Object Recognition Using a Generalized Robust Invariant Feature and Gestalt's Law of Proximity and Similarity

Kim Sungho, Kuk-Jin Yoon, Inso Kweon

Object Recognition Using a Generalized Robust Invariant Feature and Gestalt's Law of Proximity and Similarity

Sungho Kim, Kuk-Jin Yoon and In So Kweon
Korea Advanced Institute of Science and Technology
373-1 Guseong-dong Yuseong-gu Daejeon, Korea
{shkim, kjyoon}@rcv.kaist.ac.kr, iskweon@kaist.ac.kr

Abstract

In this paper, we propose a new context-based method for object recognition. We first introduce a neuro-physiologically motivated visual part detector. We found that the optimal form of the visual part detector is a combination of a radial symmetry detector and a corner-like structure detector. A general context descriptor, named G-RIF (Generalized-Robust Invariant Feature), is then proposed, which encodes edge orientation, edge density and hue information in a unified form. Finally, a context-based voting scheme is proposed. This proposed method is inspired by the function of the human visual system, called figure-ground discrimination. We use the proximity and similarity between features to support each other. The contextual feature descriptor and contextual voting method, which use contextual information, enhance the recognition performance enormously in severely cluttered environments.

1. Introduction

How to cope with image variations caused by photometric and geometric distortions is one of the main issues in object recognition. It is generally accepted that the local invariant feature-based approach is very successful in this regard. This approach is generally composed of visual part detection, description and classification.

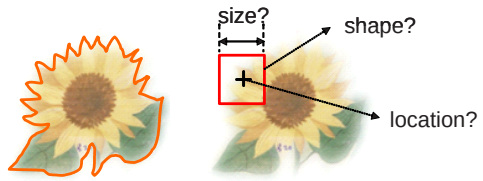
The first step in the local feature-based approach is visual part detection. Lindeberg [9] proposed a pioneering method on blob like image structure detection in scale-space. Shokoufandeh [24] extended this feature to wavelet domain. Schmid et al [21] compared various interest point detectors and concluded that the scale-reflected Harris corner detector is most robust to image variations. Mikolajczyk and Schmid [13] also compared visual part extractors and found that the Harris-Laplacian based part detector is suitable for most applications. Recently, several visual descriptors have been proposed [10],[14],[5],[26]. Most approaches try to encode local visual information such as spatial orientation or edge.

Based on these local visual features, several object recognition methods, such as the probabilistic voting method [20] and constellation model-based approaches [4], [15], have been introduced. However, those approaches occasionally fail to recognize objects with few local features in highly cluttered backgrounds. Recently, Stein and Hebert [25] proposed a background invariant object recognition method by combining local feature with the object segmentation scheme. Because this method is based on prior figure-ground information, it cannot be used in a general environment. Torralba et al. [27] introduced a completely different approach that exploits the background clutter information into object recognition. The background information is called place context or exterior context in that method. The exterior context information is very useful for practical applications such as intelligent mobile robotics systems.

Our research interest is how to efficiently extract an object's interior contextual information to enhance object recognition performance in strongly cluttered environments. We concentrate on the properties of a human visual receptive field and propose a computationally efficient local context coding method. We also propose an object recognition method based on the human visual system's figure-ground discrimination mechanism, which is accomplished by grouping interior contextual information.

2. Perceptual visual part detection

Appearance- or view-based object recognition methods were proposed in the early 1990s for face recognition and currently have become popular in the object recognition society. However, global appearance representation using a perfect support window as shown in Figure 1(a) cannot recognize objects in cluttered environments because of imperfect figure-ground segmentation. A bypassing method is to approximate an object as the sum of its sub-windows.



(a) (b)

Figure 1. (a) Perfect global object description. (b) Issues in local part-based object description.

The main issue in this approach is how to select the location, shape and size of a sub-window as shown in Figure 1(b). For successful object recognition, visual parts have to satisfy the following requirements. First, background information should be excluded as much as possible. Second, they must be perceptually meaningful. Finally, they should be robust to the effects of photometric and geometric distortions. Mahamud [12] suggested a greedy search method to find visual parts that satisfy the above requirements. Image structure-based part detectors are more efficient than regular grid-based part [21], [10]. Although these methods work well, they do not exploit the full structural information of interesting objects.

Recently, Serre and Riesenhuber [22] proposed a neuro-computational object recognition method by precisely modeling the properties of simple and complex cells. The basic concepts of the model are tuning and MAX operation on both feature detection and recognition. Although this model is a biologically good model, it is computationally inefficient because of the enormous tuning process of the location, size and phase of the Gabor filter.

We approximate the tuning process of a simple cell by two dominant Gabor filters - 0° and 90° phases [19]. A 0° phase Gabor is equivalent to the 2nd derivative of a Gaussian, and a 90° phase is equivalent to the 1st derivative of a Gaussian kernel. According to the complex cell model, location and size invariance are achieved by the MAX operation of various tuning responses. Figure 2 shows the complex cell responses using the approximated model for cross and circle structures.

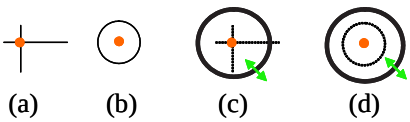


Figure 2. (a) Location tuning by spatial MAX operation of 1st derivative of Gaussian. (b) Location tuning by spatial MAX operation of 2nd derivative of Gaussian. (c) Scale tuning by scale-space MAX operation of 2nd derivative of Gaussian centered on (a). (d) Scale tuning by scale-space MAX operation of 2nd derivative of Gaussian centered on (b).

(1) Location tuning using 1st derivative of Gaussian
- Select maximal response in all orientations within a 3 × 3 complex cell.
- Suitable method: Harris corner or KLT corner extraction (both eigen values are large).
(2) Location tuning using 2nd derivative of Gaussian
- Select maximal response in all orientations within 3 × 3 complex cell.
- Suitable method: Laplacian or DoG gravity center (radial symmetry point).
(3) Scale tuning using convexity
- Select maximal response in directional scale-space [16].
- For computational efficiency, a convexity measure such as DoG is suitable. This is related to the properties of the V4 receptive field where the convex part is used to represent visual information [17].

Figure 3 shows the biologically motivated and computationally efficient model. Scale reflected 1st derivative of Gaussian can be approximated by subtracting neighboring pixels in a scale-space image, and 2nd derivative of Gaussian can be approximated by subtracting between scale-space images. This is supported by the psychophysical facts that HVS attends on gravity centers and high curvature points [18], and objects are decomposed into a perceptual part that is convex [8], [11].

Interestingly, the final approximated model of the simple and complex cells receptive field is a conjunction of the Harris-Laplace and DoG-DoG feature detectors. Figure 4 shows the results of the proposed perceptual part detector. Note that the proposed method extracts all possible visual parts complementing each other.

3. General contextual descriptor

The description of visual parts is very important for object identification or classification. Two important factors in

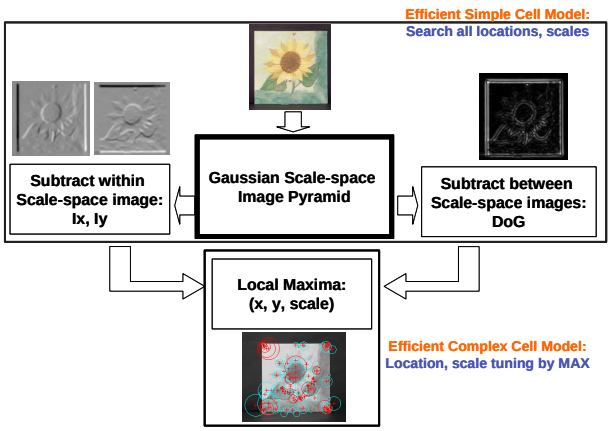
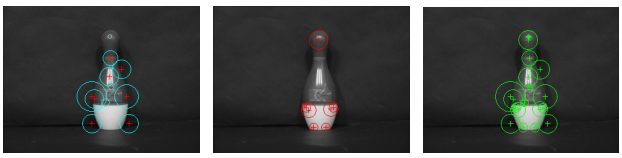


Figure 3. The final model of perceptual part detection which is an approximated simple and complex cell model of HMAX.



(a) (b) (c)

Figure 4. (a) Right pass: radial symmetry part. (b) left pass: corner-like part. (c) final visual part detector.

descriptor design are selectivity and invariance. Selectivity means that different parts or part classes can be discriminated robustly. Invariance means that the same parts or part classes can be detected regardless of photometric and geometric distortions. Descriptors must satisfy both properties appropriately. Direct use of pixel data shows very high selectivity but very low invariance. A PCA-based descriptor shows high selectivity but low invariance due to its properties that are sensitive to the translation of interest points [5]. Histogram-based descriptors show a proper compromise of selectivity and invariance [10], [26], [1].

Then, what information has to be represented? We found the solution from the properties of the receptive field in V1, V4. Through simple and complex cells, orientation response maps are generated and fine orientation adaptation occurs on the receptive field within the attended convex part in V4 [2]. The computational method of the orientation adaptation phenomenon is steering filtering [3]. Adapted orientation is calculated by the maximum response spanned by basis responses. Color blobs also exist in the hyper column where opponent color information is stored. Hue represents perceptual color information like the human visual system. Hue is invariant to affine illumination change and highlights. This orientation and color information is combined in V4 [29].

Summarizing these findings, a plausible receptive field response in V4 is shown in Figure 5. Adapted orientation, edge strength and hue field information coexist within the attended convex visual part.

Now, how does the HVS encode the receptive field responses within the attended convex part? Few facts are discovered on this point, but it is certain that larger receptive fields are used to represent broader orientation bands [7]. Figure 6 represents such a receptive field model in higher area V4. Smaller receptive field information such as edge orientation, edge strength and hue are encoded to larger receptive fields (black circular regions). The density of the black circle represents the HVS' attention level. 86% fixation occurs around the center receptive field [18].

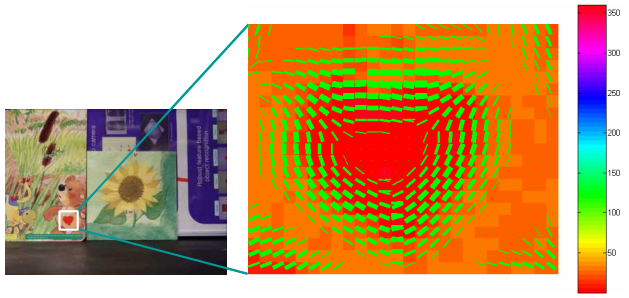


Figure 5. Plausible receptive model in V4: orientation adapted bar with edge strength on the hue field.

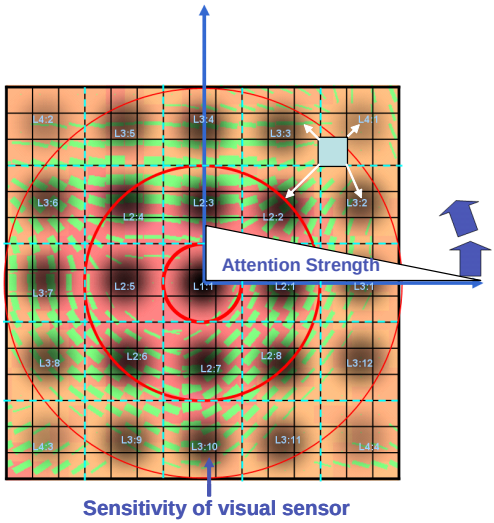


Figure 6. Plausible receptive field model in higher area V4: black circles represent larger receptive fields.

Each receptive field encodes histogram-based orientation distribution in $[0, \pi]$, hue distribution in $[0, 2\pi]$ and scalar value of edge density in $[0, 1]$. The coordinates of the receptive field are aligned to the dominant orientation of the convex visual part, which is calculated efficiently by steering filtering [3]. This encoding scheme can control the level of local context information-aperture size and feature level as in Table 1. The proposed encoding scheme is a general form of the contextual feature representations of [10], [26], [1]. We can select a suitable level of context depending on applications. In general, as the aperture size is larger, the recall is lower and the precision is higher. Figure 7 shows an example of visual part matching using the proposed part detector and general context descriptor (G-RIF: generalized robust invariant feature). Most visual parts correspond to one another by a simple Euclidean distance measure. More specific details of implementation and invariant properties can be found in [6].

4. Neighboring context-based object recognition

A conventional classifier based on local informative features is direct voting of nearest neighbors as in Figure 8(a)

Context	Contents of information
Aperture size	L1:1 L2:1,L2:2,L2:3,L2:4,L2:5,L2:6,L2:7,L2:8 L3:1,L3:2,L3:3,L3:4,L3:5,L3:6,L3:7,L3:8, L3:9,L3:10,L3:11,L3:12 L4:1,L4:2,L4:3,L4:4
Type of feature	(1) Edge orientation $[0, \pi]$, quan. level:4 (2) Edge density $[0, 1]$ (3) Hue information $[0, 2\pi]$, quan. level:4 etc: (1)+(2), (1)+(3),(2)+(3), (1)+(2)+(3)

Table 1. Level of local context.

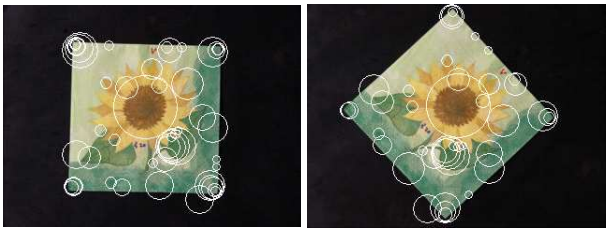


Figure 7. The matching results on the rotated objects using the perceptual part detector and local contextual descriptor [G-RIF: aperture size L3 and feature type (1)+(2)].

[20]. Although there are more sophisticated classifiers, such as SVM [28], Adaboost [30], Bayesian decision theory [15], and strong spatial constraint-based indexing [23], those methods are basically based on the concept of nearest neighbors and their voting.

The proposed object recognition method, named neighboring context-based voting, can be modeled as follows:

$$L = \arg\max_{l} P(l|\mathbf{X}) \approx \arg\max_{l} \sum_{i=1}^{N_{\mathbf{X}}} P(l|\mathbf{x}_i) \qquad (1)$$

where local feature $\mathbf{x}_i$ belongs to input feature set $\mathbf{X}$, l is an object label and $N_{\mathbf{x}}$ is the number of input local features. The posterior $P(l|\mathbf{X})$ is approximated by the sum rule. We use the following binary probability model to design $P(l|\mathbf{x}_i)$:

$$P(l|\mathbf{x}_i) = \begin{cases} 1 & i \in l, S_M(i,l) \geq \epsilon \\ 0 & otherwise \end{cases} \qquad (2)$$

where $S_M(i,l)$ is called a feature-label map which represents the strength of the match between an input feature and its label based on neighboring context information. It is calculated as follows:

$$S_M(i,l) = \sum_{j \in N(i)} w(i,j) c(j,l) \qquad (3)$$

where $N(i)$ is the support region (or neighbors) of feature i. As shown in Figure 9, neighboring context information

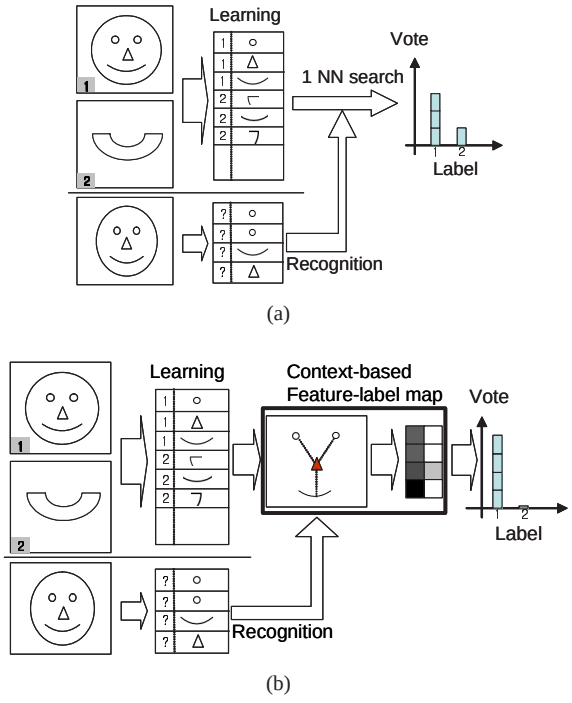


Figure 8. (a) Conventional classifier: direct voting of nearest neighbor. (b) Novel classifier: neighboring context-based voting.

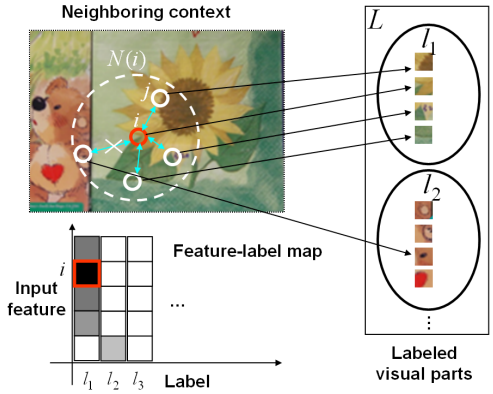


Figure 9. Concept of feature-label map (strength of match) generation using neighboring context information.

is aggregated by summing over the neighborhood using the above equation. $w(i,j)$ represents the support (contextual) weight at site j in the support region. The support from the neighborhood is valid when the neighboring visual parts have the same label (or object index) as the site of interest. This is the same concept as Gestalt properties, i.e., proximity and similarity. $c(j,l)$ is the goodness of match pair (j,l)

The support weight will be proportional to the probability, $p(O_j^l|O_i^l)$ where O_i^l represents the event that the label of a site i is l:

$$w(i,j) = k \cdot p(O_j^l|O_i^l) \qquad (4)$$

When we consider the fixed label l, it can be rewritten as:

$$w(i,j) = k \cdot p(l\mathbf{x}_i = l, l\mathbf{x}_j = l) \tag{5}$$

where $l\mathbf{x}_i$ represents the label of feature $\mathbf{x}_i$ and k is normalization factor. The probability of both features coming from the same object l is defined as:

$$p(l\mathbf{x}_i = l, l\mathbf{x}_j = l) = \frac{S(l\mathbf{x}_i = l)S(l\mathbf{x}_j = l)}{\sum_{m=1}^{N_L}\sum_{n=1}^{N_L} S(l\mathbf{x}_i = n)S(l\mathbf{x}_j = m)} \tag{6}$$

where $S(l\mathbf{x}_i = l)$ means the similarity that a local feature is mapped to a label l. This is equivalent to the χ^2 kernel value between input feature $\mathbf{x}_i$ and its nearest neighbor x^l_{1NN}, whose label is l:

$$S(l\mathbf{x}_i = l) = K_{\chi^2}(\mathbf{x}_i, x^l_{1NN}) \tag{7}$$

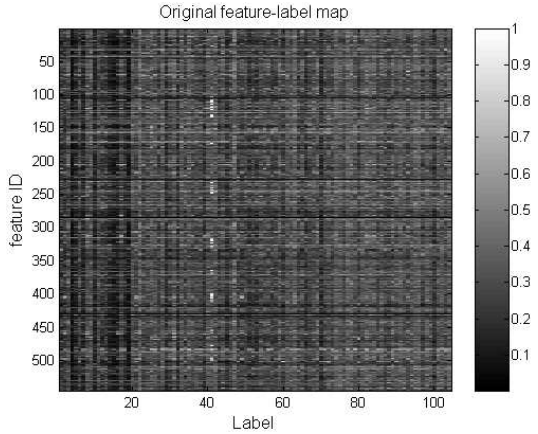
The χ^2 similarity kernel is defined as [28]:

$$K_{\chi^2}(\mathbf{x}, (y)) = exp\{-\rho\chi^2(\mathbf{x}, \mathbf{y})\} \tag{8}$$

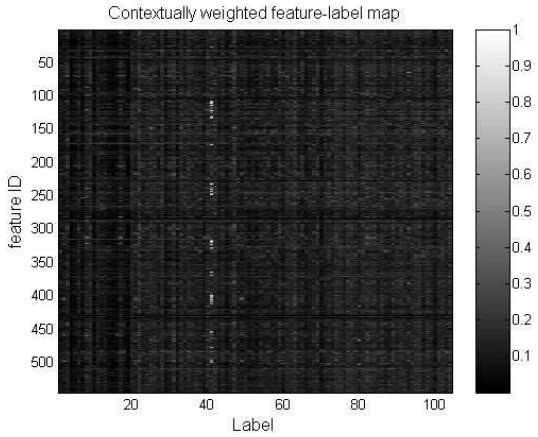
where ρ is set to 3$\sim$4 normally. Goodness of match, $c(j,l)$ is also defined as $S(l\mathbf{x}_j = l)$.

Although the feature-label map $S_M(i,l)$ is calculated using the above similarity kernel, it can be formulated in recursive form by simply reusing the feature label map in the next iteration, rather than the similarity kernel.

Figure 10(a) shows the feature label map calculated from the similarity kernel and Figure 10(b) shows the feature label map after applying the proposed method: contextually weighted feature-label map after two recursions. Figure 11 represents a close up view of Figure 10 and displays several feature locations. Note the power of neighboring contextual information, which enhances the strength of figural features (here, f.id: 527, 536) and suppresses the background features (here, f.id: 525).



(a)



(b)

Figure 10. (a) Initial feature-label map obtained from similarity kernel. (b) Contextually weighted feature-label map using the proposed method.

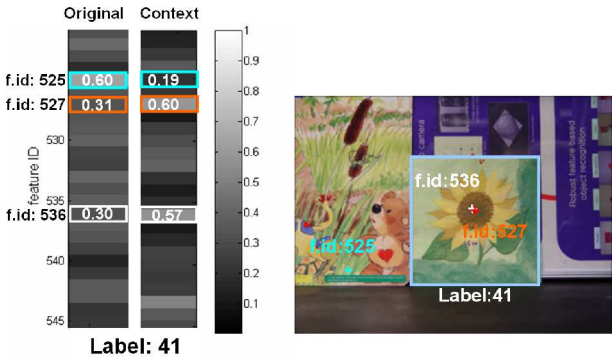


Figure 11. Close-up view of Figure 10 around Label 41 and corresponding feature locations.

5. Experimental results

Although there are several open object databases, such as Amsterdam DB, PASCAL DB, ETH-80 DB, and Caltech DB, we evaluate the proposed method using COIL-100 DB for feature comparison. We also use our DB, CMU DB (available at http://www.cs.cmu.edu/$\sim$stein/BSIFT) which contain background clutter, because our research goal is to recognize general objects in severely cluttered environments. Figure 12 (top) shows sample objects in our DB, composed of 21 textureless 2D objects, 26 textured 2D objects, 26 textureless 3D objects and 31 textured 3D objects. We stored each model's image in the canonical viewpoint. Because our research issue is how the contextual information enhances recognition, fixing a 3D viewpoint is reasonable. To effectively validate the power of the context-based object recognition method, we made test images as in Figure 12 (bottom). 104 objects are placed on a highly cluttered background. These DB and test images are acquired using a SONY F717 digital camera and are resized to 320$\times$240.

First, we compare the performance of the proposed G-RIF with SIFT, which is a state-of-the art descriptor [10],

Figure 12. (Top) examples of 104 general object models, (bottom) corresponding examples of 104 test objects on a highly cluttered background.

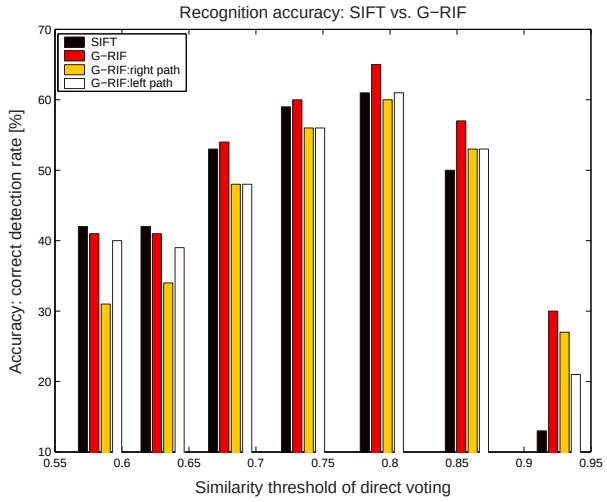


Figure 13. Comparison of NNC classification accuracy between G-RIF (left, right, both paths in Figure 3) and SIFT.

[14]. Typically, visual part detectors are evaluated only in terms of their repeatability. However, the evaluation should be in the context of a task, such as object labeling in this paper. We evaluate the features using the direct voting-based classifier which is used commonly as in Figure 8(a). We use the binary program offered by Lowe [10]. The accuracy of detection rate is used as a comparison measure which is widely used in classification or labeling problems. Figure 13 shows the evaluation results using our DB by changing the similarity threshold of direct voting. G-RIF shows better performance than SIFT and reveals the complementary properties of G-RIF: right path (gravity center point) and G-RIF: left path (high curvature point).

The 2nd and 3rd columns in Table 2 show the overall recognition results using the optimal classifier threshold (0.8). The recognition rate using G-RIF is higher than the SIFT feature by 5.77%. This good performance originates from the complementary properties of the proposed visual part detector and the effective spatial coding of multiple features in a unified way. In this test, we set the aperture level up to L3 and feature level by (1)+(2). Figure 14 shows several successful examples using G-RIF that are incorrect using the SIFT feature, although our feature dimension is 105 (84 for edge orientation + 21 for edge density) and SIFT's is 128.

Figure 14. Correct detection using G-RIF (failure cases using SIFT).

Feature	SIFT	G-RIF	G-RIF
Classifier	Direct voting	Direct voting	Contextual voting
# success /# test	62/104	68/104	74/104
Success rate [%]	59.61%	65.38%	71.15%

Table 2. Recognition accuracy (similarity threshold: 0.8) for 104 test images.

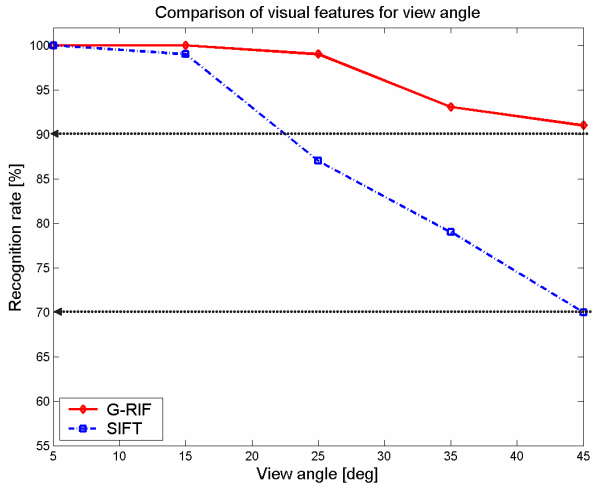


Figure 15. G-RIF vs SIFT on COIL-100 DB).

We also compared G-RIF and SIFT on COIL-100 DB. Since the DB is composed of multiple views with interval 5°, we use the frontal view (0°) as a model image and test 5 different views up to 45°. We use the NNC-based simple voting as a classifier. Figure 15 shows the evaluation results. Note that G-RIF shows upgraded performance by 20% at 45°view angle.

Next, we compared our context-based voting (Figure 8(b)) to the conventional nearest-neighbor based direct voting (Figure 8(a)). We used the same visual feature (G-RIF) proposed in this paper and our DB. The neighborhood size is set to two times the part scale. Overall test results are shown in Table 2 (3rd and 4th columns). The contextual voting-based classifier shows better recognition performance than direct voting by 5.77%. This good performance

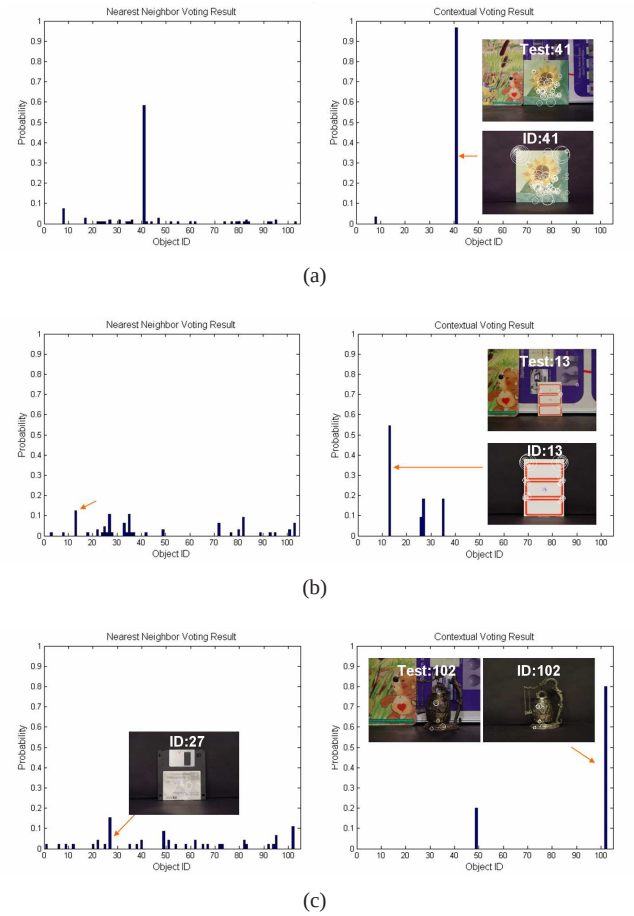


Figure 16. Recognition performance of nearest neighbor-based direct voting method (left column) and contextual voting method (right column): (a) Normal case, (b) ambiguous case and (c) failure using direct voting but success using ours.

originates from the effective use of the neighboring context. The contextual influence was explained in Figure 11.

Figure 16 shows three kinds of recognition examples with both methods. Figure 16(a) is a normal case. Both methods succeeded for textured objects. Figure 16(b) show ambiguous recognition results by direct voting but stable recognition results by contextual voting. Finally, Figure 16(c) shows the failure case by direct voting, but success with contextual voting. Figure 17 shows other results of Figure 16(c)'s case. Due to many ambiguous features on the cluttered background, the direct voting method labels the wrong object. However, the context-based voting method can reduce the effect of those cluttered features by neighboring contextual information.

Finally, we evaluated contextual and direct voting using the CMU database. This database is composed of 110 separate objects and 25 background images. We generated test images by changing objects' sizes on different backgrounds. This is equivalent to changing the background

Figure 17. Correct detection using contextual voting (failure cases using direct voting).

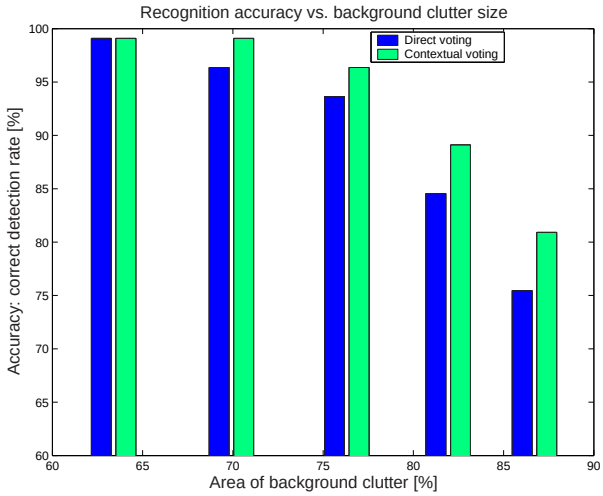


Figure 18. As the area of background clutter increases, contextual voting shows relatively higher detection accuracy than direct voting.

size. Figure 18 shows the evaluation results. Contextual voting shows an equal or better recognition rate than direct voting. Note that the neighboring contextual information has a more important role for object labeling in a severely cluttered background than in a less cluttered background.

6. Conclusions

In this paper, we propose a biologically motivated visual part selection method that extracts complementary visual parts. In addition, we propose a general contextual descriptor that encodes multi-cue information in a unified form. Recognition performance was improved by using the proposed feature. We also propose a simple and powerful object recognition method based on neighboring contextual information. Neighboring contextual information suppresses the strength of background features and enhances the strength of figural features in the feature-label map. The proposed method shows better recognition performance than other methods in a severe environment. We will extend the context-based voting method for general object-based image understanding such as labeled figure-ground segmentation.

Acknowledgements

This research has been partially supported by the Korean Ministry of Science and Technology for National Research Laboratory Program (Grant number M1-0302-00-0064) and by Microsoft Research Asia.

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and bject recognition using shape contexts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(24):509–522, 2002. 3

[2] G. M. Boynton. Adaptation and attentional selection. *Nature Neuroscience*, 7(1):8–10, 2004. 3

[3] O. Chomat, V. C. de Verdière, D. Hall, and J. L. Crowley. Local scale selection for gaussian based description techniques. In *European Conference on Computer Vision (ECCV'00)*, pages 117–133, London, UK, 2000. Springer-Verlag. 3

[4] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. pages 264–271, 2003. 1

[5] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04)*, pages 506–513, 2004. 1, 3

[6] S. Kim and I.-S. Kweon. Biologically motivated perceptual feature: Generalized robust invariant feature. In *Asian Conference on Computer Vision (ACCV'06)*, pages 305–314, 2006. 3

[7] M. Kouh and M. Riesenhuber. Investigating shape representation in area v4 with hmax: Orientation and grating selectivitiesn. Technical report. 3

[8] L. J. Latecki and R. Lakämper. Convexity rule for shape decomposition based on discrete contour evolution. *Computer Vision and Image Understanding*, 73(3):441–454, 1999. 2

[9] T. Lindeberg. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: a method for focus-of-sttention. *International Journal of Computer Vision*, 11(3):283–318, 1993. 1

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 1, 2, 3, 5, 6

[11] G. Loy and A. Zelinsky. Fast radial symmetry transform for detecting points of interest. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(8):959–973, 2003. 2

[12] S. Mahamud and M. Hebert. The optimal distance measure for object detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'03)*, 2003. 2

[13] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004. 1

[14] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005. 1, 6

[15] P. Moreels, M. Maire, and P. Perona. Recognition by probabilistic hypothesis construction. In *European Conference on Computer Vision (ECCV'04)*, pages 55–68, 2004. 1, 4

[16] K. Okada and D. Comaniciu. Scale selection for anisotropic scale-space: Application to volumetric tumor characterization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04)*, pages 594–601, 2004. 2

[17] A. Pasupathy and C. E. Connor. Shape representation in area V4: Position-specific tuning for boundary conformation. *Journal of Neurophysiology*, 86(5):2505–2519, 2001. 2

[18] D. Reisfeld, H. Wolfson, and Y. Yeshurun. Context-free attentional operators: The generalized symmetry transform. *International Journal of Computer Vision*, 14(2):119–130, 1995. 2, 3

[19] D. L. Ringach. Spatial structure and symmetry of simple-cell receptive field in macaque primary visual cortex. *Journal of Neurophysiology*, 88:455–463, 2001. 2

[20] C. Schmid. A structured probabilistic model for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'99)*, volume II, pages 485–490, June. 1, 4

[21] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000. 1, 2

[22] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005. 2

[23] A. Shokoufandeh, D. Macrini, S. Dickinson, K. Siddiqi, , and S. W. Zucker. Indexing hierarchical structures using graph spectra. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(7):1125–1140, 2005. 4

[24] A. Shokoufandeh, I. Marsic, and S. J. Dickinson. View-based object matching. In *IEEE International Conference on Computer Vision (ICCV'98)*, pages 588–595, 1998. 1

[25] A. Stein and M. Hebert. Incorporating background invariance into feature-based object recognition. In *Workshop on Applications of Computer Vision (WACV'05)*, pages 37–44, 2005. 1

[26] M. Tico and P. Kuosmanen. Fingerprint matching using an orientation-based minutia descriptor. *IEEE Tran. on Pattern Analysis and Machine Intelligence*, 25(8):1009–1014, 2003. 1, 3

[27] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *Ninth IEEE International Conference on Computer Vision (ICCV'03)*, pages 273–280, Washington, DC, USA, 2003. IEEE Computer Society. 1

[28] C. Wallraven, B. Caputo, and A. B. A. Graf. Recognition with local features: the kernel recipe. In *IEEE International Conference on Computer Vision (ICCV'03)*, pages 257–264, 2003. 4, 5

[29] R. H. Wurtz and E. R. Kandel. Perception of motion, depth and form. *Principles of Neural Science*, pages 548–571, 2000. 3

[30] L. Zhang, S. Z. Li, and Z. Y. Qu. Boosting local feature based classifiers for face recognition. In *CVPR Workshop on Face Processing in Video (FPIV'04)*. 4