Gestalt Principles for Attention and Segmentation in Natural and Artificial Vision Systems

Gert Kootstra, Niklas Bergström, and Danica Kragic

***Abstract*— Gestalt psychology studies how the human visual system organizes the complex visual input into unitary elements. In this paper we show how the Gestalt principles for perceptual grouping and for figure-ground segregation can be used in computer vision. A number of studies will be shown that demonstrate the applicability of Gestalt principles for the prediction of human visual attention and for the automatic detection and segmentation of unknown objects by a robotic system.**

I. INTRODUCTION

The Gestalt psychology studies how humans perceive the visual environment as unitary elements instead of individual visual measurements [1], [2]. Through the years, Gestaltists have suggested different Gestalt principles for perceptual grouping and for figure-ground segregation [3], [4] (see Fig. 1 for some examples).

In this paper, we discuss the use of Gestalt principles in natural and artificial vision systems and give an overview of our work in this field. We present a few applications in computer vision for 1) the prediction of human eye fixations, and 2) the detection and segmentation of unknown objects in robotic vision.

The paper is organized as follows. In section II, we discuss Gestalt theory in natural vision systems with a focus on visual attention and the prediction of human eye fixations. Next, we discuss Gestalt theory in artificial vision and present our studies on object detection, object segmentation, and segment evaluation in section III. We end with a discussion in section IV.

II. GESTALT IN NATURAL VISION SYSTEMS

The Gestalt psychology studies the tendencies of humans to group individual elements of the visual scene in to larger structures, such as objects [3]. In this section, we focus on this aspect in relation to visual attention. We first discuss some insights from visual-search experiments, and then present our research on the prediction of human eye fixations based on symmetry.

A. Visual search and configural superiority

In visual-search experiments, participants are asked to search for the odd figure among a large number of figures. The odd figure is different from the other figures in a particular feature, for instance, in color, or shape. According to [5], the efficiency of the search depends on the increase of the reaction time as a function of the number of items in the search display. If this increase is close to zero, search is said to be efficient, and the feature is associated with a pop-out effect.

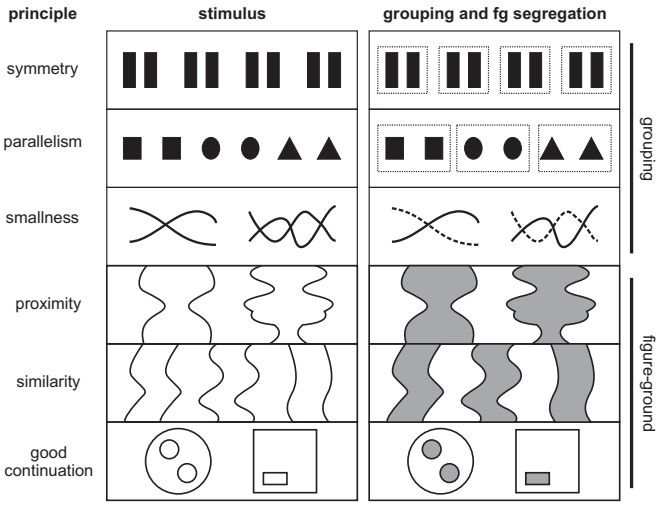

Fig. 1: Gestalt principles for perceptual grouping and figure-ground segregation

Many basic features result in efficient search, such as, brightness, color, and orientation [6], or shape [7], [8]. However, conjunction search is inefficient [6]. In conjunction search, the target is not unique in either of the features, but in the combination only, for instance, a red square among red circles and green squares. In that case, reaction times drastically increase when there are more items in the display.

Figure 2 shows an interesting case. Figure 2a shows the original search display. Search for the ')' among the 16 items is not so efficient. We therefore expect the reaction time to increase when we add another 16 items shown in Fig. 2b. However, the reverse is true. Reaction times actual decrease when participants search for the odd figure in Fig. 2c [9]. The reason for this is that humans do not perceive 32 individual basic elements, but that the basic elements are grouped into 16 configures. The fact that search for the target in Fig. 2c is more efficient than in Fig. 2a can be explained from the *emerging features* in the configures. Instead of only having the basic feature curvature, the elements now also have the configural features *symmetry* and *enclosure*, which makes the target more salient among the distractors. This example shows the *configural superiority* in visual attention.

The Gestalt principles for perceptual organization and figure-ground segregation provide a number of these configural features. In the following subsection, we show the

All three authors are with the Center for Autonomous Systems of the Royal Institute of Technology (KTH), Stockholm, Sweden. {kootstra|nbergst|danik}@kth.se

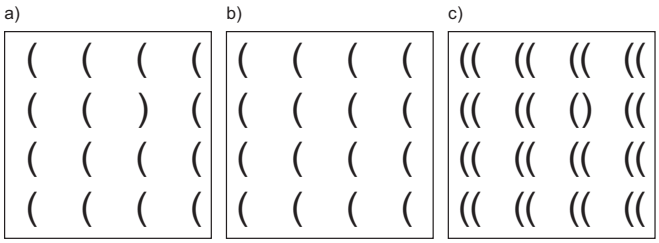

Fig. 2: Examples of configural superiority. Searching the ')' among '(' in (a) is quite challenging. Usually in visual search experiments, adding more item to the display (b) will increase the reaction time. However, the search for the target in (c) is actually more efficient. The reason for this is that apart from the basic feature *curvature*, there are now two additional emerging configural features, *symmetry* and *enclosure*. This attracts more attention. (Figure adapted with modifications from [9].)

use of one of the principles, *symmetry*, for the prediction of overt visual attention.

B. Symmetry for the prediction of human eye fixations

Computational models for the prediction of eye fixations on photographic images are often based on saliency models that determine the saliency at every point on the image based on specific features. Most of these models are based on the center-surround contrast of basic features, inspired by the Feature-Integration theory [6]. Examples are the contrast in intensity, color, and orientation [10], texture [11], and distributions of features [12].

However, the example in the section II-A illustrates that visual attention is guided by a hierarchy of features in which higher-level features like the Gestalt features precede lower-level basic features, as also suggested in [13]). We used this idea in our visual-attention model by using local symmetry to predict human eye fixations [14].

A motivation for our symmetry-saliency model is shown in Fig. 3. The figure shows some images and the distribution of human eye fixations, as well as the predictions of the contrast-saliency model [10] and the symmetry-saliency model [14]. The human fixations are clearly concentrated at the symmetrical centers of the objects. This is not reproduced by the contrast model, which puts emphasize at the borders of the objects, where the contrast with the background is higher. The symmetry model, on the other hand, gives a much better prediction.

The symmetry-saliency model is illustrated in Fig. 4. The model is based on the symmetry operator presented in [15] and extended to a multi-scale saliency model. The amount of local mirror symmetry at a point $\mathbf{p}$ in the image is determined by comparing the gradients of neighboring pixels pairs in the symmetry kernel. The symmetry contribution, $s(i,j)$ of a pixel pair $\{\mathbf{p}_i, \mathbf{p}_j\}$ is determined by:

$$s(i,j) = c(i,j) \cdot log(1+m_i) \cdot log(1+m_j) \quad (1)$$

$$c(i,j) = (1-\cos(\gamma_i+\gamma_j)) \cdot (1-\cos(\gamma_i-\gamma_j)) \quad (2)$$

where the variables are illustrated in Fig. 4. The contributions of all pixel pairs in the kernel are combined to give the amount of local symmetry for the given point at a given scale. The local symmetry is then calculated for all pixels

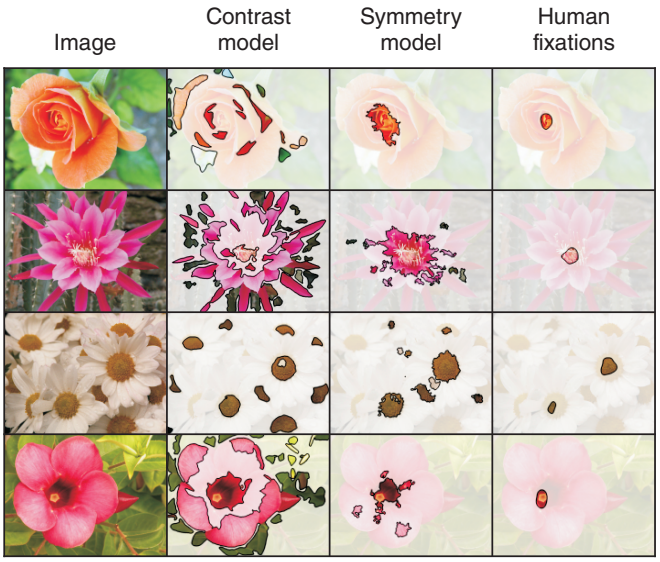

Fig. 3: Examples that a saliency model using symmetry can make better predictions about the location of human eye fixations than a contrast-saliency model. The second column shows the contrast-saliency maps, the third column gives the symmetry-saliency maps, and the human-fixation density maps are shown in the last column. The preference of humans to fixate on the center-of-symmetry of the flowers is correctly reproduced by the symmetry model, whereas the contrast model puts focus on the edges of the forms. The regions of the image that are highlighted are the parts of the maps above 50% of its maximum value. Not only for these image, but on a large variety of different photographic images, the symmetry model outperforms the contrast model [14].

and on different scales. The summation over scales gives the symmetry-saliency map.

In [14] we showed that the symmetry-saliency model outperforms the contrast-saliency model not only on the images shown in Fig. 3, but on a large number of different images. The resulting saliency maps correlate better with fixation-density maps created from the human eye fixations. The amount of symmetry is furthermore highest at the first fixations, and slowly drops for later fixations. Since earlier fixations are thought to be more bottom-up controlled, symmetry can thus be a good predictor of bottom-up visual attention.

The explanation of these results can be found in the Gestalt principles. To interpret a visual scene, humans pay attention to objects and not so much to individual features [16], [17] and symmetry, as one of the principles for figure-ground segregation, can be used as a bottom-up feature for the detection of these objects.

The success of using symmetry as a bottom-up feature for the prediction of human visual attention, motivated us to study the applicability of Gestalt principles for the detection and segmentation of unknown objects in artificial vision systems.

III. GESTALT IN ARTIFICIAL VISION SYSTEMS

The detection and segmentation of unknown objects is an important, but challenging problem in robotic vision. Since no prior knowledge is available about the objects, top-down search methods cannot be used to detect the objects. Instead, bottom-up methods need to be used to find and segment the

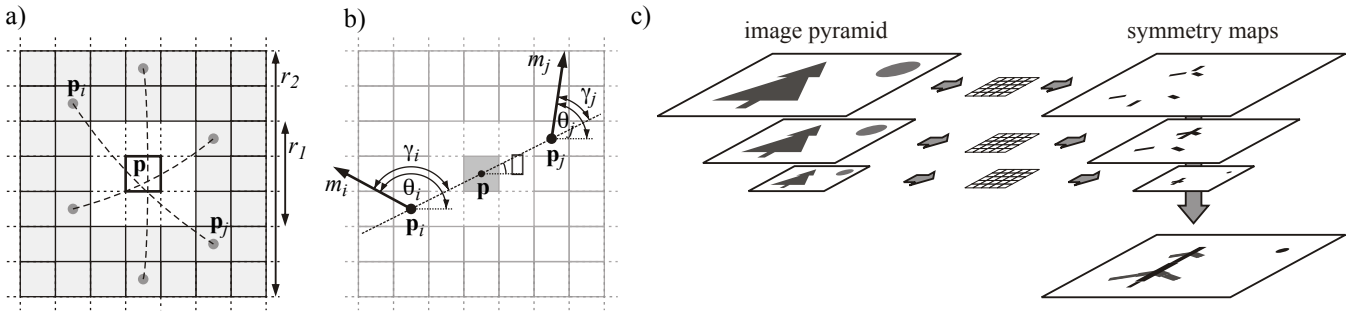

Fig. 4: The symmetry model. a) The symmetry kernel that is applied to all pixels. b) The symmetry contribution of a pixel pair is based on brightness gradients. c) The symmetry responses on different scales are combined to the symmetry-saliency map.

objects in the scene, in order to enable the robot to learn about the objects and manipulate them. This is exactly what many of the Gestalt principles for perceptual grouping and figure-ground segregation entail.

We developed a system for object detection, object segmentation, and segment evaluation of unknown objects based on Gestalt principles. The general setup of this method is depicted in Fig. 5. Firstly, the object-detection method will generate hypotheses (fixation points) about the location of objects using the principle of symmetry. Next, the segmentation method separates foreground from background based on a fixation point using the principles of proximity and similarity. The different fixation points and possibly different settings for the segmentation method result in a number of object-segment hypotheses. Finally, the segment-evaluation method selects the best segment by determining the goodness of each segment based on a number of Gestalt principles for *figural goodness*. In the following subsections, we describe each module in more detail.

A. Symmetry for the detection of unknown objects

We proposed a bottom-up method for the detection of unknown objects using local symmetry in [18]. This method is based on the symmetry-saliency model that we used to predict human eye fixations (see Section II and Fig. 4).

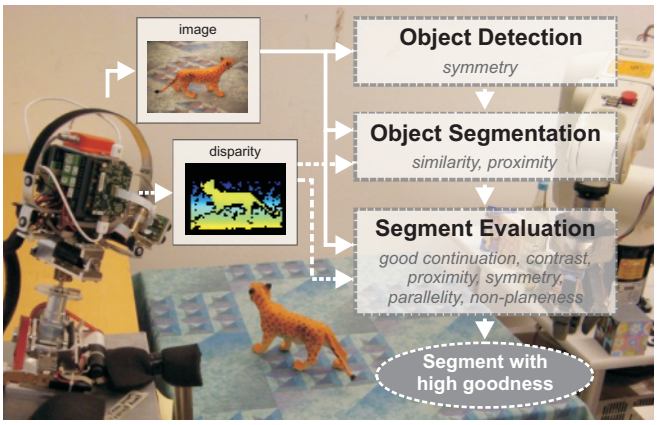

Fig. 5: The bottom-up method for the detection, segmentation, and evaluation of unknown objects. The object-detection step generates object hypotheses by selecting fixation points based on *symmetry* in the image. Next, one of these object hypotheses initializes the object segmentation. Based on *similarity* in color, *proximity* in depth, and deviations from the dominant plane, the foreground is segmented from the background. Finally, the *figural goodness* of the resulting segment is evaluated using *good continuation*, *contrast*, *proximity*, *symmetry*, *parallelity*, *color uniqueness* and deviations from the dominant plane.

Based on the saliency map, fixation points are iteratively generated by selecting the local maximum in the map with the highest saliency value. An inhibition-of-return mechanism lowers the probability to revisit an earlier fixated area. This mechanism devalues all local maxima that are in the same salient blob as the generated fixation point. This ensures that points in different parts of the image in different salient regions are selected. The fixation points are the object hypotheses, which initialize the segmentation. More details about the method can be found in [18].

We tested our object-detection method on two different databases, the MSRA Salient Object database [19] and the KTH Object and Disparity (KOD) database [1]. Both databases have hand-labeled ground truth and especially the first database contains very challenging images with a lot of visual clutter. The results have been compared to those using the contrast-saliency method [10] and are shown in Fig. 7. The figure depicts the proportion of times that the salient object in the scene is detected as a function of the number of selected fixation points. It can be seen that our symmetry method significantly outperforms the contrast method. Already for the first fixation, the detection rates are high. Additional fixations increase the proportion of any of the fixations being on the object. Not only does our method have a higher detection rate of saliency objects in the images, it has also been shown to produce fixations points that are closer to the objects' center, which is favorable for initializing object segmentation [18].

The object-detection method is furthermore fast. Running on a CPU (2.53 GHz Intel processor), the method takes approximately 50 ms for a 640×480 image, while a parallel implementation on a GPU (Nvidia GTX 480) runs in 5-10 ms .

B. Object segmentation by grouping super pixels

The object-detection method hypothesizes the location of objects by generating fixation points. To segment foreground from background, we use the segmentation method that we presented in [20]. The method pre-segments the image into super pixel, where the super-pixels are clusters of neighboring pixels with similar color. To group fore- and

[1] http://www.csc.kth.se/~kootstra/kod

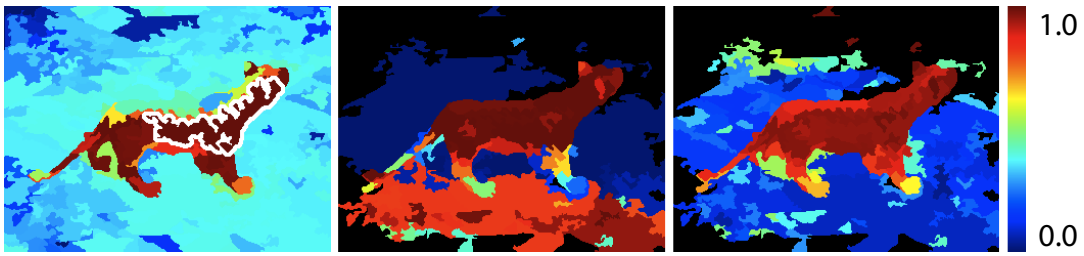

Fig. 6: Illustration of the likelihoods for every super pixel to belong to the foreground given the fixated super pixel (white boundary). Left: using similarity in color. Middle: using proximity in depth. Right: the likelihood of belonging to the foreground based on 3D information and the dominant plane. Black super pixels have no valid disparity information. The image is best viewed in color.

background super pixels in the image, the Gestalt principles of *similarity* and *proximity* are used. Super pixels that are similar to the fixated super pixel in color and/or proximal in 3D position are likely to be labeled as foreground. In addition, the estimated 3D planes of the super pixels are compared to the dominant plane, to remove elements of the supporting plane from the foreground segment.

The segmentation process is formulated as a Markov Random Field, using graph cuts to minimize the energy function. This energy function is based on color, depth, and plane information. Figure 6 gives an example of the likelihoods that the super pixels belong to the foreground based on the similarity to the fixated super pixel in color, the proximity in depth, and 3D plane information. Initially, only the fixated super pixel is labeled as foreground, but the segment is iteratively refined by comparing all super pixels to the current fore- and background information. More details about the method can be found in [20].

We compared the segmentation method with a recent method, the active-segmentation method proposed in [21], which also uses disparity information for the segmentation and initializes the process with a fixation point. Both methods have been tested on the KTH Object and Disparity (KOD) database[2]. The produced figure-ground segmentation have been compared to the hand-labeled object segments. The similarities with the ground truth given by the F1-score are shown in Fig. 8. The plots indicate that our method performs significantly better than the active-segmentation method. Fixation points selected based on contrast are often located near the borders of the object (see discussion in Section III-A). This causes less problems for our method, since color, disparity and plane information can be gather for the whole super pixel. Fig. 11 and 12 show some examples of segments.

Our method is furthermore a couple of magnitudes faster than the active-segmentation method[21]. Due to the use of super pixels, the Markov Random Field does not contain many nodes. The graph-cut minimization of the energy function is therefore very fast. Implemented on a CPU (2.53 GHz Intel processor), the segmentation is done in 4-8 ms. This allows us to perform multiple segmentations and evaluate the resulting segments. The super-pixel pre-segmentation, including the transformation to the L*ab* color space and

[2]http://www.csc.kth.se/~kootstra/kod

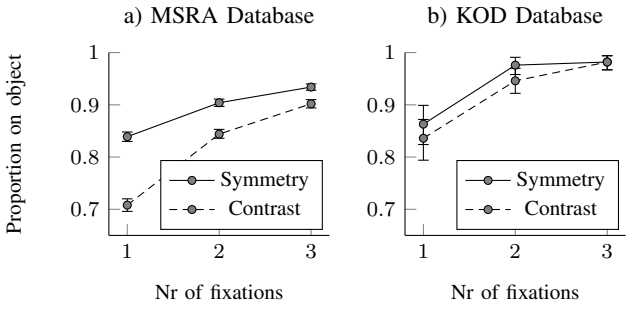

Fig. 7: The object-detection performance given by the proportion of times that any of the fixations is on the salient object in the scene as a function of the number of fixations. The symmetry-saliency method [18] is compared to the contrast-saliency method [10] on two different databases.

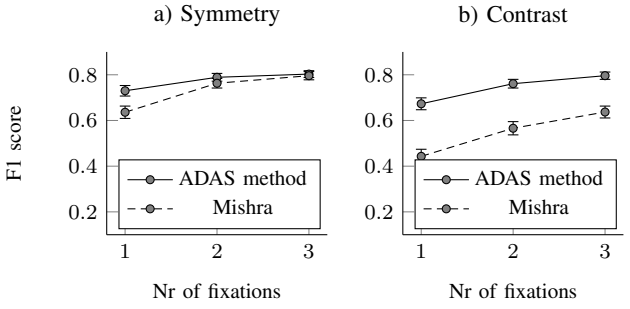

Fig. 8: The object-segmentation performance showing the best segmentation of any of the fixations selected by respectively symmetry and contrast. Our method [20] is compared to the method presented in [22]

the calculation of some color and disparity statistics, takes approximately 100 ms on the same CPU. A parallel GPU implementation of this part is quite straightforward and will greatly cut the processing time. As a comparison, the active-segmentation method[21] needs a couple of minutes for a segmentation on the same machine.

C. Gestalt principles for the evaluation of segments

Failures in both the object-detection and in the object-segmentation method can lead to incorrect object segments. The detection method can for instance propose fixation points that are located near the boundaries of the object - which generally results in a poor segmentation [18] - or that are not located on the object at all. The segmentation method can fail to find the true boundaries of the object. To detect these kind of failures, we proposed a segment evaluation method [23].

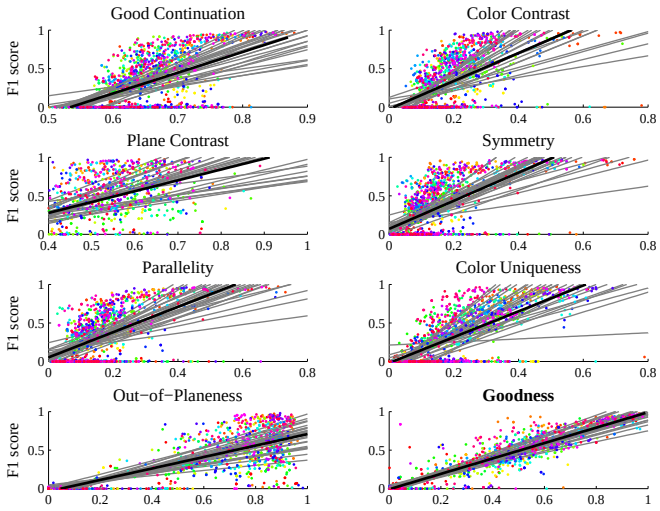

Fig. 9: The relation between the Gestalt principles and the true quality of the object segmentation measured by the F1 score. The results for the different objects are displayed with differently colored dots. The solid black lines show the linear regression model fitted to all data points, and the solid gray lines show the linear regression for the individual objects.

The method evaluates an object segment using a number of Gestalt principles for *figural goodness*. This is the Gestalt term for a measure of how good, ordered, or simple a shape is. By having the object-detection method suggesting multiple fixation points and by running the segmentation method with different parameter settings, a number of possible segments are suggested. The proposed segment-evaluation method then select the best segments among the suggestions by comparing the figural goodness values.

The figural goodness is based on the principles of *good continuation*, *contrast*, *proximity*, *symmetry*, *parallelity*, *color uniqueness* and on deviations from the dominant plane. Details about these different methods can be found in [23]. The segment is evaluated on each of the principles and the individual measures are combined into a goodness measure by training an artificial neural network.

In Fig.9, the different goodness measures are plotted against the F1 score. Given a scene and a number of segments, the goodness measures should be able to indicate the best segments, that is the measures should positively correlate with the F1 score. It can be appreciated from the plots that the measures indeed show a positive relation with the ground-truth quality measure. The correlation coefficients moreover indicate positive correlations for all measures, with high values for the explained variance (see Tab. I). This shows that all measures are good predictors of the quality of the segments [23]. Especially the correlation for the color uniqueness and the out-of-planeness is high. Also contrast in 3D plane, symmetry, and parallelity performed well. A linearly combination of the individual measures improves the results, which shows that the different measures are complimentary.

To combine the different measures in one goodness measure, a multi-layer feed-forward neural network was trained to predict the segment quality. The trained network shows a very good correlation, with a high value for the explained

TABLE I: Correlation and explained variance (R^2) of the Gestalt measures.

Measure	Correlation	R^2 measure
good continuation	0.56	0.31
color contrast	0.58	0.34
plane contrast	0.64	0.41
symmetry	0.63	0.39
parallelity	0.61	0.37
color uniqueness	0.71	0.51
out-of-planeness	0.77	0.59
Linear combination	–	*0.80*
Neural network	0.93	*0.87*

variance, outperforming the linear combination of measures (see Tab. I). This can also be appreciated by the low variance of the measurements around the diagonal in the goodness plot in Fig. 9. Some examples of segments and their calculated goodness measure are given in Fig. 11 and 12.

Based on the goodness measure, the segment-evaluation method selects the best segment from a number of suggested segments. Figure 10 shows that this greatly improves the segmentation performance. The segmentation performances for the theoretical optimal selection are given by the solid lines. The bar shows the performance of our fully automatic object detection, segmentation and segment-evaluation method. It can be seen that the performance is very close to the theoretical optimum. The results show that the Gestalt measures for the goodness of a segment correspond well with the objective quality of the segment.

IV. DISCUSSION

Many Gestalt principles have been established through decades of psychophysical experimentation. In this paper, we presented an overview of our work on the use of Gestalt principles in computational models for 1) the prediction of human visual attention, and 2) attention and segmentation in artificial vision systems. The results presented show that Gestalt principles can be successfully used to predict human eye fixations using symmetry. We furthermore demonstrated that unknown objects in a scene can be detected using symmetry, and that they can be segmented from the background using the principles of similarity and proximity. We finally showed that the quality of object segmentation can be evaluated using a large number of Gestalt principles, i.e., good continuation, contrast, proximity, symmetry, parallelity,

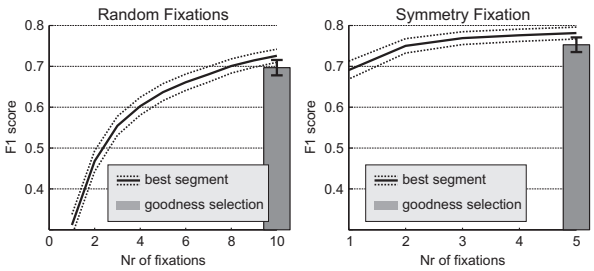

Fig. 10: Selecting the segment with the highest Gestalt goodness. The solid line show the average F1 score of the objectively best segment as a function of the number of fixation points. The gray bar displays the mean F1 score of the segment with the highest figural goodness. The left plot is based on random fixations and the right plot on fixation points selected by our object-detection method. The error bars give the 95% confidence intervals.

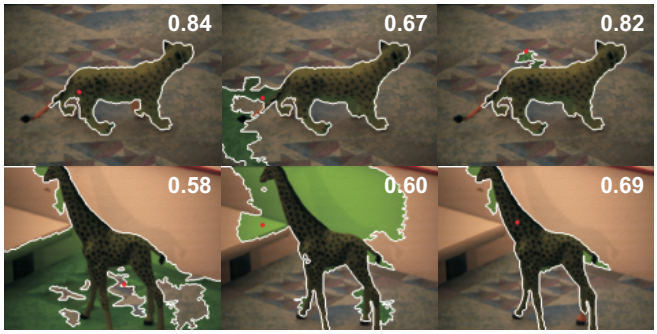

Fig. 11: Examples of segments resulting from different fixation points. The red dot indicates the fixation point, and the green area with the white border shows the segment. The goodness measures of our segment evaluation method are shown in the top-right corner of the images.

Fig. 12: Examples of scenes with multiple objects. The numbers in the top-right corners of the images indicate the goodness measures resulting from the segment-evaluation method.

color uniqueness, and deviations from the dominant plane.

Using the Gestalt principles, meaningful abstractions of the scene can be achieved through bottom-up processing of visual information. This is especially valuable in cases in which our robots have to deal with novel object and environments. However, also for processes using top-down information, for instance, visual search and recognition of known objects, such a sparse and meaningful representation is beneficial to decrease the search space.

Acknowledgments

This work was supported by the EU through the project eSMCs, IST-FP7-IP-270212 and Swedish Foundation for Strategic Research.

References

[1] M. Wertheimer, "Untersuchungen zur lehre von der gestalt ii," *Psychologische Forschung*, vol. 4, pp. 301–350, 1923, translation published in Ellis, W. (1938). A source book of Gestalt psychology (pp. 71-88). London: Routledge & Kegan Paul.

[2] K. Koffka, *Principles of Gestalt Psychology*. London: Lund Humphries, 1935.

[3] S. E. Palmer, *Vision Science. Photons to Phenomenology*. Cambridge, MA: MIT Press, 1999.

[4] ——, "Modern theories of gestalt perception," in *Understanding Vision: An Interdisciplinary Perspective – Readings in Mind and Language*, G. W. Humphreys, Ed. Oxford, England: Blackwell, 1992, pp. 39–70.

[5] J. M. Wolfe, "Visual search," in *Attention*, H. Pashler, Ed. University College London Press, 1998.

[6] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[7] J. Theeuwes, "Perceptual selectivity for color and form," *Perception & Psychophysics*, vol. 51, no. 6, pp. 599–606, 1992.

[8] B. Julesz, "A brief outline of the texton theory of human vision," *Trends in Neuroscience*, vol. 7, pp. 41–45, 1984.

[9] J. R. Pomerantz, "Colour as a gestalt: Pop out with basic features and with conjunctions," *Visual Cognition*, vol. 14, no. 4-8, pp. 619–628, 2006.

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

[11] D. J. Parkhurst and E. Niebur, "Texture contrast attracts overt visual attention in natural scenes," *European Journal of Neuroscience*, vol. 19, pp. 783–789, 2004.

[12] H. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, pp. 1–24, 2009.

[13] A. Açik, S. Onat, F. Schumann, W. Einhäuser, and P. König, "Effects of luminance contrast and its modifications on fication behavior during free viewing of images from different categories," *Vision Research*, vol. 49, pp. 1541–1553, 2009.

[14] G. Kootstra, B. de Boer, and L. R. B. Schomaker, "Predicting eye fixations on complex visual stimuli using local symmetry," *Cognitive Computation*, vol. 3, no. 1, pp. 223–240, 2011, doi: 10.1007/s12559-010-9089-5.

[15] D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Context-free attentional operators: The generalized symmetry transform," *International Journal of Computer Vision*, vol. 14, pp. 119–130, 1995.

[16] B. J. Scholl, "Objects and attention: the state of the art," *Cognition*, vol. 80, no. 1-2, pp. 1–46, 2001.

[17] Y. Yeshurun, R. Kimchi, G. Sha'shoua, and T. Carmel, "Perceptual objects capture attention," *Vision Research*, vol. 49, pp. 1329–1335, 2009.

[18] G. Kootstra, N. Bergström, and D. Kragic, "Using symmetry to select fixation points for segmentation," in *Proceedings of the International Conference on Pattern Recognition*, 2010.

[19] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *Int. Conf. on Computer Vision and Pattern Recogntion (CVPR '07)*, 2007.

[20] G. Kootstra, N. Bergström, and D. Kragic, "Fast and automatic detection and segmentation of unknown objects," in *Proceedings of the IEEE-RAS International Conference on Humanoid Robots (Humanoids 2010)*, 2010.

[21] A. Mishra, Y. Aloimonos, and C. L. Fah, "Active segmentation with fixation," in *Proceedings of the International Converence on Computer Vision (ICCV)*, 2009.

[22] A. Mishra, C. Fermuller, and Y. Aloimons, "Active segmentation for robotics," in *Proceedings of the Interation Conference on Intelligent Robots and Systems (IROS)*, 2009.

[23] G. Kootstra and D. Kragic, "Fast and bottom-up object detection and segmentation using gestalt principles," in *Proceedings of the International Conference on Robotics and Automation (ICRA)*, 2011.