In Defense of Active Part Selection for Fine-Grained Classification[1]

D. Korsch[a],* and J. Denzler[a],**

[a]*Computer Vision Group, Friedrich Schiller University Jena, Jena, Germany*

** e-mail: dimitri.korsch@uni-jena.de*

*** e-mail: joachim.denzler@uni-jena.de*

Abstract—Fine-grained classification is a recognition task where subtle differences distinguish between different classes. To tackle this classification problem, part-based classification methods are mostly used. Part-based methods learn an algorithm to detect parts of the observed object and extract local part features for the detected part regions. In this paper we show that not all extracted part features are always useful for the classification. Furthermore, given a part selection algorithm that actively selects parts for the classification we estimate the upper bound for the fine-grained recognition performance. This upper bound lies way above the current state-of-the-art recognition performances which shows the need for such an active part selection method. Though we do not present such an active part selection algorithm in this work, we propose a novel method that is required by active part selection and enables sequential part-based classification. This method uses a support vector machine (SVM) ensemble and allows to classify an image based on arbitrary number of part features. Additionally, the training time of our method does not increase with the amount of possible part features. This fact allows to extend the SVM ensemble with an active part selection component that operates on a large amount of part feature proposals without suffering from increasing training time.

Keywords: fine-grained recognition, SVM, ensemble, bagging

DOI: 10.1134/S105466181804020X

1. INTRODUCTION

With the growing digitalization the amount of digital images rapidly increases. Consequentially, the need for precise automatic classification of these images was never so present. The newest hardware and software developments in the field of computer vision solve challenges from the last decades with ease and new challenges with more sophisticated tasks are created. It can be noticed on the latest benchmark datasets for object recognition like CUB-200-2011 [14]. This dataset consists of bird images from 200 species. The large number of different species and the fact that some species are only characterized by subtle differences in their appearance create a need for novel classification approaches. Therefore, fine-grained algorithms get more and more attention.

The classical way of extracting features from the entire image achieves remarkable results [6], but most related works [4, 5, 8, 11] on fine-grained image classification use a part-based approach. The main idea is to combine the global feature with additional part features extracted from different locations of the image. Based on the extracted part features and the global feature the image is classified.

One of the drawbacks of considering all existing part features is that in some cases the features extracted to distinguish different classes do not contribute to the classification decision. Furthermore, the additional information in form of these part features often leads to an increased training complexity, since all features are considered regardless of their impact on the classification result. A possible solution is a part selection component which actively selects the part features for the classification process with the objective to increase the classification performance. This is similar to former research on optimal sensor data selection with the help of a sequential decision process [3].

An active part selection component requires a classifier that is capable of processing an arbitrary number of part features. Hence, we suggest in this work a sequential classification method that observes the part features one after another and classifies them. Additionally, since the classifier takes only one part feature as input, its training time does not increase with the number of existing part features. This allows to train a classifier on a large amount of part features. Since the active part selection component is not the focus of this work, we simulate the part selection by computing the classification results for all possible part feature combinations. With this simulation we show that given such a selection component the upper bound of the suggested method's recognition performance lies far above the current state-of-the-art part-based classification approaches.

[1] The article is published in the original.

Received July 9, 2018

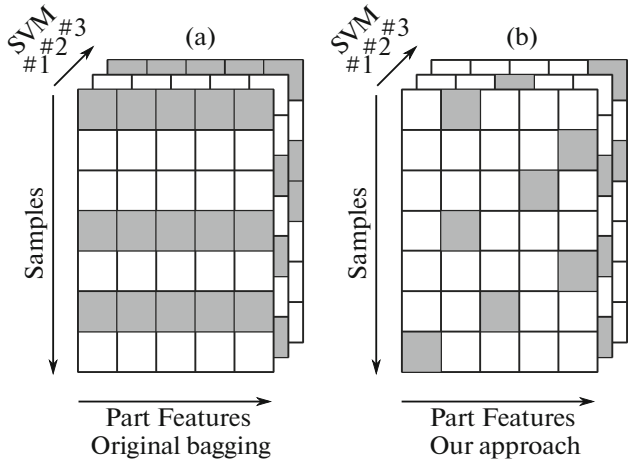

Fig. 1. Two SVM bagging methods. (a) In the original bagging algorithm each SVM trains on a random subset of training samples. In case of part-based approach for fine-grained recognition a sample consists of all part features. (b) We suggest to train each SVM on all samples, but to select a random part feature for each sample.

The sequential classification is realized with an SVM ensemble, more specifically SVM bagging [2]. Works like [7, 15] have already used SVM ensembles to either improve the classification performance or decrease the training complexity. These approaches as well as the original bagging algorithm train the single classifiers on a random subset of samples. In contrast, we suggest to train a single SVM classifier on all training samples, but randomly select a single part feature for each sample, as shown in Fig. 1.

The resulting ensemble of weak classifiers performs a classification on a single part feature, creating a classification decision for each classifier in the ensemble. In the end, the ensemble decisions for one part feature are aggregated to a single decision. Furthermore, the classification based on a part feature combination can be performed iteratively by observing one part feature after each other. This also allows each combination to have a different and also arbitrary number of part features.

2. RELATED WORK

2.1. Part-Based Fine-Grained Recognition

Most of the part based fine-grained recognition methods [4, 5, 8, 11] fit a part detection model, which identifies part locations. At these locations, part features are extracted and the object is classified based on these features.

Works like [5, 11] use a constellation or pose based approach. The authors estimate relative positions of some part proposals and find best groupings based on the training images. Based on these groupings, which defines the part model, the positions are estimated and used for feature extraction.

Jaderberg et al. [4] and Liu et al. [8] on the other hand train multiple parallel networks. Each of these networks operates on one part and either transforms it [4] or extracts features around its position [8]. Similarly, Zheng et al. [16] trains a multi-attention network and extracts attentions based on the attentions. These methods are trained end-to-end.

All of these methods perform a passive part selection since the parts are either selected based on the location information [5, 11] or totally independent from each other [4, 8, 16]. In this work we suggest a simulation of an active part feature selection component that computes the decisions based on prior knowledge about the observed image. Additionally, all of the mentioned methods have not investigated the upper bound recognition performance of their part-based approaches.

An active part selection component could be realized with recurrent attention models (RAMs) [1, 9, 10]. RAMs are iterative attention models that predict positions where the system should extract features next based on previously seen information. This information can be seen as prior knowledge, but all our experiments with RAMs so far could not outperform the previously mentioned state-of-the-art methods.

2.2. SVM Bagging

Bagging predictors were introduced by Breiman [2]. It is a "method for generating multiple versions of a predictor and using these to get an aggregated predictor." In the original paper a single predictor was trained on a random subset of the training data, as shown in Fig. 1a. The performance of such resulting classifier is mediocre, but if the prediction results are aggregated, for instance with a majority voting, then such an classifier ensemble is able to outperform a single classifier trained on the whole training dataset.

Wang et al. [15] illustrate also some empirical results on using SVMs as base classifier. In their work the authors were able to show some minor advantages of SVM bagging compared to regular SVMs and other classifiers. Linghu et al. [7] have used in their work SVM ensembles to reduce the training complexity and improve the recognition performance of their system. Additionally, they have evaluated different aggregation methods for the ensemble decisions.

3. METHOD

3.1. Part Feature Extraction

Our part feature extraction builds on the work of Simon et al. [11]. As shown in Fig. 2, the authors estimate a part constellation model in unsupervised manner with the help of gradient maps of a pre-trained CNN. Then, they sort the part locations by their fitting score to the constellation model. Finally, the best ten part locations are selected and another pre-trained

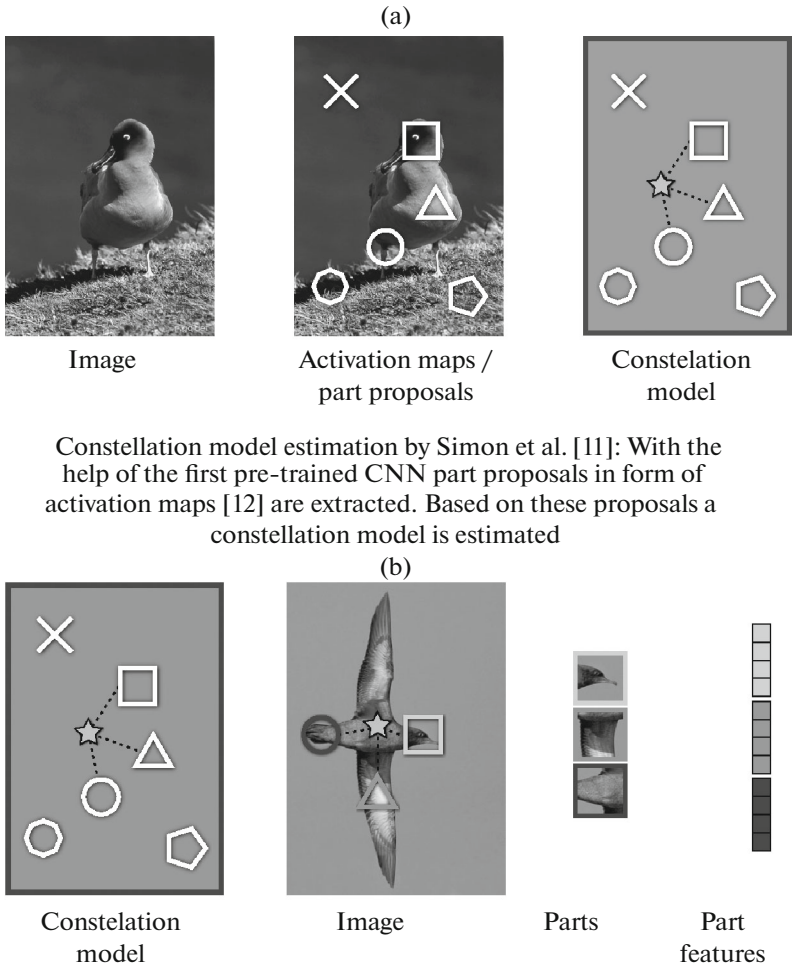

Fig. 2. Part feature extraction method. First the constellation model is estimated and based on this model part features are extracted (square: head, circle: leg, triangle: wing).

CNN extracts the part features on these locations. The extraction is performed on two different scales, resulting in 20 part features. These part features are concatenated together with the global feature to one single feature vector and a single SVM is trained.

3.2. SVM Ensemble Training

In this work, we consider the part features as distinct features, since we assume that not all present part features impact the classification result in a positive way. In order to allow an active part selection, the fine-grained classifier has to be able to handle changing number of part features as input. Hence, we propose an ensemble of SVMs as a sequential classifier.

We utilize the bagging algorithm [2], but instead of randomly selecting a subset of training samples, we select for each sample a random part feature, as illustrated in Fig. 1. After an arbitrary amount of SVMs has been trained, the aggregated ensemble decision for a single part feature is computed by a majority voting over the ensemble classifiers.

Next, we define how to compute a prediction for a given part feature combination $\langle f \rangle_t = \langle f_1, f_2 \ldots f_t \rangle$. Given such part feature combination and an classifier k we perform a maximum likelihood estimation under the uniform prior assumption:

$$c_{\text{est}} = \operatorname*{argmax}_c p_k(c|\langle f \rangle_t) \tag{1}$$

$$= \operatorname*{argmax}_c \prod_{x=1}^{t} p_k(c|f_x) \tag{2}$$

$$= \operatorname*{argmax}_c \sum_{x=1}^{t} \log p_k(c|f_x). \tag{3}$$

This way of class estimation for a specific part feature combination allows compute the decisions in the

Table 1. Comparison of our work with other part-based methods on the CUB-200-2011 dataset

Method			Accuracy, %	No. of parts
Simon et al. [11]			81.0	21
Krause et al. [5]			82.0	31
Jaderberg et al. [4]			84.1	4
Liu et al. [8]			84.3	3
Zheng et al. [16]			86.5	5
Ours	Ground truth parts		77.6	16
	Random part selection (GT parts)		77.3 ± 0.2	4
	Best combination (GT parts)	overall	79.8 ± 0.1	5 ± 0.2
		per class	87.4 ± 0.1	2.12 ± 0.03
		per sample	92.8 ± 0.1	3.9 ± 0.02
	All parts (single scale)		77.4%	11
	Random part selection (single scale)		77.4 ± 0.3	4
	Best combination (single scale)	overall	79.1 ± 0.1	5.3 ± 0.6
		per class	**85.4** ± 0.1	1.97 ± 0.02
		per sample	**89.4** ± 0.1	3.77 ± 0.09
	All parts		78.4	21
	Random part selection		78.2 ± 0.2	4
	Best combination	overall	80.8 ± 0.1	9.3 ± 0.7
		per class	**87.2** ± 0.2	2.06 ± 0.02
		per sample	**92.1** ± 0.2	3.74 ± 0.07

form of class probabilities for each part feature separately and fuse these decisions. It also allows the usage of the SVM ensemble presented in this work as a classifier k, because the input of the SVM ensemble is a single part feature.

3.3. Part Feature Combination Analysis

The second part of the work analyzes the resulting classifier ensemble. Therefore, we check how well the ensemble performs when it classifies the images based on all part features as well as only on a specific subset of the part features. As already mentioned before, we simulate the active part selection component by computing the classification results for all possible part feature combinations. Since we use 20 part features, we compute the predictions for $2^{20} = 1048576$ part feature combinations. With these combinations we are able to estimate the upper bound for the recognition performance for these part features. First, we select the best combination for all training samples, then best combinations for each class and finally for each single sample separately. This way we simulate three different active part selection methods. The first one can be constructed without any prior knowledge about the observed image. The later two methods would use a prior information about the given image in form of a class distribution or any other sample related information.

4. EXPERIMENTS AND RESULTS

In our experiments we used the CUB-200-2011 [14] dataset. The SVM ensemble consists of 48 classifiers. The number of the classifiers can be chosen arbitrary. In our case, 48 SVMs could be trained in parallel most efficiently on our testing machine, a dual-socket server with two Intel Xeon E5-2650 v4 processors. Further investigation may also be needed to check the impact of this hyperparameter on the classification performance.

The experiments were repeated ten times to observe the variance of the approach. Every test run took seven hours for training the classifier ensemble and computing predictions for all possible part feature combinations.

As mentioned before, we are using the ten part locations matching best the constellation model of Simon et al. [11]. On these location we extract the part features with a VGG19 [13] network from the last pre-classification fully connected layer (*fc7*). This results in a single feature having the dimensionality of 4096. Like Simon et al. [11], we extract the features from the locations on two different scales, which yields 20 part features. In our experiments we also consider a setup with a single scale, namely only ten part features. Furthermore, in order to observe how good our method works with perfectly selected parts, we considered a setup with ground truth parts.

The results of the mentioned setups can be seen in Table 1. First, we compare our baseline results where all parts were selected with the reference work of Simon et al. [11]. Here, the original work clearly yields better performance. Same observation can be seen if we select one combination that performed best for all samples or if we randomly select four arbitrary parts. The original strategy of using all part features still performs slightly better. However, the selected combination contains less than half of the additional information, namely only nine part features instead of all 21 (20 part and one global feature).

If we go further and select the best combination for each class or for each sample separately, then we can clearly outperform not only the reference work of Simon et al., but also other part-based approaches. The best combinations achieve state-of-the-art performances not only on the more comprehensive two-scale setup, but also in the single scale scenario. Thereby, at most only four parts are selected from the given 21, 16, or 11. It is important to notice that the

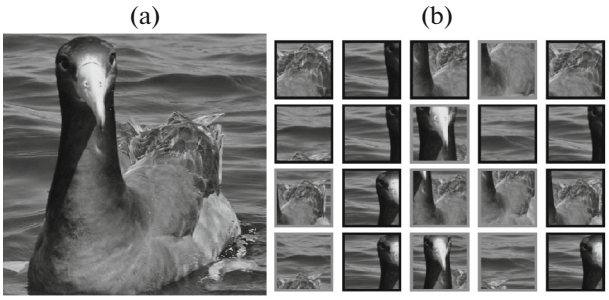

Fig. 3. Example of an incorrectly classified image, if all of the parts are selected. If only the marked (light gray) parts are selected, the classification becomes correct.

additional scale does not impact the mean number of selected parts, but the performance. Hence, we conclude that the extraction scale is as important as the active part selection.

Furthermore, in the setup with ground truth parts, we can also observe a performance improvement. As expected, the results with these parts are the best. Nevertheless, similar to the previous setups, the amount of selected parts drops to roughly four. Again, this argues in favor of an active part selection, even if the part detector is already perfect.

Additionally, we have visualized how extracted parts may distract the classifier. In Fig. 3 a sample from the CUB-200-2011 [14] dataset is shown on the left (Fig. 3a). On the right, Fig. 3b illustrates the parts that were extracted with the constellation model of Simon et al. [11]. The first two rows are the ten parts extracted with the one scale and the last two rows are extracted with the second scale. In this particular example, if all of the parts are selected, the image is classified incorrectly. The prediction becomes correct if only the marked parts are selected.

This qualitative result shows how active part selection can improve the classification by leaving out non-informative parts. It illustrates that the unsupervised part extraction algorithm used in this work is not perfect. Some of the extracted parts do not even cover the observed object (second row, first and fourth column) or only cover a the object marginally (second column). On the other hand, the parts leading to the correct prediction mostly cover the object. The final understanding why these parts lead to a correct decision and how to select them automatically is an open research objective.

CONCLUSIONS

In this work we presented a novel approach for sequential fine-grained classification. The sequential classifier is an ensemble of SVMs implementing the bagging algorithm. In contrast to previous works on bagging algorithm, we used random part features instead of a random sample subset for the training of the ensemble. The advantage of the presented algorithm is the independence of the SVM ensemble from the number of the observed part features.

To compare our approach with other part-based fine-grained classification methods, we simulated an active part feature selection component. The simulation was achieved by computing the classification result for all possible part feature combinations. Afterwards, best combinations were estimated for all samples, for each class and for each sample separately. These best combinations represent an upper bound recognition performance given specific part features.

Based on the part features of Simon et al. [11] we showed that the upper bound recognition performance is way above the current state-of-the-art results. This leads to the conclusion that an active part feature selection component is the best way to improve the fine-grained classification.

REFERENCES

1. J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *CoRR*, abs/1412.7755 (2014). https://arxiv.org/abs/1412.7755
2. L. Breiman. "Bagging predictors," Mach. Learn. **24** (2), 123–140 (1996).
3. J. Denzler and C. M. Brown. "Information theoretic sensor data selection for active object recognition and state estimation," IEEE Trans. Pattern Anal. Mach. Intell. **24** (2), 145–157 (2002).
4. M. Jaderberg, K. Simonyan, A. Zisserman, et al. "Spatial transformer networks," in *Advances in Neural Information Processing Systems 28: Proc. Annual Conf. NIPS 2015* (Montreal, Canada, 2015), pp. 2017–2025.
5. J. Krause, H. Jin, J. Yang, and L. Fei-Fei, "Fine-grained recognition without part annotations," in *Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) (Boston, MA, 2015), pp. 5546–5555.
6. J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei, "The unreasonable effectiveness of noisy data for fine-grained recognition," in *Computer Vision–ECCV 2016, Proc. 14th European Conf., Part II*, Ed. by B. Leibe et al., Lecture Notes in Computer Science (Springer, Cham, 2016), Vol. 9906, pp. 301–320.
7. B. Linghu and B.-Y. Sun, "Constructing effective SVM ensembles for image classification," in *Proc. 2010 3rd International Symposium on Knowledge Acquisition and Modeling* (*KAM*) (Wuhan, China, 2010), IEEE, pp. 80–83.
8. X. Liu, T. Xia, J. Wang, and Y. Lin, "Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition," arXiv:1603.06765 (2016). https://arxiv.org/abs/ 1603.06765
9. V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Advances in Neural Information Processing Systems 27: Proc. Annual*

Conf. NIPS 2014 (Montreal, Canada, 2014), pp. 2204–2212.

10. P. Sermanet, A. Frome, and E. Real, "Attention for fine-grained categorization," arXiv:*1412.7054* (2014). https://arxiv.org/abs/1412.7054
11. M. Simon and E. Rodner, "Neural activation constellations: Unsupervised part model discovery with convolutional networks," in *Proc. 2015 IEEE Int. Conf. on Computer Vision* (*ICCV*) (Santiago, Chile, 2015), pp. 1143–1151.
12. M. Simon, E. Rodner, and J. Denzler, "Part detector discovery in deep convolutional neural networks," in Computer Vision–ACCV 2014, Proc. 12th Asian Conference on Computer Vision, Ed. by D. Cremers et al., Lecture Notes in Computer Science (Springer, Cham, 2014), Vol. 9004, pp. 162–177.
13. K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition," *CoRR*, abs/1409.1556 (2014). https://arxiv.org/abs/1409.1556
14. C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, *The Caltech-UCSD Birds-200-2011 Dataset*, Technical Report CNS-TR-2011-001 (California Institute of Technology, 2011).
15. S.-J. Wang, A. Mathew, Y. Chen, L.-F. Xi, L. Ma, and J. Lee, "Empirical analysis of support vector machine ensemble classifiers," Expert Syst. Appl. **36** (3), Part 2, 6466–6476 (2009).
16. H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. 2017 IEEE Int. Conf. on Computer Vision* (*ICCV*) (Venice, Italy, 2017), pp. 5219–5227.

Dimitri Korsch is a research associate in the Computer Vision Group at Friedrich Schiller University Jena, Germany. He received his BSc and MSc degree in IT-Systems Engineering from University of Potsdam in 2013 and 2016, respectively. His research interests include unsupervised learning, reinforcement learning as well as fine-grained visual categorization.

Joachim Denzler earned the degrees "Diplom-Infor-matiker," "Dr.-Ing." and "Habilitation" from the University of Erlangen, Germany, in years 1992, 1997, and 2003, respectively. Currently, he holds a position as full professor for computer science and is head of the Computer Vision Group at the Friedrich Schiller University Jena, Germany. He is also Director of the Michael Stifel Center for Data-Driven and Simulation Science, Jena. His research interests comprise the automatic analysis, fusion, and understanding of sensor data, especially development of methods for visual recognition tasks and dynamic scene analysis. He contributed in the area of active vision, 3D reconstruction, as well as object recognition and tracking. He is author and co-author of over 300 journal and conference papers as well as technical articles. He is a member of IEEE, IEEE computer society, DAGM, and GI.