Characterizing the SOM Feature Detectors Under Various Input Conditions

Macario O. Cordel II(✉) and Arnulfo P. Azcarraga

College of Computer Studies, De La Salle University,
2401 Taft Avenue, 1004 Manila, Philippines
{macario.cordel,arnulfo.azcarraga}@dlsu.edu.ph

Abstract. A classifier with self-organizing maps (SOM) as feature detectors resembles the biological visual system learning mechanism. Each SOM feature detector is defined over a limited domain of viewing condition, such that its nodes instantiate the presence of an object's part in the corresponding domain. The weights of the SOM nodes are trained via competition, similar to the development of the visual system. We argue that to approach human pattern recognition performance, we must look for a more accurate model of the visual system, not only in terms of the architecture, but also on how the node connections are developed, such as that of the SOM's feature detectors. This work characterizes SOM as feature detectors to test the similarity of its response vis-á-vis the response of the biological visual system, and to benchmark its performance vis-á-vis the performance of the traditional feature detector convolution filter. We use various input environments i.e. inputs with limited patterns, inputs with various input perturbation and inputs with complex objects, as test cases for evaluation.

Keywords: Feature detectors · Self-organizing maps · Multilayer perceptron · Pattern recognition

1 Introduction

The ability of living organisms to detect salient or target objects regardless of the background or lighting condition inspires most of the recent computational models for pattern recognition. For example, the results of the experiments to map the functional architecture of the monkey and cat's visual system [3,14,16] have been the bases for the layered architecture of successful machine vision systems [8,12,23]. Specifically, the Neocognitron [9,10] and the convolutional neural network (CNN) [17] networks rely on their layered architecture that are directly analog of the complex connection of neurons in the visual system.

With the success of CNN in pattern recognition [20], several applications have been proposed which leverage on the representational power of the trained feature detectors and address pattern recognition problems e.g. natural face

Q. Yang et al. (Eds.): PAKDD 2019, LNAI 11441, pp. 144–155, 2019.
https://doi.org/10.1007/978-3-030-16142-2_12

detection and recognition [19,22] and action recognition [2,24]. Recently, the Mask Regional-CNN [13] has shown to have promising performance in automatic object detection and segmentation. These current successes in such tasks raise the question: is pattern recognition a solved problem?

In this work, we argue that to approach the visual system pattern recognition performance, the computational model should consider the visual system behavior, including the manner in which the receptive fields are developed. The work in [5] proposes self-organizing maps (SOM)-based feature detectors for pattern recognition which exhibit competition-based development of its weight akin to the development of connection between neurons of the visual cortex [3,15]. They showed that indeed SOM feature detectors could be used in pattern recognition.

However, SOM feature detectors [5] performance remains to be inconclusive as its response to various cases e.g. constrained input environment, input perturbation and complex patterns has not been investigated, thus the proposed network is more of a blackbox. Evidences [3,15] show that the development of the visual system receptive fields, both simple and complex, have significant dependence on the kind of environment during its early development. That is, receptors which are exposed to a specific pattern, e.g. horizontal lines of different width, color, length, small distortion and the like, for a long time will develop receptors which are highly specialized to detect horizontal lines. The previous work [5] also has no verification that their proposed SOM feature detectors indeed capture the pattern information, i.e. the spatial relationship of pixels that form the patterns in the input image, as opposed to the convolution filter which was verified and quantified in [6].

This work examines the performance of SOM as the basic feature detectors for pattern recognition under various constrained and perturbed input environment. The classification performance of the feature detectors are determined and misclassification of patterns are carefully observed vis-á-vis the type of the training input patterns. In addition, the SOM feature detectors ability to extract pattern information is also verified by gradually removing the pixel spatial relationship of the input pattern. The classification performance of the two feature detectors are evaluated using simple (MNIST [17]) and complex patterns (vehicle dataset). MNIST is commonly used to evaluate the potential of several proposed pattern recognition algorithms, e.g. [4,17,21]. The following are the contributions of our work:

- *We verified that the proposed SOM feature detectors exhibit similar decline in pattern recognition performance of the visual system, when trained with limited input pattern during the training phase.* Towards biological visual system pattern recognition performance, we argue that it is important to have a similar visual system characteristics, not only in terms of the layered architecture, but also in terms of how the connections are developed. We discovered that increasing the number and proper positioning of the receptive fields result in robustness of the classifier from a set of feature detectors exposed to limited input pattern.

- *We show that for various input perturbation, the classifier using SOM feature detector was able to correctly classify the input.* We showed that the canonic forms help in the equivariance as the input conditions change and the entity is rotated over the appearance manifold or the SOM receptive field. Although convolution filter has better performance, the accuracy of SOM feature detectors at this early stage shows potential.

2 Related Works

Early studies [3,15] revealed that the development of the visual system exhibits competition by limiting the input patterns that pass through the visual pathway during their development. For example in an experiment [15] which sutured the left eyelid of a kitten from birth, results show profound cell atrophy in layers receiving light input from the covered eye. A similar experiment was performed [3] which sutured the left eye of an infant monkey to observe the effect in the development of its striate cortex. Their experiment showed that the population of cells favors the open eye and the earlier and longer the exposure of the open eye, the greater shift in eye preference is observed. Further investigation reveals that eyes which are exposed to a certain pattern e.g. vertical stripes have more cells orientation like the input pattern.

Several feature detectors which exhibit competition during their development have been proposed as an alternative to convolution filters. SOM nodes are used as the feature map nodes in CNN architecture in [18]. They additionally introduce algorithm which determine the locations of high density data. Another work is proposed by Arevalo et al. [1] which uses topographic independent component analysis (TICA) for CNN learning. Their system is used in detecting basal cell carcinoma in medical images. A semi-supervised learning proposed by Dong et al. [7] uses Sparse Laplacian Filtering learning (SLFL) during the training of the convolution layer. This allows much less data during the network training. Their proposed work is applied to vehicle type classification which takes in high resolution video frames. These systems, however, presented the feature detectors in black box, i.e. did not show any verification as to how these extract the pattern information.

Current image pattern recognition systems [13,22,24] based on convolution filters show promising performance and demonstrate applicability to several pattern recognition tasks e.g. speech and text recognition. We argue that to continue to approach human-level pattern recognition performance, we should probe various computational models that better represent the biological visual system. We investigate further in this paper SOM as competition-based feature detectors for pattern recognition.

3 Test Setup

Three experiments are conducted in this work. The first experiment simulates the constrained input environment conducted to observe the response of the

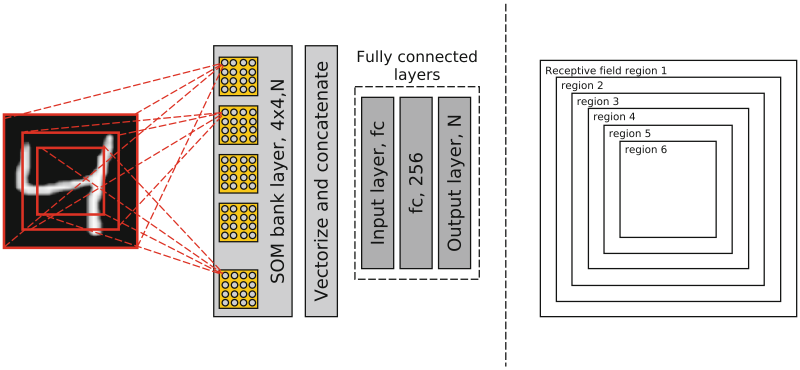


Fig. 1. Pattern recognition with SOM feature detectors (left). Six (not explicitly shown) SOMs are used as feature detector. The output of the feature detector is a similarity vector whose elements are equal to the cosine similarity of the node weights and the pixels in the corresponding limited domain. Shown at the right is the arrangement of the limited domains for each SOM feature detector.

visual system. It uses limited input pattern in training the feature detectors namely (a) vertical only, (b) horizontal only, and (c) circular only. The second experiment determines the robustness of the SOM feature detectors to input perturbation particularly (a) rotation, (b) variation of pattern size and stroke thickness and (c) affine transformation. The third experiment observes the pattern extraction capability of the feature detectors. The spatial relationships of the pixels which form the digit, e.g. curves and edges, are gradually removed by randomly repositioning the pixels. Note that the repositioning is random with respect to the image, but the new pixel arrangement is fixed for the training and test images [6].

3.1 Architecture

Shown in Fig. 1 (left) the architecture with SOM feature detectors under test. The map size of these SOM feature detectors is set to have 4×4 nodes. Six SOMs are used as the feature detectors which are trained separately from the classifier. The SOMs are arranged such that the limited domain or the receptive field for each SOM are focused at the center of the image as illustrated in Fig. 1 (right). This type of limited domain arrangement is used as visual system receptive fields encode the information at the center of the scene [11].

The performance of this architecture is then compared to the conventional feature detector with a single layer of convolution filters and mean pooling kernels. Twenty (20) 9 × 9 convolution kernels and 20 mean pooling kernel are used for the conventional architecture. A simple CNN architecture is used as we are after the comparison of the basic behavior of the two feature detectors, i.e. comparing deeper CNN will be unfair for the SOM feature detector. For both networks using convolution filters (CNN) and SOM feature detectors, the

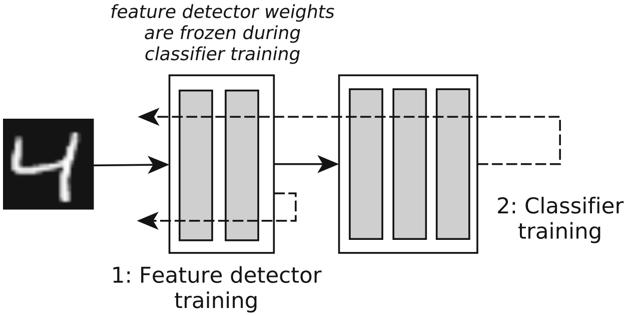

feature detector weights are frozen during classifier training
1: Feature detector training
2: Classifier training

Fig. 2. Training sequence for the evaluation of the feature detector trained in limited input environment. The feature detectors are first trained using specific patterns. Afterwards, the feature detector weights are frozen and the classifier training proceeds. In the classifier training, input patterns are unconstrained.

number of hidden nodes in the fully-connected layer is 256. For digit classification task, the output node is 10 while for vehicle classification task, the output node is 3.

3.2 Training and Testing

In the first experiment, the feature detectors are trained separately from the classifier to allow the learning on limited pattern i.e. vertical, horizontal and circular patterns (first training). Afterwards, the learned feature detector weights are frozen, and the classifier are then trained to classify the handwritten digits (second training) as illustrated in Fig. 2. MNIST test set is used to evaluate the performance of this experiment.

In the second experiment, the feature detectors are trained together with the classifier using the MNIST training set for 60 epochs and batch size of 256. Afterwards, we use the test sets called the rotated-NIST, the size-NIST and the Affine-NIST[1] to evaluate the robustness of the SOM feature detectors and to compare it with the convolution filter. The rotated-NIST are simply the MNIST test set randomly rotated from $-15°$ to $15°$, the size-NIST are generated from the test set with varying dilation and resizing of the digit.

Finally, in the third experiment, the feature detectors are trained and the classifier using the MNIST for the digit classification and the vehicle training set for the vehicle classification. For each randomization of the pixel position (please refer to [6]), the two architectures are retrained using the new set repositioned pixels. In this way, we ensure that the pixel values as attributes of the dataset are preserved, and only the spatial correlation of the pixel forming the patterns are removed.

[1] Downloaded from: https://www.cs.toronto.edu/~tijmen/affNIST/.

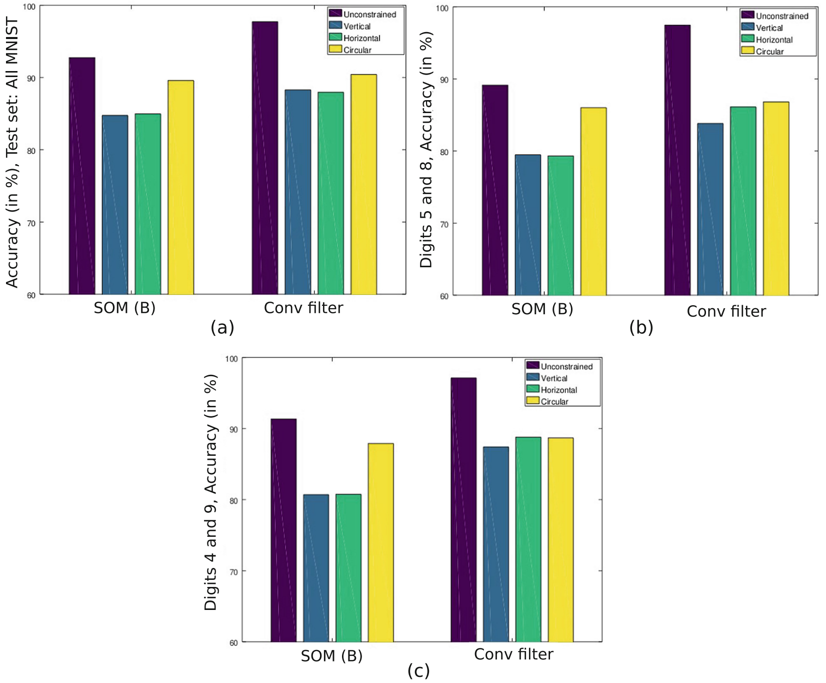


Fig. 3. The performance of the pattern recognition models with SOM feature detectors whose receptive fields are the SOM and the convolution filters. The average accuracies in classifying the MNIST digits using unconstrained input pattern are significantly higher when the feature detectors are trained using limited patterns. Also shown the commonly misclassified pairs digits 5 and 8 and digits 4 and 9.

4 Results and Analysis

4.1 Performance in Constrained Input Environment

Generally, as summarized in Fig. 3(a), the accuracies of SOM feature detectors and the convolution filter feature detector show expected decrease in classification ability when the feature detectors are trained with limited patterns. This was first observed in the visual system which manifests as cell atrophy when the visual receptors are exposed to limited environment during its early development stage [3,15]. The results also imply that both types of feature detectors exhibit similar response as that of the biological system visual receptors developed in a constrained environment.

Interestingly, however, for SOM feature detectors whose receptive fields are distributed all over the input image (shown in Fig. 4 (left)), the ability to identify the input pattern seems to be robust, see Fig. 4 (right), despite the limited training pattern of the feature detectors, presumably due to the richness of the sampled viewing domain or receptive fields. In the confusion matrix of these

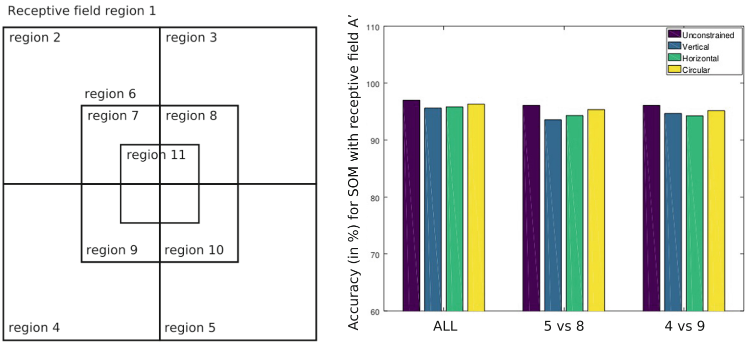


Fig. 4. Another experiment is conducted using distributed receptive fields as opposed to the previous center-focused receptive fields. Although each SOM has limited domain in the input, the distributed domain allows the architecture to look at the other parts of the input pattern rather than concentrating at the center. The respective average accuracies (right) for classifying all the digits trained using unconstrained and constrained patterns show that distributing the receptive field of the feature detector increases the robustness to the dependency to limited input patterns.

three evaluation setups, digit pair 5 and 8; and digit pair 4 and 9 are often confused and misclassified in the center-focused SOM receptive fields and convolution filter receptive fields, but not in distributed SOM receptive fields. The misclassification between the digit pairs are more common in center-focused SOM feature detectors and convolution filter (refer to Fig. 3(b) and (c)). Intuitively, these pairs (digits 5 and 8, digits 4 and 9) are very similar when the horizontal or vertical components are missing.

Both SOM and the convolution filter as feature detector exhibits similar behavior as biological detectors, such that when the detectors are exposed to limited pattern, the pattern recognition ability drops significantly. However, when SOM feature detectors have distributed receptive field, the classifier does not experience significant decrease in pattern recognition (see the three graphs of Fig. 3 A') – which implies that this type of receptive field arrangement overcomes the limitation of the biological visual system which only develops feature detectors depending on the input environment. For our experiments, the classifier whose feature detectors are arranged to focus on the center details, shows dependence on what the feature detectors have seen. For the case where the SOM feature detectors are distributed across the input pattern, the classifier has more viewing points or spatial sampling which add to the classifier input information.

Using the commonly confused input digit pair 5 and 8, we compared the feature maps of the digit pairs for the two receptive field arrangements, shown in Fig. 5 for the distributed and in Fig. 6 center-focused receptive fields. These feature maps are rendered from the element-by-element multiplication of the input pattern and the node's weight vector with the highest response or activation.

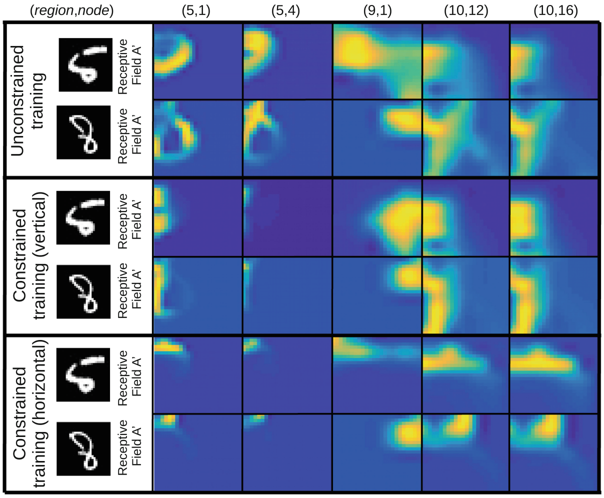


Fig. 5. The feature maps from distributed receptive fields (or arrangement A') trained using the 10 digits from MNIST (first two rows), trained using vertical patterns only (3rd to 4th rows) and trained using horizontal patterns only (5th to last row). The feature maps are formed by rendering the input image (digits 5 and 8) and the weight vector of the corresponding receptive field node with the highest response. The unconstrained feature maps of 8 and 5 has significant difference as compared with the corresponding constrained feature maps.

Note that rendered portion in Figs. 5 and 6 which are near yellow implies high value of similarity of the input pattern and the canonical weight value, while the feature map portions which are near blue means approaching zero similarity. In the test setup when the training is unconstrained, using both receptive field arrangements, the feature maps of 5 and 8 with significant activation differ visually when rendered, see the first two rows of Figs. 5 and 6.

For the constrained input pattern however, the difference between the rendered feature maps of 5 and 8 decreases significantly for the center-focused receptive field (see Fig. 6 trained with vertical and horizontal patterns) but not in distributed receptive field (see Fig. 5 trained with vertical and horizontal patterns) – thus, the classifier with center-focused receptive fields frequently fails to discriminate the input pattern of digits 5 and 8.

4.2 Performance in Perturbed Input Patterns

SOM as feature detectors was able to allow the classifier to detect the input pattern with small random rotation from −15° to 15°. For distributed receptive fields the accuracy is 97.97% and for the center-focused receptive fields the accuracy is 89.39% both of which are comparable to the performance when there is input no rotation. This implies that SOM feature detectors performance is robust to small rotation. For various stroke size and thickness, however, a significant

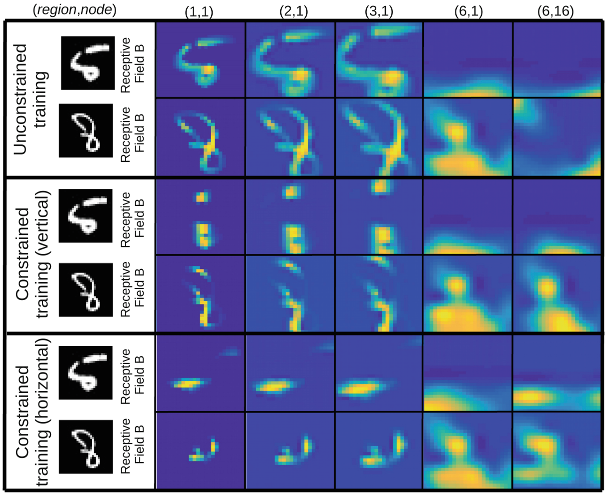


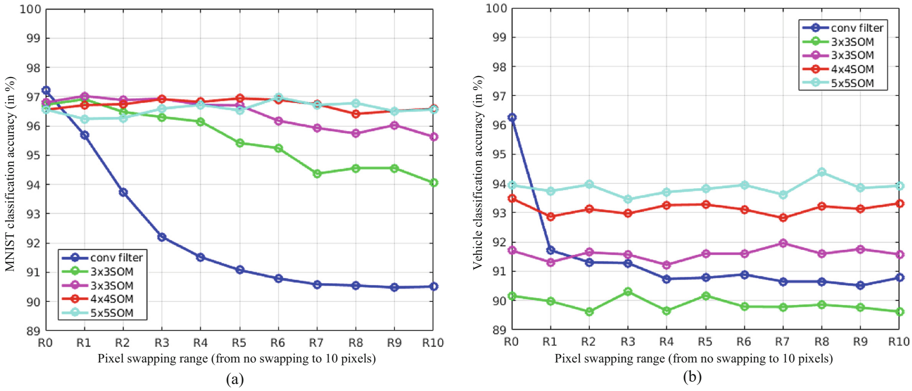
Fig. 6. The feature maps from center-focused feature detectors (or arrangement B) trained using the 10 digits from MNIST (first two rows), trained using vertical patterns only (3rd to 4th rows) and trained using horizontal patterns only (5th to last row). The unconstrained feature maps of 8 and 5 has significant difference as compared with the corresponding constrained feature maps.

Table 1. Average accuracy of 4 × 4 SOM feature detectors with distributed receptive field (A'), center-focused receptive field (B) and convolution filter feature detectors, under for various perturbed and complex patterns

Dataset	SOM (A')	SOM (B)	Conv filter
Perturbed NIST (small rotation)	97.97%	89.39%	96.08%
Perturbed NIST (stroke thickness)	72.59%	62.80%	96.98%
Affine-NIST	10.41%	9.25%	39.38%

decrease in classification accuracy is seen for both receptive field arrangements. For the distributed receptive fields, the accuracy drops to 72.59% and for the center-focused receptive fields, the accuracy becomes 62.80% only. This performance becomes even worse when the input patterns went through affine transformation. For both arrangements of receptive fields for SOM feature detectors, the accuracy was no better than chance (Table 1).

For small rotation the convolution filter was able to detect 96.08%, which is lower than the accuracy of the distributed SOM receptive field. For various size and thickness however, the convolution filters were able to extract the needed information for the classifier to achieve the accuracy of 96.98%. Finally, for the Affine-NIST, convolution filters show accuracy of 39.38% which is much better than the SOM feature detectors.



(a)

(b)

Fig. 7. As the spatial correlation of the (a) MNIST and the (b) vehicle datasets are gradually removed, the primitive pattern information e.g. edges and corners are also gradually removed. The plot shows gradual decrease in CNN performance using convolution filter as the pattern information from left to right are gradually removed for both datasets. The classifier using SOM as feature detectors shows consistent accuracy even in the absence of the input pattern.

4.3 Performance in Complex Vehicle Dataset

We also verify the performance of SOM feature detectors in complex dataset i.e. vehicle dataset. In addition to this, we gradually remove the pattern from the input image and observe the classification performance of SOM A' and SOM B. Previous work [6] shows that by gradually randomizing the position of the pixels in an image, while fixing these new randomized position of the pixels for all the images in the dataset, the primitive pattern information e.g. edges and corners, are removed while retaining the pixel value information as the only image attribute.

Figure 7 shows the classification accuracy of SOM feature detectors as for (a) MNIST and (b) vehicle datasets as the pixel positions are randomized to remove the resemblance of the object, from left to right. The convolution filter feature detector of the CNN shows this dependence to the spatial correlation of the pixel forming the edges as shown by the gradual decrease in the classification accuracy for both the MNIST and vehicle dataset. The consistent performance of the classifier using SOM as feature detector implies that SOM feature detector is robust to such removal of spatial correlation of pixels. However, as R increases from left to right, no canonical information, as to the kind of input patterns, could be obtained when the weights of the SOM nodes are rendered.

5 Discussion

We performed the evaluation of SOM as feature detectors for different input environment conditions. We showed that both the SOM and convolution filters suffers misclassification if these detectors are trained under constrained input environment. Particularly, feature detectors trained on vertical patterns could only extract vertical patterns and feature detectors trained on horizontal patterns could only extract horizontal patterns from the input image, such that any differentiating traits between two categories other than the vertical (or the horizontal) pattern, are not regarded. Although SOM feature detectors exhibit this behavior of the biological feature detectors, we discovered that the arrangement of the receptive fields of SOM feature detectors allows the classifier to be robust to the removal of the primitive patterns, e.g. edges and curves, in the input image.

SOM feature detectors also have better robustness to small rotation of the input pattern as compared to the convolution filter. However, for various stroke sizes and thickness and affine transformation of the input pattern, the convolution filter shows better performance. SOM feature detectors show promising results however when the resemblance of primitive patterns e.g. edges and curves, are slowly removed.

The remarkable performance of SOM feature detectors over the conventional convolution filters exhibits its potential. SOM feature detectors are still far from perfect. Examining the different receptive field arrangements and tweaking the connection of this feature detector to fit the dynamic routing algorithm [21] could help SOM feature detectors reach its full potential.

References

1. Arevalo, J., Cruz-Roa, A., Arias, V., Romero, E., Gonzalez, F.: An unsupervised feature learning framework for basal cell carcinoma image analysis. Artif. Intell. Med. **64**, 131–145 (2015)
2. Bilen, H., Fernando, B., Gavves, E., Vedaldi, A., Gould, S.: Dynamic image networks for action recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3034–3042, June 2016
3. Carlson, M., Hubel, D.H., Wiesel, T.N.: Effects of monocular exposure to oriented lines on monkey striate cortex. Dev. Brain Res. **25**(1), 71–81 (1986)
4. Ciresan, D., Meier, U., Schmidhuber, J.: Multi-column deep neural networks for image classification. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3642–3649 (2012)
5. Cordel, M.O., Antioquia, A.M.C., Azcarraga, A.P.: Self-organizing maps as feature detectors for supervised neural network pattern recognition. In: Hirose, A., Ozawa, S., Doya, K., Ikeda, K., Lee, M., Liu, D. (eds.) ICONIP 2016. LNCS, vol. 9950, pp. 618–625. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46681-1_73
6. Cordel, M.O., Azcarraga, A.P.: Measuring the contribution of filter bank layer to performance of convolutional neural networks. Int. J. Knowl.-Based Intell. Eng. Syst. **21**(1), 15–27 (2017)
7. Dong, Z., Wu, Y., Pei, M., Jia, Y.: Vehicle type classification using semisupervised convolutional neural network. IEEE Trans. Intell. Transp. Syst. **16**, 2247–2256 (2015)
8. Fu, M., Xu, P., Li, X., Liu, Q., Ye, M., Zhu, C.: Fast crowd density estimation with convolutional neural networks. Eng. Appl. Artif. Intell. **43**, 81–88 (2015)

9. Fukushima, K.: Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern. **36**, 193–202 (1980)
10. Fukushima, K.: Artificial vision by multi-layered neural networks: neocognitron and its advances. Neural Netw. **37**, 103–119 (2013)
11. Haines, D.E., Mihailoff, G.A.: The visual system. In: Fundamental Neuroscience for Basic and Clinical Applications, Chap. 20. Elsevier (2018)
12. Haoxiang, L., Zhe, L., Xiaohui, S., Jonathan, B., Gang, H.: A convolutional neural network cascade for face detection. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5325–5334 (2015)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the International Conference on Computer Vision (ICCV) (2017)
14. Hubel, D.H., Wiesel, T.N.: Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. J. Physiol. **106**, 106–154 (1962)
15. Hubel, D.H., Wiesel, T.N.: Effects of visual deprivation on morphology and physiology of cells in the cats lateral geniculate body. J. Neurophysiol. **26**, 978–993 (1963)
16. Hubel, D.H., Wiesel, T.N.: Receptive fields and functional architecture in two nonstriate visual areas of the cat. J. Neurophysiol. **28**, 229–289 (1965)
17. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**, 2278–2324 (1998)
18. Mohebi, E., Bagirov, A.: A convolutional recursive modified self organizing map for handwritten digits recognition. Neural Netw. **60**, 104–118 (2014)
19. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference (2015)
20. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**(3), 211–252 (2015)
21. Sabour, S., Frosst, N., Hinton, G.: Dynamic routing for between capsules. In: Advances in Neural Information Processing Systems, pp. 3859–3869 (2017)
22. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: CVPR, pp. 815–823. IEEE Computer Society (2015)
23. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: OverFeat: integrated recognition, localization and detection using convolution networks. In: International Conference on Learning Representations (2014)
24. Tu, Z., et al.: Multi-stream CNN: learning representations based on human-related regions for action recognition. Pattern Recogn. **79**, 32–43 (2018)