Harris-Laplace, Harris-Affine and SIFT (DoG) detectors perform equally and the MSER detector slightly worse than the other detectors.

The state-of-the-art detector by Judd et al. [11] performs better than the Hessian-Laplace based detector and the central biased saliency.

3.3 Summary of results

In this work, we made two experiments; 1) with Natural images; and 2) with Abstract/surreal images. These results are summarised in Table 1. In general, the results were aligned in both experiment, but there were some differences. For example, in Experiment 2 with Abstract Images, saliency maps generated based on central bias did not predict the locations of human fixations as accurately as maps generated from regions detected using the Hessian-Laplace and Hessian-Affine detectors. On the other hand, in Experiment 1 (Natural Images), maps generated based on central bias predicted human fixations more accurately than any of the methods based on local feature detectors. We can also notice that the AUCs for saliency detectors, inter-subject and central biased saliency are on average 0.07 less with Abstract images than with Natural images.

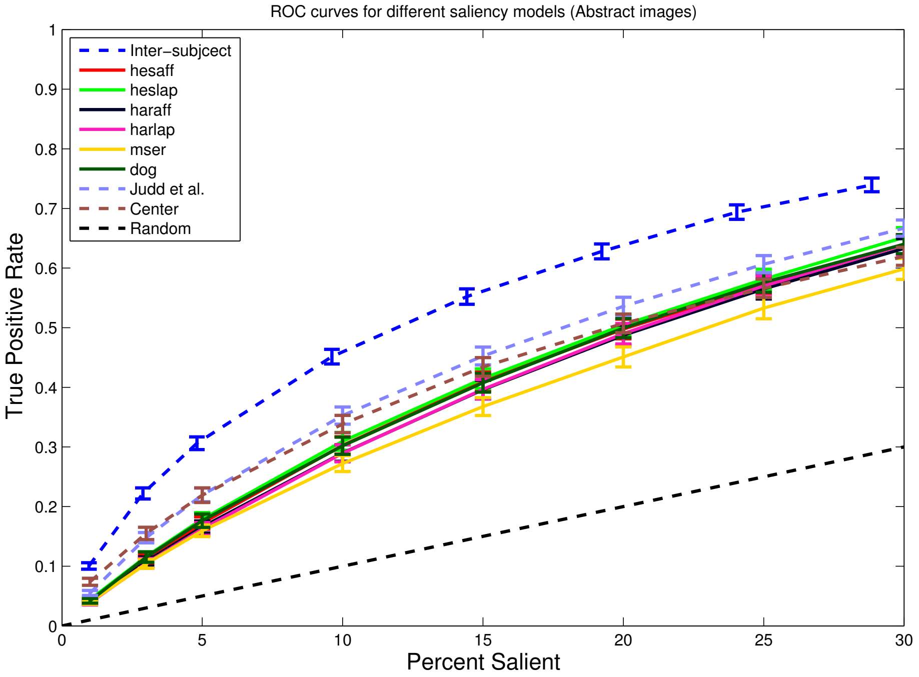

Fig. 4: ROC curves with standard error bars for saliency detection for Abstract image set.

4 Discussion

According to the results presented in the previous section, we can see that the saliency maps generated using the Hessian-Laplace and Hessian-Affine and Harris-Laplace and Harris-Affine detectors and SIFT (DoG) detectors perform almost equally, Hessian-Laplace being slightly better than the others. The MSER detector detects fewest local features from the salient regions, i.e. regions where most of the fixations from human observers are located.

We also noticed that the state-of-the-art detector by Judd et al. [11] performs better than any of the detectors based on a local feature detector, but the difference is surprisingly small. The difference between the state-of-the-art method by Judd et al. and local feature detectors is smaller for the Abstract image set than for the Natural image set. One of the reasons might be related to the fact that the central bias is smaller for the Abstract images than for Natural images, which is evident from Table 1, where means of AUCs from the experiments are compared.

When comparing the results of this study with the one presented by Akshat et al. [1], we can notice an interesting contradiction. Akshat et al. found that local feature detectors perform equally or worse than random selection of interest points. Moreover, they found only a small correlation between the local feature

Table 1: Mean AUC and standard errors for both experiment

Method	**Natural images** [11]	**Abstract images** [13]
Inter-subject	0.8904 (0.0035)	0.7757 (0.0065)
Central bias	0.7842 (0.0103)	0.6865 (0.0090)
Judd et al. [11]	0.8221 (0.0095)	0.7090 (0.0095)
HesAff	0.7625 (0.0129)	0.6880 (0.0095)
HesLap	**0.7708** (0.0130)	**0.6941** (0.0096)
HarAff	0.7452 (0.0122)	0.6828 (0.0097)
HarLap	0.7479 (0.0124)	0.6852 (0.0098)
MSER	0.6892 (0.0140)	0.6639 (0.0100)
SIFT (DoG)	0.7571 (0.0130)	0.6884 (0.0094)

detectors (i.e. interest point detectors) and human fixations. On the other hand, we found a clear relation between the regions of detected local features and human fixations. Our results also indicate that local feature detectors can predict locations of fixations clearly better than random selection. The reason behind the different results is that in this work we converted detected regions into predicted attention maps instead of using only the spatial location of the centre of the detected feature. We claim that our approach is more justified than the method presented earlier, because we take the scale of the detected region into account and use all the detected regions of local features instead of randomly chosen features as in the original study by Akshat et al. In addition, we used local feature detectors that have been succesful in various visual object categorisation studies [21, 16] instead of detectors that have performed worse in the same studies.

5 Summary

In this work, we studied local feature detection and saliency. We carried out two experiments where we compared detected regions and fixations obtained from human participants.

We found out that the local feature detectors can detect local features from salient regions clearly better than random selection. This result contradicts an earlier study by Akshat et al. [1] where it was found that detectors performed worse or equal to random selection. We used a different method to compare local feature detectors, one that is based on earlier studies for saliency predictors. We consider our approach to be more justified than the previous approach, because we take all the detected regions into account and we also consider the scale of the local feature instead of only its origin.

In the future, we are going to study also the descriptor part of the local features and explore if there exists such a thing as a “salient local feature”.

References

1. Akshat, D., Rachit, D., Bernard, G.: Do Humans Fixate on Interest Points? In: Proc. of International Conference on Pattern Recognition. Tsukuba Science City,

Japan (2012)

2. Bay, H., Tuytelaars, T., Gool, L.: Surf: Speeded up robust features. In: Proc. of European Conference on Computer Vision. pp. 404–417 (2006)
3. Biederman, I.: Recognition-by-components: A theory of human image understanding. Psychological Review 94(2), 115–147 (1987)
4. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(1), 185–207 (2013)
5. Chum, O., Mikulik, A., Perdoch, M., Matas, J.: Total recall ii: Query expansion revisited. In: Proc. of Computer Vision and Pattern Recognition. pp. 889 –896 (2011)
6. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes (VOC) challenge. International Journal of Computer Vision 88(2), 303–338 (2010)
7. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: Proc. of Computer Vision and Pattern Recognition (2008)
8. Harding, P., Robertson, N.: A comparison of feature detectors with passive and task-based visual saliency. In: Scandinavian Conference on Image Analysis, Lecture Notes in Computer Science, vol. 5575, pp. 716–725. Springer Berlin Heidelberg (2009)
9. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Proc. of Neural Information Processing Systems. pp. 545–552. MIT Press (2007)
10. Itti, L., Koch, C.: Computational modelling of visual attention. Nature Reviews Neuroscience 2(3), 194–203 (2001)
11. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: Proc. of International Conference on Computer Vision (2009)
12. Judd, T., Durand, F.d., Torralba, A.: A Benchmark of Computational Models of Saliency to Predict Human Fixations. Tech. rep. (2012)
13. Laine-Hernandez, M., Kinnunen, T., Kamarainen, J.K., Lensu, L., Kälviäinen, H., Oittinen, P.: Visual Saliency and Categorisation of Abstract Images. In: Proc. of International Conference on Pattern Recognition. Tsukuba Science City, Japan (2012)
14. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 20, 91–110 (2004)
15. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. In: Proc. of British Machine Vision Conference. pp. 384–393 (2002)
16. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. International Journal of Computer Vision 65(1/2), 43–72 (2005)
17. Mikolajczyk, K., Schmid, C.: Indexing based on scale invariant interest points. In: Proc. of International Conference on Computer Vision. pp. 525–531 (2001)
18. Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. In: Proc. of European Conference on Computer Vision. pp. 128–142 (2002)
19. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International Journal of Computer Vision 60, 63–86 (2004)
20. Viola, P., Jones, M.: Robust real time object detection. International Journal of Computer Vision (2001)
21. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision 73(2) (2007)