Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 369613, 6 pages
http://dx.doi.org/10.1155/2014/369613

Research Article

Sampling Based Average Classifier Fusion

Jian Hou,[1] Wei-Xue Liu,[1] and Hamid Reza Karimi[2]

[1] *School of Information Science and Technology, Bohai University, Jinzhou 121013, China*
[2] *Department of Engineering, Faculty of Engineering and Science, University of Agder, 4898 Grimstad, Norway*

Correspondence should be addressed to Jian Hou; dr.houjian@gmail.com

Received 22 December 2013; Accepted 22 January 2014; Published 24 February 2014

Academic Editor: Xudong Zhao

Classifier fusion is used to combine multiple classification decisions and improve classification performance. While various classifier fusion algorithms have been proposed in literature, average fusion is almost always selected as the baseline for comparison. Little is done on exploring the potential of average fusion and proposing a better baseline. In this paper we empirically investigate the behavior of soft labels and classifiers in average fusion. As a result, we find that; by proper sampling of soft labels and classifiers, the average fusion performance can be evidently improved. This result presents sampling based average fusion as a better baseline; that is, a newly proposed classifier fusion algorithm should at least perform better than this baseline in order to demonstrate its effectiveness.

1. Introduction

Object classification is an important task in pattern recognition. Due to the difference in lighting conditions, viewing angles and occlusions, and so forth, there usually exist large intraclass diversity and interclass similarity in real image datasets. This presents great challenges to designing practical object classification systems. While many feature detectors, descriptors, and classification algorithms have been proposed in literature, it is evident that none of these algorithms is able to generate satisfactory classification results for real image datasets. In this case, classifier fusion and feature combination [1] are proposed to combine the decisions of multiple complementary classifiers and produce better performance than any single classifier. In this paper we focus on classifier fusion.

Majority voting is one of the most simple algorithms in classifier fusion. This algorithm uses only the class labeling and discards the probability information of the labels and thus may lead to performance loss. In order to make use of the class probability, average fusion combines the posterior probability of all training classes, that is, the soft labels. Some popular algorithms in this aspect also include weighted sum [2], logistic regression [3], Dempster-Shafer rules [4], and neural networks [5]. In this paper we focus on image classification. However, the classifier fusion algorithms are also applicable to other domains [6–8].

In proposing a new classifier fusion algorithm, researchers usually choose to compare it with average fusion to show the advantage of the new algorithms. While being simple, average fusion assigns equal weights to all classifiers regardless of their powerfulness. Intuitively this harms the discriminative power of this algorithm and then makes the claimed advantage of newly proposed classifier fusion algorithms less convincing. With this consideration in mind, in this paper we empirically investigate the impact of soft labels and classifiers on classifier fusion performance. As a result, we find that the behaviors of soft labels and classifiers in average fusion can be explained in the framework of kNN classification. This framework gives rise to a sampling based average fusion algorithm, which is shown to outperform the ordinary average fusion evidently in experiments on four diverse image datasets. This result enables us to believe that our sampling based average fusion algorithm explores the potential of average fusion and qualifies as a better baseline. A newly proposed algorithm should be compared with this new baseline to demonstrate its advantage.

The remainder of this paper is organized as follows. In Section 2 we introduce the experimental setups used in our classifier fusion experiments. Sections 3 and 4 present

the details of our work on investigating the behaviors of soft labels and classifiers in average fusion, respectively. In Section 5 we present the sampling based average fusion algorithm based on the experimental results in Sections 3 and 4. Finally, Section 6 concludes the paper.

2. Experimental Setups

We use SVM in classification experiments on four diverse datasets. The regulation parameter C is fixed to be 1000 and the multiclass SVM is trained in a one-versus-all manner. In all experiments we test with 10 different training-testing splits and report the average of recognition rates.

2.1. Datasets. We use the following four datasets in experiments.

The Event-8 dataset [10] contains images from 8 categories of sports events. Each category is composed of 130 to 250 images with different lighting conditions and postures and so forth. Following the experimental setup in [10], we randomly select 70 images per class as training and another 60 images as testing and report the overall recognition rate.

The Scene-15 dataset [11] is composed of images from 15 scene categories with 200 to 400 images in each category. We use the same experimental setup as in [15], that is, randomly selected 100 images per class as training and all the others as testing, and report the mean recognition rate per class.

Oxford Flower-17 dataset [16] consists of 1360 flower images evenly distributed in 17 categories. Similar as in [16], we randomly select 40 images per class as training examples and 20 images as testing images. The overall accuracy is reported as the results.

With the well-known Caltech-101 dataset [15], we use 30 images per class for all the 102 classes in training, and select up to 15 images per class in the remaining for testing. The mean recognition rates per class are reported as the results.

2.2. Features. We use the following features to build the kernels used in SVM classification. These features are popular due to their discriminative power in object classification, for example, in [13, 17, 18]. This makes our conclusions drawn from experiments convincing and meaningful.

PHOG Shape Descriptor. We construct oriented (20 bins) and unoriented (40 bins) PHOG descriptors [19] from level 0 to 3 and obtain 8 descriptors in total. Unlike the implementation in [19], in this paper the descriptor in level L is formed only by its 2^L windows.

Bag-of-SIFT. The SIFT descriptors [20] on patches of radius r with spacing of 8 pixels are extracted and quantized into a 500-bin vocabulary, and we select $r = 4, 8, 12, 16$ to allow for scalability. These descriptors are extracted in gray space for Scene-15 dataset which contains only gray images and, in gray, HSV and CIE-Lab spaces for Flower-17, Event-8, and Caltech-101. We build the visual words histograms from level 0 to 2 and obtain 3 or 9 descriptors.

Locally Binary Patterns. The histograms of the basic locally binary patterns (LBP) [21] are adopted from level 0 to 2.

Gist Descriptor. We extract the global gist descriptor [22] from level 0 to 1.

Self-Similarity Descriptor. The self-similarity descriptors [23] of 30 dimensions (10 orientations and 3 radial bins) are extracted and used to build a 500-bin vocabulary. The histograms are then built from level 0 to 2.

Gabor and RFS Filters. We use two texture features, that is, Gabor and RFS filters [23], to build histograms of 500 bins from level 0 to 2.

Gray Value Histogram. We also use the 64-bin gray value histograms from level 0 to 3.

For all these features, we use χ^2 distance to build kernels in the form of $k(x, y) = \exp(-d_0^{-1} d(x, y))$, where d is the pairwise distances and d_0 is the mean of pairwise distances. Here χ^2 distance is selected due to its great distinctive power, as illustrated in [13, 24–26].

3. Behavior of Soft Labels

In majority voting, each classifier assigns only one label with the largest probability to the testing image. We count the times of each label being selected and adopt the label with the maximum times as the correct one. This approach discards the probability of each label, which may be useful in classifier fusion. Therefore soft labels, that is, the posterior probability of each training label, are proposed to be used in classifier fusion. Between the two extremes, that is, using only the most probable label and using all soft labels in fusion, we are interested to know if it is possible to achieve better performance by adopting a sample of all soft labels.

We evaluate the impact of soft labels sampling on average fusion performance as follows. For each classifier, we sort all labels in descending order according to their posterior probability. Then we use in average fusion only the top k labels, that is, the labels corresponding to k largest probabilities, where k ranges from 1 to the number of all training labels. The experimental results are reported in Figure 1.

It is evident from Figure 1 that, for average fusion, neither adopting only the most probable label nor using all the soft labels is the best choice. Instead, several most probable soft labels generate the best classification results. This is a little similar to the kNN classification framework as the top k most probable soft labels produce the best classification results. Although the best k is different for 4 datasets, $k = 2$ seems an appropriate option as it produces the best or near-best performance for all 4 datasets.

Another interesting observation is that, with the increase of object categories, the performance gain obtained using k most probable soft labels instead of all soft labels is enlarged. From Event-8 to Caltech-101, the performance gain ranges from 0.1 to 10 roughly. This indicates the importance of soft labels sampling, especially for large datasets with a large number of object categories. On the other hand, this

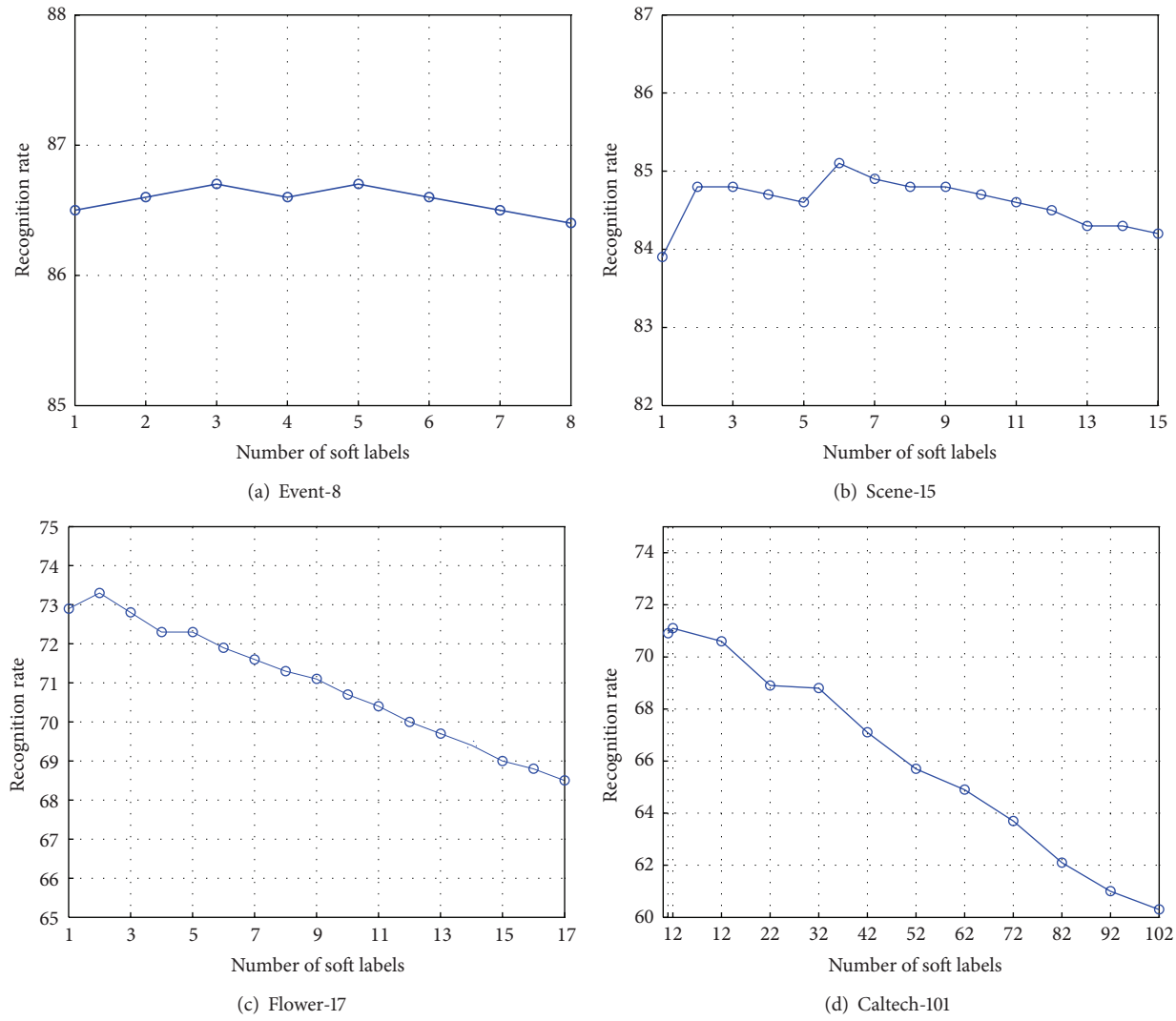

Figure 1: Recognition rates from classifier fusion with different numbers of most probable soft labels.

observation highlights the necessity of exploring the potential of average fusion and proposing a better baseline algorithm.

4. Behavior of Classifiers

As classifier fusion is to use multiple classifiers to improve classification performance, another problem of interest is if more classifiers definitely lead to better average fusion performance.

We evaluate the impact of the amount of classifiers on average fusion performance as follows. Firstly, we use the recognition rate of 10-fold cross-validation to estimate the powerfulness of each classifier. In the second step, we sort the classifiers in descending order according to the powerfulness. Then we add the classifiers into fusion one by one and record the fusion performance. The performance with different amounts of classifiers is reported in Figure 2. Note that, in this experiment, we firstly fuse the classifiers from different levels of the same features, for example, all 3 levels of LBP, and regard the fused decision as of one single classifier. In this way we have 11 classifiers for Caltech-101, Event-8, and Flower-17 and 9 classifiers for Scene-15. This is to compare different classifiers (features) more evidently.

Similar as in the case of soft labels, Figure 2 shows that with average fusion, the best performance is obtained with several most powerful classifiers. Adding more classifiers of less powerfulness into fusion only decreases the final classification performance. It is easy to see that $k = 4$ can be an appropriate selection for the number of classifiers.

5. Sampling Based Average Fusion

In the last two sections we find that using a small sample of most probable soft labels and most powerful classifiers separately helps produce the best fusion performance. Although the optimal number of soft labels, that is, 2, and the optimal number of classifiers, that is, 4, are obtained empirically, they are applied to all the four datasets without special

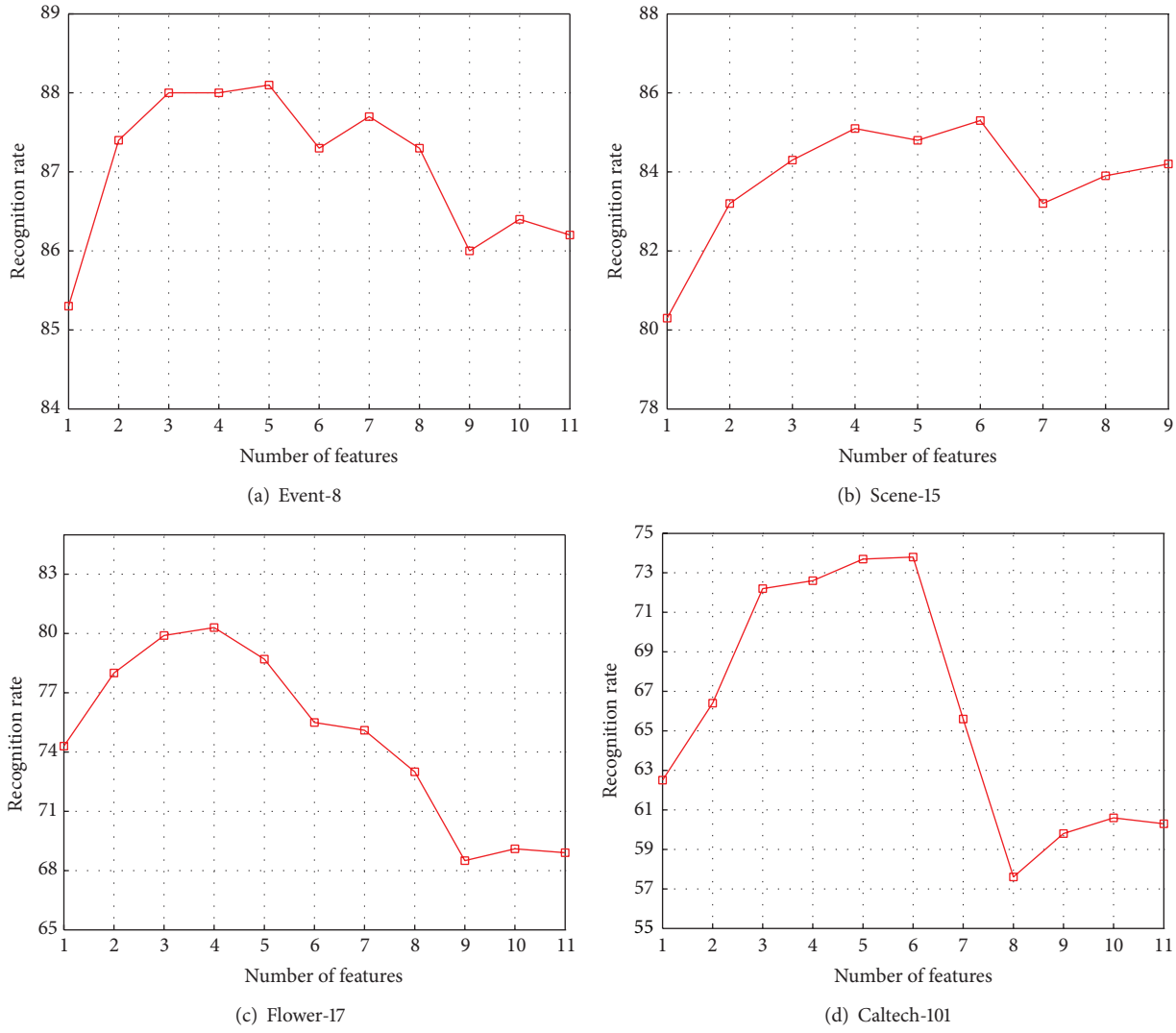

Figure 2: Recognition rates from classifier fusion with different numbers of most powerful classifiers.

tuning to individual datasets. Now we test the combined performance of the sampling of both soft labels and classifiers. In experiments on the four datasets, we compare the average fusion performance with and without sampling of soft labels and classifiers and show the results in Tables 1 and 2. In the tables "average1" means recognition rates from average fusion with all classifiers and all soft labels, whereas "average2" indicates corresponding results with sampling of soft labels and classifiers, that is, 2 most probable soft labels and 4 most powerful classifiers. We also compare our algorithm with the state-of-the-art ones on these datasets.

From the comparison we see that, with average fusion, using a small sample of soft labels and classifiers always produces a significant improvement in object classification performance. This means that the sampling based average fusion (SBAF) can serve as a better baseline than ordinary average fusion. Our algorithm performs also comparably to the state-of-the-art ones on these datasets. In fact, on Event-8 and Scene-15 our algorithm produces better results than the state-of-the-art ones, and on Flower-17 and Caltech-101 our results are close to the best ones to date. Noticing that in experiments we only use simple features and average fusion, we believe that this is a very encouraging result which validates the effectiveness of our SBAF algorithm. Since in this paper we present SBAF as a better baseline but not a novel fusion method, we only compare this algorithm with the

Table 1: Event-8 and Scene-15 recognition rates and comparison.

Event-8		Scene-15	
Method	Accuracy	Method	Accuracy
Best single	84.6 ± 1.7	Best single	79.3 ± 0.7
Average1	86.4 ± 1.7	Average1	84.2 ± 0.4
Average2	**87.9 ± 1.3**	Average2	**85.0 ± 0.5**
[9]	84.2 ± 1.0	[9]	84.1 ± 0.5
[10]	73.4	[11]	81.4 ± 0.5

Table 2: Flower-17 and Caltech-101 recognition rates and comparison.

Flower-17		Caltech-101	
Method	Accuracy	Method	Accuracy
Best single	75.1 ± 1.5	Best single	66.2 ± 1.2
Average1	77.5 ± 1.7	Average1	58.8 ± 1.7
Average2	86.0 ± 1.5	Average2	71.0 ± 1.2
[12]	**88.3 ± 0.3**	[13]	**77.8 ± 0.4**
[13]	85.5 ± 3.0	[14]	66.2 ± 0.5

ordinary average fusion and not with other fusion methods, for example, [2, 4].

Another observation from experiments is that the behaviors of soft labels and classifiers can be explained in the framework of *k*NN classification. Regarding the most probable soft labels and most powerful classifiers as the nearest neighbors, we can explain all the observations from experiments based on the *k*NN framework easily. This framework provides theoretical support to our following conclusions. Firstly, the best performance of average fusion is not achieved with all soft labels and all classifiers, but with a sample of most probable soft labels and most powerful classifiers. This gives rise to SBAF as a better baseline. Secondly, with a dataset of tens to hundreds of categories, the performance gain of SBAF over average fusion can be rather large (over 10 for Caltech-101). Since in modern time there is an explosive increase in the amounts and categories of images, this observation highlights the importance of soft label and classifier sampling and the necessity to adopt SBAF as the baseline.

Although in this paper we focus our work on image classification, the idea of classifier fusion is also useful to some other related domains, for example, document classification, speech recognition, and fault diagnosis [27–29]. In the next step we plan to explore the possibility of extending the work to more domains [30–32].

6. Conclusion

In this paper we investigated the impact of soft labels and classifiers sampling on average classifier fusion performance through experiments on four diverse datasets. As a result, we found that the behaviors of soft labels and classifiers in average fusion can be elegantly explained in the framework of *k*NN classification. This framework further gives rise to a sampling based average fusion method, that is, using a sample of most probable soft labels and most powerful classifiers in fusion to obtain the best performance. Experiments indicate that this sampling based average fusion performs evidently better than the ordinary one and thus can serve as a better baseline to be compared with. Our results on the four datasets are also comparable to the state of the art in literature. As the *k*NN framework elegantly captures the behaviors of soft labels and classifiers in classifier fusion, we believe that it can be helpful in designing novel classifier fusion methods.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant no. 61304102), Natural Science Foundation of Liaoning Province of China (Grant no. 2013020002), and Scientific Research Fund of Liaoning Provincial Education Department under Grant no. L2012400.

References

[1] X. Li, W. Yang, and J. Dezert, "An airplane image target's multi-feature fusion recognition method," *Acta Automatic Sinica*, vol. 38, pp. 1298–1307, 2012.

[2] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "Adaptive classifier integration for robust pattern recognition," *IEEE Transactions on Systems, Man, and Cybernetics B*, vol. 29, no. 6, pp. 902–907, 1999.

[3] T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 1, pp. 66–75, 1994.

[4] L. Xu, A. Krzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, no. 3, pp. 418–435, 1992.

[5] D. S. Lee, *Theory of classifier combination: the neural network approach [Ph.D. thesis]*, SUNY, Buffalo, NY, USA, 1995.

[6] S. Yin, S. Ding, A. Haghani, H. Hao, and P. Zhang, "A comparison study of basic datadriven fault diagnosis and process monitoring methods on the benchmark tennessee eastman process," *Journal of Process Control*, vol. 22, pp. 1567–1581, 2012.

[7] S. Yin, X. Yang, and H. R. Karimi, "Data-driven adaptive observer for fault diagnosis," *Mathematical Problems in Engineering*, vol. 2012, Article ID 832836, 21 pages, 2012.

[8] S. Yin, G. Wang, and H. Karimi, "Data-driven design of robust fault detection system for wind turbines," *Mechatronics*, 2013.

[9] J. X. Wu and J. M. Rehg, "Beyond the Euclidean distance: creating effective visual codebooks using the histogram intersection kernel," in *Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV '09)*, pp. 630–637, October 2009.

[10] L.-J. Li and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, pp. 1–8, October 2007.

[11] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, vol. 2, pp. 2169–2178, June 2006.

[12] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proceedings of the 6th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP '08)*, pp. 722–729, December 2008.

[13] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 221–228, 2009.

[14] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "SVM-KNN: discriminative nearest neighbor classification for visual category recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 2126–2136, June 2006.

[15] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from fewtraining examples: an incremental bayesian approach tested on 101 object categories," in *Workshop on Generative-Model Based Vision (CVPR '04)*, p. 178, 2004.

[16] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 1447–1454, June 2006.

[17] J. Yang, Y. N. Li, Y. H. Tian, L. Y. Duan, and W. Gao, "Group-sensitive multiple kernel learning for object categorization," in *Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV '09)*, pp. 436–443, October 2009.

[18] J. Hou, W. X. Liu, and H. R. Karimi, "Exploring the best classification from average feature combination," *Abstract and Applied Analysis*, vol. 2014, Article ID 602763, 7 pages, 2014.

[19] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR '07)*, pp. 401–408, July 2007.

[20] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[21] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[22] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[23] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, June 2007.

[24] J. Hou, B. P. Zhang, N. M. Qi, and Y. Yang, "Evaluating feature combination in object classification," in *Proceedings of the International Symposium on Visual Computing*, pp. 597–606, 2011.

[25] J. Hou and M. Pelillo, "A simple feature combination method based on dominant sets," *Pattern Recognition*, vol. 46, pp. 3129–3139, 2013.

[26] J. Hou, W. X. Liu, E. Xu, Q. Xia, and N. M. Qi, "An experimental study on the universality of visual vocabularies," *Journal of Visual Communication and Image Representation*, vol. 24, pp. 1204–1211, 2013.

[27] S. Yin, H. Luo, and S. Ding, "Real-time implementation of fault-tolerant control systems with performance optimization," *IEEE Transactions on Industrial Electronics*, vol. 61, pp. 2402–2411, 2014.

[28] A. Baak, M. Muller, G. Bharaj, H.-P. Seidel, and C. Theobalt, "A data-driven approach for real-time full body pose reconstruction from a depth camera," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*, pp. 1092–1099, November 2011.

[29] S. Yin, S. X. Ding, A. H. A. Sari, and H. Hao, "Data-driven monitoring for stochastic systems and its application on batch process," *International Journal of Systems Science*, vol. 44, no. 7, pp. 1366–1376, 2013.

[30] X. Zhao, L. Zhang, P. Shi, and M. Liu, "Stability and stabilization of switched linear systems with mode-dependent average dwell time," *IEEE Transactions on Automatic Control*, vol. 57, no. 7, pp. 1809–1815, 2012.

[31] X. Zhao, H. Liu, and J. Zhang, "Multiple-mode observer design for a class of switched linear systems linear systems," *IEEE Transactions on Automation Science and Engineering*, 2013.

[32] X. Zhao, L. Zhang, P. Shi, and H. R. Karimi, "Novel stability criteria for t-s fuzzy systems," *IEEE Transactions on Fuzzy Systems*, vol. 21, pp. 1–11, 2013.