Salient image regions as a guide for useful visual features

Hannes Saal[1], Nora Nortmann[1], Norbert Krüger[2] and Peter König[1]
[1]Institute of Cognitive Science, University of Osnabrück,
Osnabrück, Germany
hsaal@uos.de nnortman@uos.de pkoenig@uos.de
[2]The Maersk Mc-Kinney Moller Institute, Syddansk University,
Odense, Denmark.
nk@imi.aau.dk

Abstract - *Feature selection plays a crucial part for computer vision algorithms. By examining properties of fixation points in natural images collected via eye-tracking, we take the human visual system as a model. Here we make use of intrinsic dimensionality (iD): Image patches with high i0D values correspond to homogeneous surfaces, whereas patches with high i1D values contain edges, lines or parallel lines. A high i2D value corresponds to corners, junctions, and the like. A comparison of mean iD feature values at actual fixations points with those at control points shows significant effects: Mean i0D values are lower for actual than for control, while mean i1D and i2D values are higher. A subsequent information-theoretic analysis shows entropy to be higher at actual than at control points. Mutual information between successive fixation points is found to be significantly lower for actual than control sequences. This effect stems mainly from short saccades, as is revealed by calculation of mutual information from saccades falling into different length ranges. As it turns out, the human visual system prefers scarce features like corners or junctions over common ones like homogeneous surfaces. iD provides an effective computational means to extract and classify such features.*

Keywords: Computer Vision, Eye-Tracking, Information Theory

1 Introduction

Computer vision aims at developing algorithms that are capable of giving concise descriptions and interpretations of images. The selection of image features plays a crucial part in this process and turns out to be a key part when moving from low-level to higher-level abstract descriptions.. In previous studies [5] we defined local image properties via their intrinsic dimensionality (iD), which gives a continuous estimation of the numbers of dimensions that would be needed to describe an image patch properly.

In this study, we use the human visual system as a model for feature extraction by examining properties of fixation points in natural images. We hypothesize that useful and important features are frequently fixated by human subjects and that iD may indicate salient image regions. As the human visual system is restricted to repeatedly sampling small image regions in order to study a particular visual scene, fixation points have to be chosen wisely to maximize the amount of incoming new information with each fixation [cf. 6]. Thus, information theory and algorithms originating in machine vision may prove to be a useful tool when examining the human visual system. As human vision as well as computer vision systems face similar problems, we use methods originating in machine vision to help assess fixation point selection strategies as done by humans.

Correlation of certain simple image features like luminance contrast [8] with fixation is firmly established. While multiple such features (e.g. color, occurrences of edges, disparity) normally are used to create a saliency map [2], that is a map that highlights interesting regions for the observer to have a closer look at, we rely on a single unified method to extract a multi-dimensional feature from image data.

2 Methods

The following section gives a more detailed background on intrinsic dimensionality, then explains our eye-tracking setup and the correlational as well as information-theoretic analyses applied to the data in more detail.

2.1 Intrinsic Dimensionality

Intrinsic dimensionality is computed from local image gradient and orientation measures [12]. Properties of image regions are mapped into a triangular-shaped space. The corners of the iD triangle refer to extreme values regarding the dimensionality of the image region in question: Patches with high i0D values correspond to homogeneous surfaces, whereas patches with high i1D values contain edges or parallel lines. A high i2D value corresponds to corners, junctions, and the like. (see Figure 1 for some example image patches as well as their position within the iD triangle). Projections of iD values onto the triangle's axes indicate the contributions from each corner point and thereby give confidences as to whether a specific image region contains flat surfaces, edges or corners. Applying the iD algorithm to an image leads to iD maps, as depicted in Figure 2.

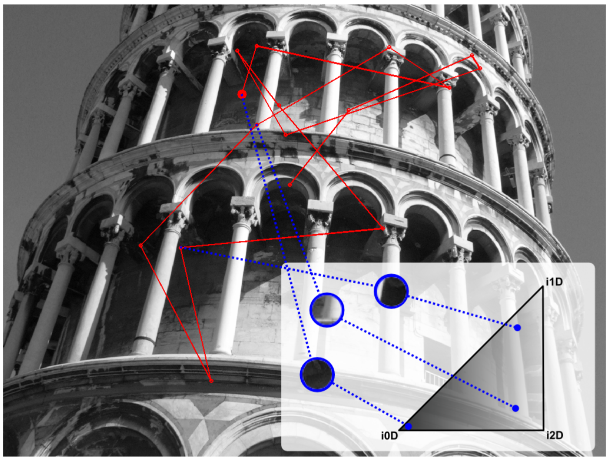

Figure 1. Sample stimulus picture showing a fixation trajectory. Three fixated regions have been enlarged and their iD values are indicated in the iD triangle.

Our implementation of the iD algorithm [5] makes use of the structure tensor, which is well-known in image processing and provides measures of gradient and orientation variances. Mathematically, the structure tensor is a 2x2 matrix and computed as the outer product of the image gradient matrix (containing both the gradients from x and y axis as row vectors) with itself. The structure tensor can be interpreted by examining its eigensystem. The larger of its two eigenvalues indicates the strength of the prominent orientation in the patch, with its corresponding eigenvector pointing in that direction. The second eigenvalue gives the strength orthogonal to that direction. From the eigenvalues both orientation and gradient variances can be computed. While the gradient variance provides a measure of image patch homogeneousness, the orientation variance distinguishes between lines oriented in the same direction and lines directed in multiple directions. Combining these two measures and normalizing them yields the aforementioned iD triangle.

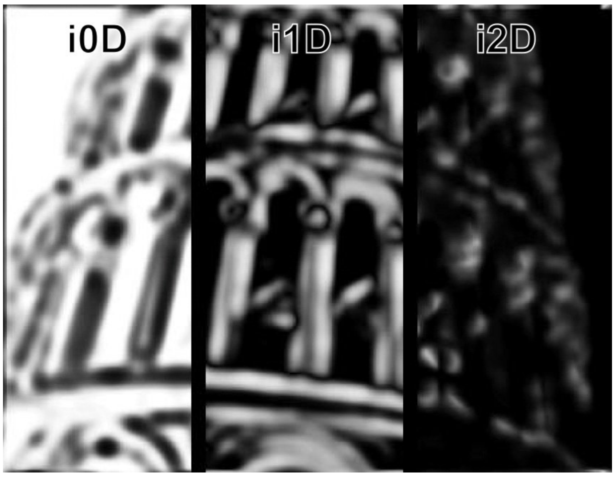

Figure 2: *ix*D maps of picture shown in Figure 1. Brighter pixels denote higher *ix*D values respectively.

2.2 Eye-Tracking

We use eye-tracking on free-viewing subjects. A total of 64 gray-scale images are shown to 14 subjects in randomized order for 5 seconds each, while eye fixations are recorded. Images come from one of four categories: Fractals (mostly Mandelbrot sets), man-made objects (all outside, different size ranges from whole buildings to smaller objects), natural scenes (trees, bushes, leaves at different scales) and faces (frontal against wall, all roughly same size and shape). See Figure 3 for thumbnails. Images were displayed at a screen resolution of 1024x768 pixels on a 21 inch monitor. A picture as shown on the screen spanned 28.16° horizontally and 21.12° vertically of a subject's visual field. Control fixation points as later used in the analysis are generated by shuffling complete participants' eye-traces over images regardless of image category membership when the whole image set is analyzed, while eye-traces are shuffled within categories when single image categories are examined. Fixation points lying near the image border are discarded from later analysis to avoid border effects. Mean saccade length was 148px or 4° of the visual field. Short saccades are abundant, with longer ones becoming increasingly rare. The longest saccade that was included in the analysis spanned 1083px.

Stimulus pictures are converted into iD feature maps for analysis. For noise reduction purposes images are low-pass filtered before the iD algorithm is applied. The Gaussian window function that is used when computing the structure tensor for selecting image sub-parts has a standard deviation of 0.28° of the subjects' visual field. Averaging iD values from their feature maps over all images shows a mean value of 67% for i0D projections, 22% for i1D and 11% for i2D. Junctions and corners are therefore the rarest features, edges occur twice as much, and flat surfaces dominate heavily.

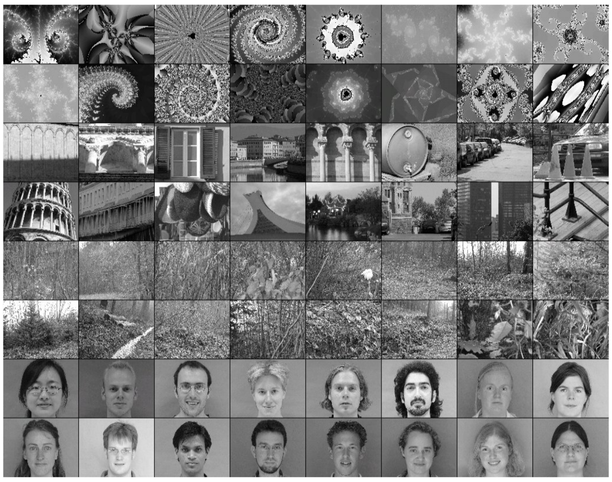

Figure 3: Gray-scale pictures that were used as stimuli. Upper two rows show fractals, then man-made objects, natural scenes and last two rows depict faces.

2.3 Correlation of features and fixation

Correlation of features and fixations is determined by applying two-sided sign tests for single iD projections, comparing values of actual and control condition. We also computed the percentage of actual values exceeding their control ones. A total of 12,972 fixation points (all subjects, whole image set) are used, i*x*D values are averaged within each of the three iD projections over subjects and images. Results are calculated category-wise as well, with each image category incorporating 16 images and roughly 4,000 fixations each.

When extracting single features from an image like the occurrences of edges or luminance contrast, usually their absolute values are taken to build or contribute to a saliency map [2]. As iD is a multi-dimensional feature and therefore there doesn't exist one single feature value this is not possible, so we use a different conversion scheme to obtain a saliency map. The concept of self-information provides a way to perform this transformation that is backed up by neuropsychological data as well. The concept of self-information has successfully been employed as a transformation yielding salient image regions. Topper [10] showed that saliency maps based on self-information transformations of image feature maps that are known to correlate with fixation give better results in terms of similarity with actual fixation point densities than do absolute feature values. Self-information is higher for rare events than for common ones. Hence, the human visual system seems to respond to surprising and unusual visual events rather than just high absolute values of certain features.

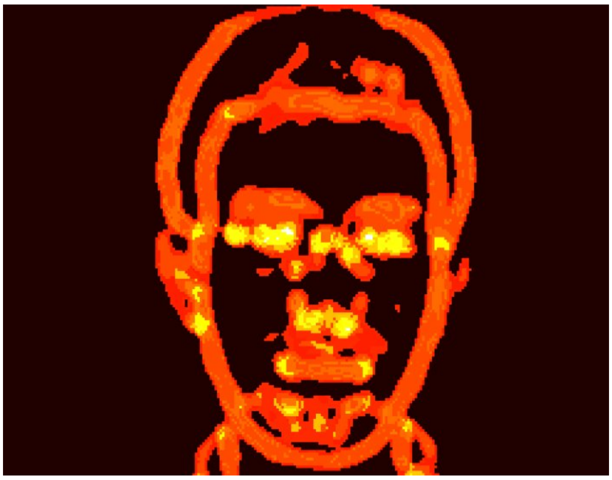

Figure 4: Computed self-information map of one of the stimulus pictures in the faces category. Brighter pixels denote higher self-information and therefore higher salience.

We use a local (picture-wise) approach to compute the self-information of a specific pixel within an iD map. To estimate the probability distribution of iD values, feature map entries are placed in 100 equally sized bins, with their relative frequency being used as an estimate of probability. Self-information values at actual fixation points are compared with control ones by applying two-sided sign tests. See Figure 4 for an example of a self-information map.

2.4 Information-theoretic Analysis

Entropy of actual and control conditions is calculated by allocating i*x*D values into 100 bins respectively. 150 different control permutations of eye-traces are generated and results deemed significant when the actual value exceeds or falls below *all* control values. These analyses were carried out for the whole image set as well as for single image categories independently.

An analysis of mutual information values between consecutive fixation points is carried out in order to examine (first-order) dependencies of newly selected regions to previous ones. Hence, sequences of two immediately consecutive fixation points are extracted and mutual information between first and second points is calculated. We use 11,860 sequences in this analysis. iD values are allocated to ten bins only. In a later analysis, fixation sequences are grouped together according to the saccade length between the two fixation points and mutual information is calculated separately. Each of the resulting five groups that are used here contains the same number of fixation sequences. Group boundaries are 56, 105, 166, 248 and 434 pixels of saccade length. As short saccades are prevalent saccade length ranges become bigger for longer saccades to ensure an equal number of sequences. This analysis is carried out for the whole image set only.

3 Results

In the following section results from the correlational and information-theoretic analyses are presented.

3.1 Correlation of features and fixation

Comparing iD values of actual and control fixation points shows i0D values to be lower, but i1D and i2D values to be higher at actual fixation points. All results are highly significant (P<0.0001). i0D projections show the biggest effect size, with only 27.8% of actual values being higher than the control values. i1D values are higher than the control one in 69.1% of cases, and i2D in 70.3%. This indicates that while image regions consisting only of flat surfaces are avoided, those with higher intrinsic dimensionality are fixated more often than would be expected from chance. Figure 5 shows the results of this analysis within the iD triangle as well as actual vs control values for the i2D projection (inset).

The results from examining single image categories mainly underline the effect found when analyzing the whole picture set, differing slightly in effect sizes. It is worth noting that most pronounced effects are obtained for the picture categories man-made objects (P<0.0011 for all projections), faces (P<0.0001 for all), and fractals (P<0.0001 for all). Natural scenes show significant results for the i0D projection only (P=0.0268), but not for the i1D (P=0.0516) and i2D (P=0.3141) ones. While faces and man-made objects show the highest amounts of actual values exceeding control ones in case of i1D (64.4%

and 61.2%) and i2D (73.0% and 72.2%) projections and the lowest amount surpassing them in the i0D (27.0% and 26.8%) projection respectively, results for the fractals category are somewhat less pronounced (32.9% i0D, 59.0% i1D, 66.7% i2D). Effect sizes for the natural scenes category are by way the smallest (42.3% i0D, 56.8% i1D, 53.6% i2D).

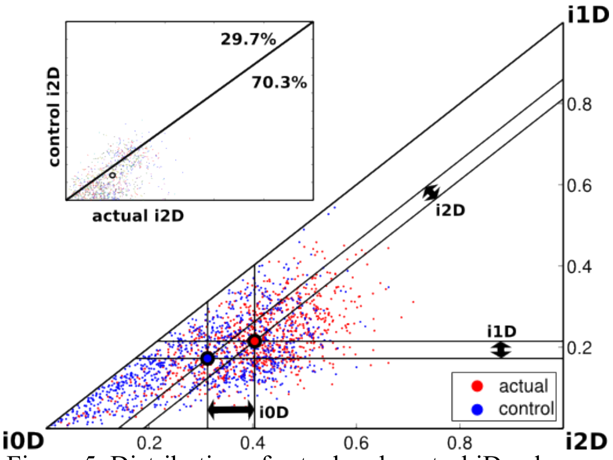

Figure 5. Distribution of actual and control iD values as well as their respective means within the iD triangle. *ix*D values stay constant along the straight lines. The inset shows i2D actual vs control values.

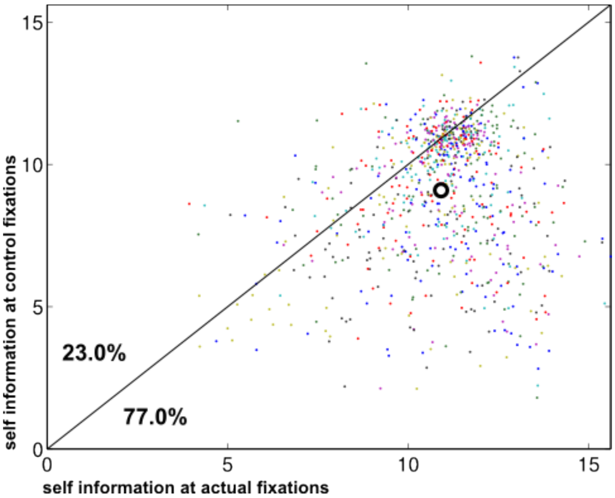

Figure 6: Self-information at actual vs control fixations for the whole image set (averaged over images and subjects). The big dot denotes actual vs control mean.

The self-information transformation of iD values is intended to give an indication of the saliency of a certain image point. Results show that actual values are highly significantly bigger than control ones, both for the whole image set as well as single image categories. The results show that 77.0% of actual self-information values for the complete image set are higher than control ones (see Figure 6). Category-wise examination shows the same effects within the single image categories, however to different degrees (73.7% man-made objects, 71.6% faces, 66.7% fractals, 58.6 natural scenes). As in the correlational analysis concerning single iD projections, self-information results show biggest effect sizes for the faces and man-made objects categories. The effect for fractals is smaller, that for natural scenes even less so. However, deviations of actual from control values for all image categories are highly significant.

3.2 Information-theoretic analysis

Entropy is significantly higher at actually fixated than at control points, meaning that histograms of iD values are more equalized for actual than for control points. This effect is visible for all three iD projections and consistent over image categories as well. Differences between actual and control distributions amount to 0.50 bit for i0D, 0.52 bit for i1D, and 0.64 bit for i2D projections or the whole image set (see Figure 7). Absolute entropy values are decreasing from i0D to i2D projections. Effect sizes for different image categories show a similar hierarchy than in the correlation analysis carried out above: Effects are most pronounced for man-made objects (0.33 to 0.55 bit difference between actual value and control maximum) and fractals (0.27 to 0.37 bit). Faces (0.16 to 0.52 bit) show a smaller effect size and natural scenes (0.01 to 0.07 bit), although meeting the criteria for significance for every iD projection exhibit only a tiny effect compared to the other image categories.

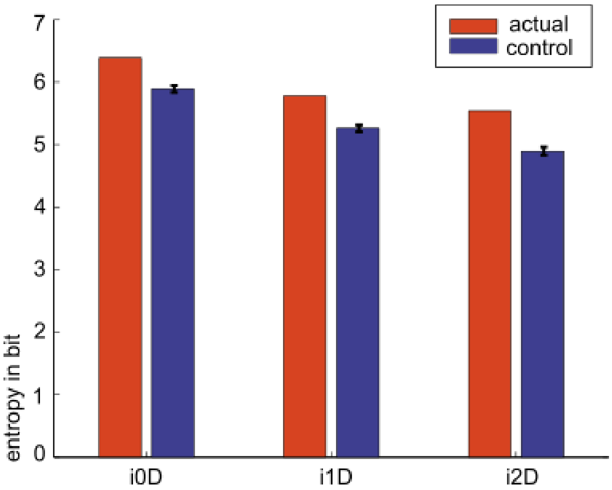

Figure 7: Entropy for actual and control conditions for the whole image set. Error bar indicate max/min control entropy values.

Mutual information is significantly lower at fixated than at control points. Effects are biggest for the i2D class (0.04 bit difference), smaller when calculated for the i0D class (0.03 bit) and barely noticeable for the i1D class (0.01 bit). The effect of mutual information being lower turns out to stem mainly from short saccades (see Figure 8), with the group containing the shortest saccades (up to 56px) returning significant results for all three iD projections. Differences between actual and control condition amount to 0.12 bit for i0D, 0.04 bit for i1D and 0.08 bit for i2D projections. The next group (56-105px) exhibits significantly lower mutual information values for actual than control sequences for i0D and i2D projections, while the third one does so for i0D values alone. In the control conditions, mutual information is decreasing as saccade lengths are getting longer. This is to be expected, as image regions near each other would be more correlated than regions further apart. This effect is much less pronounced – if present at all – for the actual condition though.

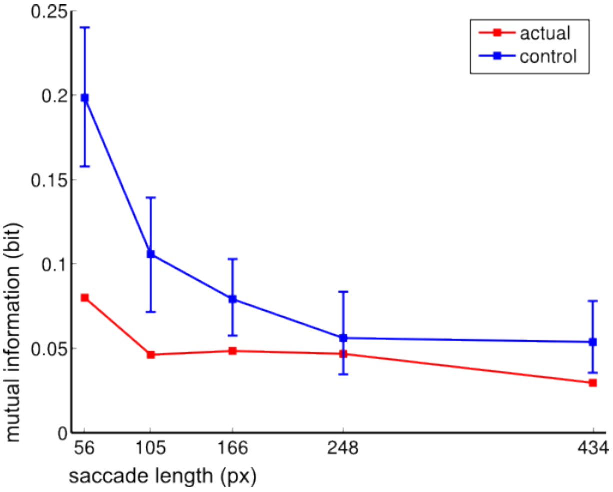

Figure 8. i0D mutual information as a function of saccade length in pixels. Error bars indicate max/min control mutual information values.

4 Conclusions

As shown, the iD algorithm provides a useful way to encode image features that goes along with what regions are determined to be salient by the human visual system. iD features are low-level and can be extracted by relatively simple computational means. Also, they are able to combine edge and corner classificators into a single non-discrete approach and therefore constitute a powerful tool in machine vision. It has already been applied in optic flow statistics [3] and as part of a larger framework in multi-modal matching [7].

Our results suggest that the human visual system tends to prefer edges and junctions/corners over homogeneous surfaces [cf. 4, 8]. In fact rare iD feature values are favored, while frequently occurring ones are neglected as shown by entropy analysis. The idea that it is especially two-dimensional image structures exhibiting high curvature like corners or junctions that are attracting fixations is based on the high informativeness of these regions. Krieger et al. [4] compared the bispectra (third-order statistics) of image patches at fixation points with those at randomly selected points and found a higher amount of highly curved image features. Also, as shown above, i2D features are rare in natural images. By exploiting similar features as the human visual system – that is concentrating on few but highly informative regions – the concept of iD has great potential for applications in computer vision systems.

The key concept of the self-information maps relates to the idea that it is not the absolute value of certain feature that determines its salience but rather this specific feature value in context to all other occurring feature values – and in this respect the surprise that is associated with encountering this specific value. In this study we applied the self-information transformation to iD maps that encode the intrinsic dimensionality of image patches. This procedure results in self-information maps that can readily be seen as saliency or interestingness maps since they highlight image regions that are worth looking at as these regions contain image features different from most of the rest of the image [cf. 1]. The self-information maps show bigger effect sizes than any of the iD features, when comparing actual with control fixation points. This goes along with the previously mentioned saliency enhancing effect of the self-information transformation on feature maps containing absolute strengths of certain image features [10].

There exist quantitative differences in the results pertained from different image categories – found both in the fixation and feature value correlation analysis as well as the information theoretic part. These could be due to the human visual system treating different image classes in different ways – either because of differences in the low-level properties or because of the application of different strategies after rapid image classification. Torralba et al. [11] showed that different image categories come along with distinct differences in their power spectrum, by means of which rapid image classification becomes possible. They used a variety of different image classes, containing natural scenes (mountain, beach, forest), man-made objects (highway, street, indoor), objects (face, car, chair) and so on. Another idea is that rapid scene identification is made possible by texture recognition that is highly parallelized [9].

Selection of a fixation point shows less dependence on the iD value of the point previously attended to than would be expected from chance. That is, actual mutual information values are lower for consecutive fixation points than control ones. This may suggest the existence of a decorrelation mechanism underlying the selection of consecutive fixation points. Thus, while a bottom-up approach may partly explain which points are taken into account for fixation – or at least why they are – there is indication that the temporal order in which fixations happen follows more complex rules.

5 Acknowledgments

This research has been funded by the European project PACOplus (IST-FP6-IP-027657) and the Danish project NISA..

References

[1]N.D.B. Bruce, “Features that Draw Visual Attention”, Neurocomputing, Vol 65-66, pp. 125-133, 2005.

[2]L. Itti, and C. Koch, “Computational Modelling of Visual Attention”, Nature Reviews Neuroscience, Vol 2, No. 3, pp. 194-203, 2001.

[3]S. Kalkan, D. Calow, M. Felsberg, F. Wörgötter, M. Lappe, N. Krüger, “Local Image Structures and Optic Flow Estimation”, Network: Computation in Neural Systems, Vol 14, No. 4, pp. 341-356, 2005.

[4]G. Krieger, I. Rentschler, G. Hauske, K. Schill, and C. Zetzsche, “Object and scene analysis by saccadic eye-movements: An investigation with higher order statistics”, Spatial Vision, Vol 13, No. 2/3, pp.201-214, 2000.

[5]N. Krüger, and M. Felsberg, “A continuous formulation of intrinsic dimension”, Proc. British Machine Vision Conference, 2003.

[6]J. Najemnik, and W.S Geisler, “Optimal eye movement strategies in visual search”, Nature, Vol 434, pp. 381-391, 2005.

[7]N. Pugeault, and N. Krüger, “Multi-Modal Matching Applied to Stereo”, Proceedings of the British Machine Vision Conference (BMVC) 2003, pp. 271-280, 2003.

[8]P. Reinagel, and A.M. Zador, “Natural scene statistics at the centre of gaze”, Network: Computation in Neural Systems, Vol 10, pp. 341-350, 1999.

[9]L.W. Renninger and J. Malik, “When is Scene Recognition just Texture Recognition?”, Vision Research, Vol 44, pp. 2301-2311, 2004.

[10]T.N. Topper, “Selections Mechanisms in Human and Machine Vision”, PhD thesis, University of Waterloo, Canada, 1991.

[11]A. Torralba and A. Oliva, “Statistics of Natural Image Categories”, Network: Computation in Neural Systems, Vol 14, pp. 391-412, 2003.

[12]J. van de Weijer, and Th. Gevers, “Tensor based feature detection for color images”, Proc. 12th Color Imaging Conference: Color Science and Engineering Systems, Technologies, Applications, pp. 100-105, 2004.