Mid-level Action Representation with Association Rule Mining for Still Images

Ozge Yalcinkaya
ozge@cs.hacettepe.edu.tr

ABSTRACT

Finding discriminative and representative parts of images is essential in order to describe an object or an action. The benefits of these distinctive parts on image classification has been proven in many studies. In this work, we find specific parts of action images by utilizing a proposed method on this topic which focuses on association rules between feature dimensions of images. We call these parts as mid-level visual elements and we choose them with respect to two properties: discriminativeness and representativeness. Image patches that include the computed association rule are selected to be representative for their action category. Then, we use those patches for a whole action representation. We experiment on a subset of Stanford 40 Action Dataset and we extract GoogLeNet activations for feature definition. According to both qualitative and quantitative results, association rule mining is useful to determine representative action parts from images.

1 INTRODUCTION

Action recognition is a challenging and widely studied problem because of the varieties of human motions for each different action. With the developments of deep neural networks, image classification and object recognition problems are highly solved on large scale datasets such as ImageNet [7, 12, 17, 19]. Similarly, human activity recognition has been improved with proposed deep architectures such as 2D, 3D or two steam networks [11, 16, 21] generally for videos.

However, even deep level representations may not be enough to describe the action due to the challenge of recognizing the activity itself from the still images. Action recognition from still images is highly based on finding clues from human poses or their interactions with objects. Therefore, discriminative methods are suggested which are proposed for image classification or object recognition [3, 18]. According to results, instead of using the entire image, extracting some meaningful patches and using them in classification or image representation gives higher prediction accuracies than traditional whole image classification for the still image action problem.

This patches, which are called as mid-level visual elements, are selected to be discriminative and representative for that category of action. Representativeness means mid-level visual elements should frequently occur in the target category, while discriminativeness implies that they should be visually discriminative against the natural world that contains other actions.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

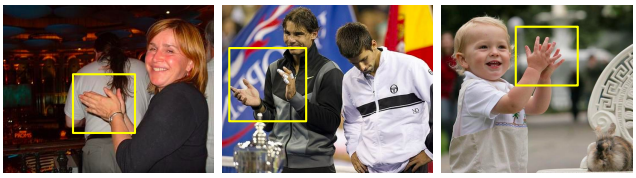

Figure 1: Still action images from Stanford40 dataset's "applauding" category. The yellow boxes show the representative and discriminative patches for applauding action.

For example in Figure 1, for the given actions, representative parts are showed in a box. Here, we can see that for the "applauding" action, hands are more representative than the other parts. Therefore, using this patch in order to recognize the applauding action will be more beneficial than using the entire image which can give some redundant and misleading information for that action.

For that purpose, many different approaches are introduced for finding discriminative parts (see Section 2). In this work, we follow the same idea of [13], which uses association rule mining to determine mid-level elements. Association rule mining is a widely used pattern mining method which discovers the if-then rules in given transactions.

In the context of mid-level visual element discovery, as noted by [2], finding discriminative patches usually involves searching through tens of thousands of patches, which is a bottleneck in recent works. In this sense, if appropriately used, association rule mining can be an appealing solution for handling "Big Data" in mid-level visual element discovery.

We utilize Apriori algorithm to find the patterns between feature dimensions and then we determine the patches that include this rule. Hence, these discovered patches are used in model learning for action recognition. By examining the firing test patches, we show to enhancements over baseline results.

2 RELATED WORKS

This work is related to three topics: Action recognition from still images, mid-level visual element discovery and association rule mining for vision. In the following we explore the recent works for each topic separately.

Action recognition from still images: Recognizing actions from still images continues to be a problem since it is essential to detect specific action patterns without the motion information. Firstly, Ikizler et al. [9] propose to find rectangles over the pose silhouettes and then represent the motion with Histogram of Oriented Rectangles. Furthermore, Thurau et al. [20] extend a Histogram of Oriented Gradient (HOG) based descriptor to recognize the pose.

More recently, Sharma et al. [15] propose to use discriminative parts of action images similar to our work. They imply that HOG

Figure 2: Discriminative and representative action patches of training set which are obtained from association rules.

is not enough to recognize action since it is only able to determine shape and it is useful for pose estimation. However, we also need to detect the object interaction to estimate the motion. In addition, localization and attention based works are proposed [24] with the development of deep networks.

Mid-level element discovery: Discovering discriminative elements from images is found to be useful especially for still image action recognition since it is preferable to classify motion itself rather than the entire image. Singh et al. [18] show the achievements of discriminative pathches for image classification and this method used in many works [2–4, 14]. These methods are based on generally splitting images into patches and then determining the ones with highest scores in model learning. More recently, discriminative methods are applied to deep networks for better classification results [1, 8].

Association rule mining for vision: Pattern mining algorithms are proposed for data mining techniques. However, they are also used in image classification [1, 6, 23]. The main advantage of pattern mining lies its ability to process massive volumes of data, which is particularly appealing in this era of information explosion. Li et al. [13] propose to use association rule mining in order to find discriminative patches of images. After splitting into thousands patches, images are represented with only the representative parts which include the discovered pattern.

3 APPROACH

In this section, the utilized method, which is based on [13], will be explored in detail. First, we explain how to represent image patches and how to extract them from images. Then, we continue with transaction creation for rule mining. Finally, we give the details of finding association rules between feature dimensions.

3.1 Patch Representation

First of all, we resize all images into dimension of $256x256$. Then, we extract $128x128$ patches with the stride of 32. Hence, for each image we get 16 patches. After that, we extract GoogLeNet [19] "ave" activations from each patch which are in the dimension of $1x1024$. For a whole image representation, while calculating the baseline, basically max pooling is done on related patch activations.

3.2 Transaction Creation

In order to create transaction database, each patch is considered as one transaction and each activation index determined as an item. Consequently, we have 1024 items in total. It is implied in [13] that for an image patch, the discriminative information within its CNN activation is mostly embedded in the dimension indices of the k largest magnitudes. Therefore, for transaction creation, we take dimension indices which have "k" largest magnitudes. This "k" is determined as 20 which is proven to be useful in [13]. Hence, in one transaction there are 20 items at most.

Since we want to find discriminative pattern rules, we apply the same process that is suggested in [13]. While calculating patterns, we consider two different sets, positive and negative so that the association rule will have the ability to discriminate between target category and natural world. Here, we calculate patterns for each action category separately.

Therefore, if the target category is the "applauding" action class, natural world transactions will be from other action classes. Then, we combine them in order to find association rules with Apriori algorithm. This is done only for training set and we get $3200x20$ transaction matrix where 1600 transactions come from target and others are selected randomly from natural world. We add class label item at the end of each transaction in order to imply which ones belong to target category (item:1025) and natural world (item:1026).

This approach provides a discriminization for the rule that will be discovered. The reason is that, the last items (1025, 1026) that determine the category of transactions, avoid some frequent items to be discovered as a pattern if they are also frequent in other categories. Therefore, this situation provides us to have discriminative and representative patterns for that action category.

In more detail, values of $support(P)$ and $confidence(P)$ ensure that pattern P is found frequently in the target category, akin to the representativeness requirement [13]. A high value of $confidence(P)$ will also ensure that pattern P is more likely to be found in the target category rather than the natural world, reflecting the discriminativeness requirement.

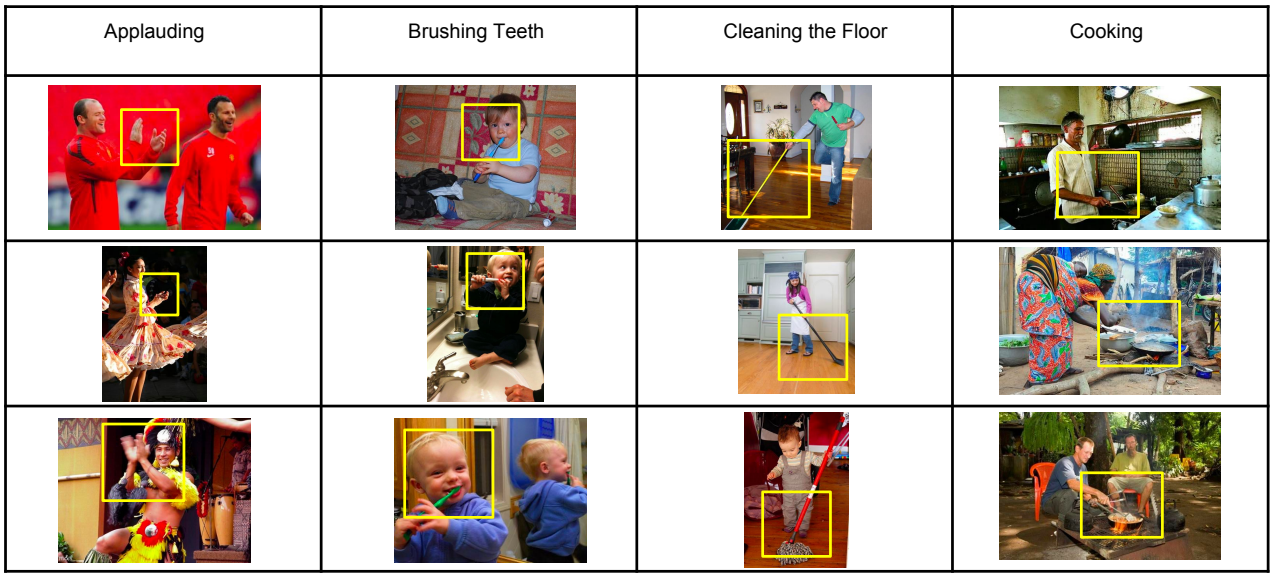

Figure 3: Predicted test patches for related actions. Illustrated by using the whole image.

3.3 Association Rule Mining

We feed created transactions to Apriori algorithm for each category and extract related association rules between indices of CNN activations. Then, by taking the union of discovered rules, we determine the feature dimensions which are discriminative and representative for that category. By utilizing inverted index approach, we find transactions that contain discovered pattern. Following those transactions, we obtain representative image patches. As a result, a mid-level visual element V contains the image patches sharing the same pattern P.

4 EXPERIMENTS

4.1 Dataset

In this work, we use a subset of Stanford 40 Action Dataset [22] which includes images of humans performing 40 different actions. There are 9532 images in total with 180-300 images per action class. We experiment on only first 10 categories. We create train and test sets by selecting randomly 100 images for train and the rest for test.

4.2 Implementation

We resize all images into dimension of $256x256$. Then, we extract $128x128$ patches with the stride of 32. In order to extract CNN activations, we use Caffe [10] and the model of GoogLeNet by using "ave" outputs. For training classifiers, we use LIBLINEAR [5] toolbox and apply multi-class classification with one-vs-all approach. The parameter C is selected as 100 for all experiments.

4.3 Parameter Tuning

In the Apriori algorithm, we determine the $minSup$ and $minConf$ parameters according to experiments. Before training the classifiers for parameter tuning, we max pool discovered image patch activations of an image both for train and test sets. Then, we obtain prediction results. In Table 1, we give the results for different $minSup$ values by taking $minConf = 0.7$. Moreover, in Table 2, we give experimental results for varying $minConf$ values by taking $minSup = 0.3$.

Table 1: Effect of different minimum support values on classification results.

Min Support Value	Classification Accuracy
0.2	57.26%
0.3	**84.81%**
0.4	75.79

Table 2: Effect of different minimum confidence values on classification results.

Min Confidence Value	Classification Accuracy
0.5	76.45%
0.7	**84.81%**
0.9	83.85

We can see that we get the best accuracies when $minSup = 0.3$ and $minConf = 0.7$. It is observed that when we choose $minSup = 0.2$, discovered image patch number is less than when $minSup = 0.3$. One may be think that, when $minSup = 0.2$, we get more discriminative patches since many of them are eliminated. However, according to results, the optimal patch number is obtained when $minSup = 0.3$. Thus, we select $minSup = 0.3$ and $minConf = 0.7$ for the rest of the experiments.

Table 3: Classification results with max pooling and patches directly.

	Baseline	Mid-level
Max pooling	75.4%	**84.81%**
Patch based	55.99%	**70.12%**

4.4 Results

We give both qualitative and quantitative results for different experiments. First of all, we take max pool of image patch activations for each image in order to obtain an image level representation. Then, we calculate the baseline by applying multi-class SVM with one-vs-all approach. Similarly, after discovering discriminative patches and max pooling only these discovered ones, we calculate accuracy for the same test set.

The results are given in Table 3. We can see that, by using only the discriminative and representative patches instead of the entire image itself, we get 9% higher accuracy while predicting test images. In addition, we give the discovered mid-level visual elements for specific action categories in Figure 2. It can be seen that all of the selected patches are discriminative and representative for their related action category and their image representation, which is calculated by max pooling, is able to predict test images more accurately.

Secondly, we use extracted image patches directly without computing the max pool of them and we train classifiers for these patches. Similarly, after getting discriminative patches, we again use them to train more robust classifiers. From the Table 3, results show that by learning models with patch level representations, we are able to increase the total prediction score about 15%. Furthermore, we give the firing test patches which are predicted with patch-level models in Figure 3. We can see that related actions are detected accurately by using the given detectors from training set which are formed from discriminative and selective action patches (in Figure 2).

Moreover, we give class based accuracy results in Table 4. We get higher accuracies for all action categories by using mid-level representations of images. Especially, for some of the challenging actions such as "brushing teeth", "drinking", "cutting vegetables", the improvement is significantly high which shows the importance of discriminative and representative patch selection.

5 CONCLUSION

In this work, we investigated the effect of mid-level representative definitions of still images for action recognition. By utilizing the proposed method, after extracting patches, we find some patterns between indices of activation's dimensions with association rule mining which is able to find the most discriminative and representative rules for the action category of interest. With the conducted experiments, we show the importance of using discriminative parts instead of the whole image representation for action recognition from still images. The results are good, both qualitatively and quantitatively and the resulting visual elements are found to capture strong semantic meanings, and are transferable to several important vision tasks. The transferability to more tasks can be studied in the future.

Table 4: Class based classification accuracy comparisons for max pooled representations.

	Baseline	Mid-level
Applauding	72.82%	82.60%
Blowing Bubbles	76.10%	83.64%
Brushing Teeth	69%	85%
Cleaning the Floor	84.82%	94.64%
Climbing	86.15%	93.33%
Cooking	72.34%	75%
Cutting Trees	90.29%	93.20%
Cutting Vegetables	68.53%	83.14%
Drinking	54.48%	72.43%
Feeding a Horse	92.22%	94.44%
Overall	75.4%	84.81%

REFERENCES

[1] Ali Diba, Ali Mohammad Pazandeh, Hamed Pirsiavash, and Luc Van Gool. 2016. Deepcamp: Deep convolutional action & attribute mid-level patterns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3557–3565.

[2] Carl Doersch, Abhinav Gupta, and Alexei A Efros. 2013. Mid-level visual element discovery as discriminative mode seeking. In *Advances in neural information processing systems*. 494–502.

[3] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. 2012. What makes paris look like paris? *ACM Transactions on Graphics* 31, 4 (2012).

[4] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition.. In *Icml*, Vol. 32. 647–655.

[5] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research* 9 (2008), 1871–1874.

[6] Basura Fernando, Elisa Fromont, and Tinne Tuytelaars. 2012. Effective use of frequent itemset mining for image classification. In *European conference on computer vision*. Springer, 214–227.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[8] Chen Huang, Chen Change Loy, and Xiaoou Tang. 2016. Unsupervised learning of discriminative attributes and visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5175–5184.

[9] Nazli Ikizler, R Gokberk Cinbis, Selen Pehlivan, and Pinar Duygulu. 2008. Recognizing actions from still images. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 1–4.

[10] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093* (2014).

[11] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725–1732.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[13] Yao Li, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. 2015. Mid-level deep pattern mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 971–980.

[14] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1717–1724.

[15] Gaurav Sharma, Frédéric Jurie, and Cordelia Schmid. 2013. Expanded parts model for human attribute and action recognition in still images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 652–659.

[16] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*. 568–576.
[17] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[18] Saurabh Singh, Abhinav Gupta, and Alexei Efros. 2012. Unsupervised discovery of mid-level discriminative patches. *Computer Vision–ECCV 2012* (2012), 73–86.
[19] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.
[20] Christian Thurau and Václav Hlavác. 2008. Pose primitive based human action recognition in videos or still images. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 1–8.
[21] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 4489–4497.
[22] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. 2011. Human action recognition by learning bases of action attributes and parts. In *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 1331–1338.
[23] Junsong Yuan, Ming Yang, and Ying Wu. 2011. Mining discriminative co-occurrence patterns for visual recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2777–2784.
[24] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2921–2929.