Pedestrian Detection and Scene Recognition With Co-occurrence Features and Ada Boosting Classifier

0029033989
School of Electrical and Computer Engineering
Purdue University
xxx.purdue.edu

Abstract

This paper is divided into two parts. The first part is the critiques of three papers, including "Learning Query and Images Similarities with Ranking Canonical Correlation Analysis" by Yao et al, "Mining And-Or Graphs for Graph Matching and Object Discovery" by Zhang et al, and "Object Detection Using Generalization and Efficiency Balanced Co-occurrence Features" by Ren et al. All three of them are focusing on computer vision region, from graph matching, image searching to object detecting.

The second part of this paper is the description of my implementation on a small portion of the paper Ren et al [3] accomplished in their paper. I construct gray-level co-occurrence features. Extending from the paper, I evaluate them not only on human detection but also on scene recognition. By introducing CoSIFT and CoSURF features, I compare their performance with Bags of SIFT features and GIST descriptor on scene recognition.

1. Critique

Learning Query and Images Similarities with Ranking Canonical Correlation Analysis

Yao et al [1] broke through the limits of image search. Instead of relying on image ranker learning which requires human labeling; and feature-based vector model which relies on the surrounding texts, they proposed Ranking Canonical Correlation Analysis (RCCA). Initially, Standard Canonical Correlation Analysis (CCA) learns the share common subspace between query and image spaces by maximizing the correlations. In addition, RCCA is proposed to simultaneously learns a bilinear similarity function and adjusts it to preserve the preference relation in the click-through data. Consequently, with the finalized subspace, similarities between query and image can be computed. Therefore, image search is being further optimized.

Not only does RCCA optimized the image search, it also reduces the training complexity. It is linear to the number of triplets, which would not dramatically slow down the process with updates of new triplets. As for online search, RCCA is fast enough on a regular PC, and can achieve instant response on searching.

The result of experiments on keyword-based image search shows that RCCA outperforms CCA, CCL, KPCA_CCA, KCCA, PSI, and PA. As for query-by-question image search, its result exhibits better performance than CCA, CCL, PCA. RCCA preserves the correlations between two views. The outcomes reveal that the similarities between image mappings could better reflect their semantic relations by maximizing the query and image correlations in CCA. And with the preserving of the corporate preference relations in click-through data, RCCA is capable of learning the subspace and similarity function at the same time and separating the images with different semantics, which leads to a better performance.

In the era of big data, Yao et al successfully utilize the large scale click based image dataset to learn the preference of CCA upon query and image set. However, the limitation of Yao et al's attempt is that the subspace learned in the first stage is mostly depending on query-image and image-image similarity. If query-query similarities are also implemented and learnt, the result of query-by-question image search might significantly improve and might also reduce searching time with the implementation of RCCA.

Mining And-Or Graphs for Graph Matching and Object Discovery

Zhang et al [2] broaden the scope of application of graph matching to computer vision. Instead of originally developed discovering frequent entity relations from data. They cut in with a different angle. With a set of unannotated At-

tributed Relational Graph(ARG) containing different unary attributes, they define a hierarchical And-Or Graph(AoG) to model the common subgraphs in the set. The top AND node is composed by a set of OR nodes and each OR node represent alternative local patterns and has several alternate terminal nodes.

An AoG model can represent distinct visual data. Zhang et al aim to mine the AoG from a number of ARGs. During the mining process, new OR nodes and terminal nodes are discovered; redundant nodes are deleted; attributes and matching parameters are trained and updated. While learning the model, the template is modified into the target AoG with the maximal number of OR nodes. The mined AoG can therefore used for graph-matching upon previously unseen ARGs and can therefore be used for object inference.

Zhang et al extend the concepts from both the fields of graph mining and unsupervised learning. The contribution of this technique is that can be used upon various types of visual data including 2D and 3D graphs; object occlusions are considered in the method, intra-category variations can be automatically mined; it is efficient to directly discover the AoG model without node numeration; and does not require supervised learning.

In comparison to 13 competing methods include image/ graph matching approaches, unsupervised learning for graph matching, object discovery, and segmentation methods, their method shows significant performance to the competing methods. Their method exhibits lower error matching rate and have the advantage of the ability to change pattern size which leads to a better matching performance.

Since this work can apply to different optical data, it can further specify and be optimized upon different attempt instead of a general solution. For instance, while encountering video mining, it can also be combined with tracking technique to guide it which can be made more efficient.

Object Detection Using Generalization and Efficiency Balanced Co-occurrence Features

Ren et al [3] propose a object detector based on three kinds of local co-occurrence features constructed by Histogram of Oriented Gradients (HOG) [7], Haar [8], and Local Binary Patterns (LBP) [9], which are able to capture complicated object characteristic and can be computed efficiently. Instead of using more powerful features like the higher-order gradient features, heterogeneous features, and feature fusion, the three features Ren et al chose to reduce the heavy computational cost which is cost by the dense feature vector.

The features correspond with a local image region to build weak classifiers in the boosted detector. Where the detector is trained by a newly proposed Generalization and Efficiency Balanced (GEB) framework which is based on RealAdaBoost [4][5][13] algorithm. It is utilized to evaluate the robustness, accuracy and efficiency of different weak classifiers at the same time. Therefore the detection based on GEB not only achieves high accuracy, but also has considerably effect and good generalization power.

For pedestrian detection, with the GEB framework, features Ren et al proposed work better compared to traditional features and their combination. It can balance the discriminative ability and the generalization power of the detector, which then contributes to the accuracy of the resulting classifier. While comparing with the state-of-the-art pedestrian detection results, the detectors with a single co-occurrence feature achieve lower accuracy. The reason is that the feature is extracted on a single scale, so the discriminative ability might be lower compared to the evolution of channel features. But this can be compensated by the combination of multiple co-occurrence features selected by GEB. The RealAdaBoost can eliminate redundant information to get meaningful patterns and improve efficiency.

The contribution of this technique is the low computation cost with a high accuracy that is competitive with other state-of-the-art methods while boosting tasks require appropriate features to effectively describe the object characteristic and most of the co-occurrence information extraction methods are time consuming.

While the features Ren et al select are specifically used upon object and pedestrian detection. It might be further optimized on different attempt. For example, they can evaluate the framework on scene detection.

2. Implementation

The paper Ren et al [3] accomplished in their paper focus on object and pedestrian detection, as for this paper, I construct gray-level co-occurrence features and evaluate the method on pedestrian and scene recognition tasks to compare its method upon different kind of detection and features.

I choose HOG and LBP for implementation from the paper and exclude Haar features since the extraction of Haar features requires too many images in one category and takes a great amount of time to achieve feature set with well performance.

Aside from the features Ren et al choose, I introduce two more feature descriptors including Scale-Invariant Feature Transform (SIFT) [9] and Speeded Up Robust Features (SURF) [10] for comparison of the recognition accuracy among different features and different classification types. The main reason I introduce SIFT and SURF features in this paper is for researching on the effects of co-occurrence features in scene recognition. And I also tried to compare the results of scene recognition between the usage

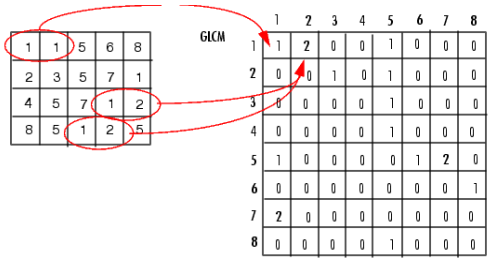

Figure 1. Example of calculating the values in the GLCM of a 4-by-5 image.

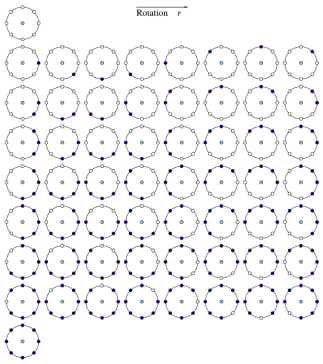

Figure 2. 58 uniform patterns of $LBP_{8,1}$ feature. Each row corresponds to a cluster in the extraction of CoLBP features.

of the bag of words model and co-occurrence model upon features.

2.1 Gray-level Co-occurrence patterns

To produce gray-level co-occurrence (GLCM) features is by calculating how often a pixel with gray-level (gray scale intensity) value i occurs horizontally adjacent to a pixel with the value j. Each element (i,j) in the return matrix specifies the number of times that the pixel with value i occurred horizontally adjacent to a pixel with value j. For the pixel pair distance offsets, I assigned it to be 5. The matrix is symmetric which means the value in (i,j) would be the same as the value in (j,i). I also normalize the outcome by dividing the total number of accumulated co-occurrences for the given offset, therefore the elements of the resulting matrix sum to 1.

For all of the production of the following features, both intensity and gradient domain (gradient-x and gradient-y) of an image are used.

CoLBP

The traditional Local Binary Patterns (LBP) is designed for texture classification. The LBP feature vector is created by comparing each pixel to each of its 8 neighbors and when the pixel's value is greater than the neighbor's value, write "0". Otherwise, write "1". This makes an 8-digits binary number for each pixel.

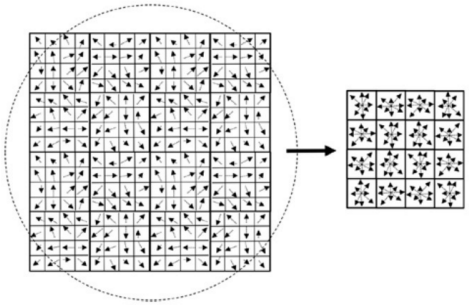

Figure 3. An example of the process of generating the SIFT feature, the left is the image gradient, with it one can further generate the key point descriptor (right image).

Uniform LBP is a subset of LBP. It can be interpreted as corners of edges. Figure 2 shows all uniform patterns of $LBP_{8,1}$. In this experiment, all non-uniform patterns are merged into another pattern. The LBP extraction is applied on both the intensity and gradient domain.

When proposing GLCM to LBP features, I merge the 58 uniform $LBP_{8,1}$ patterns to 8 clusters based on the number of '1' values same as Ren et al did. All the non-uniform patterns construct another cluster. As a result, the CoLBP histogram consists of 9 x 9 dimensions.

$$LBP_{d,r} = \sum_{i=1}^{d} sign(I_i - I_e) \times 2^{i-1}, \qquad (1)$$

CoHOG

Histogram of Oriented Gradients (HOG) is a feature descriptor first described as a use of pedestrian detection and now mostly used in a sliding window fashion for the propose of object detection.

HOG counts the occurrences of gradient orientation in localized portions of an image. It breaks the image into blocks and generates histogram based on the gradient orientation. This method computed on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for improved accuracy.

I applied HOG on both the intensity and gradient domain of the images. And while applying GLCM, the features are quantized to 8 bins, which gives 8 x 8 levels in the co-occurrence histogram.

CoSIFT

Scale-invariant feature transform (or SIFT) is an algorithm to detect and describe local features in images. It extracts distinctive image features from scale-invariant, which extract key points and compute its descriptor. SIFT is typically used for matching local regions in two images for purposes of alignment, reconstruction, and structure from mo-

Figure 4. Example scenes from each category in the 15 scene dataset. Figure from Lazebnik et al. 2006.

tion, where it also has been used for recognition as well (e.g. in a Bag-of-features fashion).

Extraction of SIFT key point descriptor can be divided into three steps including scale-space extrema detection, key point localization and orientation assignment.

After key points with the same location and scale, but different directions are built. A 16 x 16 neighborhood around the key point is taken. It is divided into 16 sub-blocks. For each sub-block, 8 bins orientation histogram is created. So a total of 128 bin values are available. It is represented as a vector to form key point descriptor. In addition to this, several measures are taken to achieve robustness against illumination changes, rotation etc.

For the co-occurrence of the SIFT features, I divided the SIFT features into 8 bins so that there are 8 x 8 elements in the CoSIFT features.

CoSURF

Comparing to SIFT features, Speeded Up Robust Features (SURF) is considered to be a speeded-up version of SIFT as its name suggests. Experiments show that SURF is three times faster than SIFT while its performance is comparable to SIFT. SURF is achieved by relying on integral images and existing detectors and is good for handling blurry image with rotation.

SURF adds a lot of features to increase the speed in every step of SIFT feature extraction. For instance, SURF approximates LoG with different size box filter; uses wavelet response in horizontal and vertical direction for a neighborhood of size 6s; and the use of sign of Laplacian for underlying interest point. All of these attempts to speed up the process of extracting features.

For the co-occurrence of the SURF features, I also divided them into 8 bins so that there are 8 x 8 elements in the CoSURF features.

2.2 AdaBoost classifier

AdaBoost is short for Adaptive Boosting. It is an estimator that begins by fitting a classifier on the original dataset and then fits more copies of the classifier on the same dataset. All the classifiers are perceived as weak classifiers. All of

	Pedestrian		Scenes	
	AdaBoost	**SVM**	**AdaBoost**	**SVM**
CoLBP	97.73%	92.49%	27.53%	44.60%
CoHOG	77.34%	60.53%	24.75%	24.72%
CoSIFT	81.23%	61.80%	16.68%	29.42%
CoSURF	72.67%	61.58	17.07%	18.33%
CoLBP, CoHOG	97.62%	93.27%	28.31%	48.60%
CoSIFT, CoSURF	82.67%	69.91%	18.36%	35.89%
CoLBP, CoHOG, CoSIFT, CoSURF	97.58%	94.71%	21.78%	55.80%

Table 1. Showing the prediction results of different co-occurrence and combinational feature pairs with AdaBoost/SVM classifier among pedestrian detection and scene recognition.

which is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost is adaptive since that the weights of weak classifiers' inadequate classified instances are adjusted so that the subsequent classifiers can be utilized in more difficult cases.

Individual learners can be weak, but as long as the performance of each one is slightly better than random guessing, the final model can converge to a strong learner.

In my experiment, I choose decision trees as the weak learners since decision trees are non-linear and are generally a very good fit for boosting.

Experimental Results

Datasets

For pedestrian recognition, INRIA dataset [5] is utilized. It contains 1,774 human annotations (3,548 with reflections) and 1,671 person free images. I use 64 x 128 pedestrians with pixel pair distance offsets equal to five and number of levels equal to 9 x 9 for CoLBP features and 8 x 8 for rest of the features. There are 2416 positive, and 1211 negative training images. As for testing images, 1126 of them are positive images and 450 of them are negative images.

For scene recognition, 15 scene database introduced by Lazebnik et al[12] is used. For training, there are 100 images from each category(i.e. 1500 training images total).

Comparison with different co-occurrence natures and feature combination

I categorize CoLBP/CoHOG as one feature class and CoSIFT/CoSURF as another since LBP/HOG are preferred features for human/object classification and SIFT/SURF are mostly used in corner detection and scene recognition.

Pedestrian Recognition

In human recognition, comparing between co-occurrence features' prediction result, I found that since CoLBP's result is better than that of LBP's (97.73% for CoLBP and 91.27% for LBP only). On the other hand, for HOG features, the result of the co-occurrence feature is not as good

as that of the original feature (77.34% for CoHOG and 94.49% for HOG feature). From this one can conclude that while co-occurrence features might reduce calculation complexity. It does not necessarily improve the recognition accuracy.

The results of the co-occurrence features are still pretty impressive even without the combination of features. With CoLBP and Adaboost classifier, the accuracy reaches 97.73%, even higher than the prediction using its combination with other features, which in this case do not match the statement of Ren et al's paper. The prediction result of the combination of CoLBP and CoHOG is 97.62% and the combination result of all four features is 97.58%. The reduction of the resulting accuracy rate might occur due to the curse of dimensionality.

Scene Recognition

For scene recognition, the results of co-occurrence features are not so satisfying comparing the results with pedestrian recognition. Since the feature of the same scene changes a lot while perspective changes.

In this experiment, for a single feature, CoLBP gives the highest accuracy rate (27.53%). CoSIFT/CoSURF features did not perform better than CoLBP/CoHOG features as expected (16.68% for CoSIFT and 17.07% for CoSURF). However, this time, the combination of CoLBP and CoHOG gives a better result than that of a single feature (28.31%). But, the performance in general is not satisfying.

Comparison between Adaboost classifier and SVM classifier

Since the results of the combination of co-occurrence features and AdaBoost classifier on scene recognition is not so satisfying. I tried to implement co-occurrence features with SVM classifier to see if the performance can be improved.

Support vector machines (SVM) [11] are supervised learning models that analyze data used for classification, regression, and outliers detection. A SVM model is a representation of the examples as points in space, mapped so that the examples of different categories are separated by a clear gap that is as wide as possible.

AdaBoost and SVM classifier are both non-parametric models; can both capture complex decision boundaries while avoiding over-fitting in many cases; and both produce a model that is not really interpretable.

The advantages of SVM are the effectiveness in high dimensional spaces and SVM are quicker to train and evaluate than AdaBoost. Nonetheless, AdaBoost almost always gives a better accuracy upon predicting accuracy.

From table 1, one can tell that for pedestrian recognition, AdaBoost gives a better result than that of SVM's. Which is predictable, since boosting algorithm usually gives the best prediction results among other classifiers.

However, for scene recognition, SVM gives the prediction accuracy almost twice as good as that of AdaBoost's. It might results from the characteristic that AdaBoost is

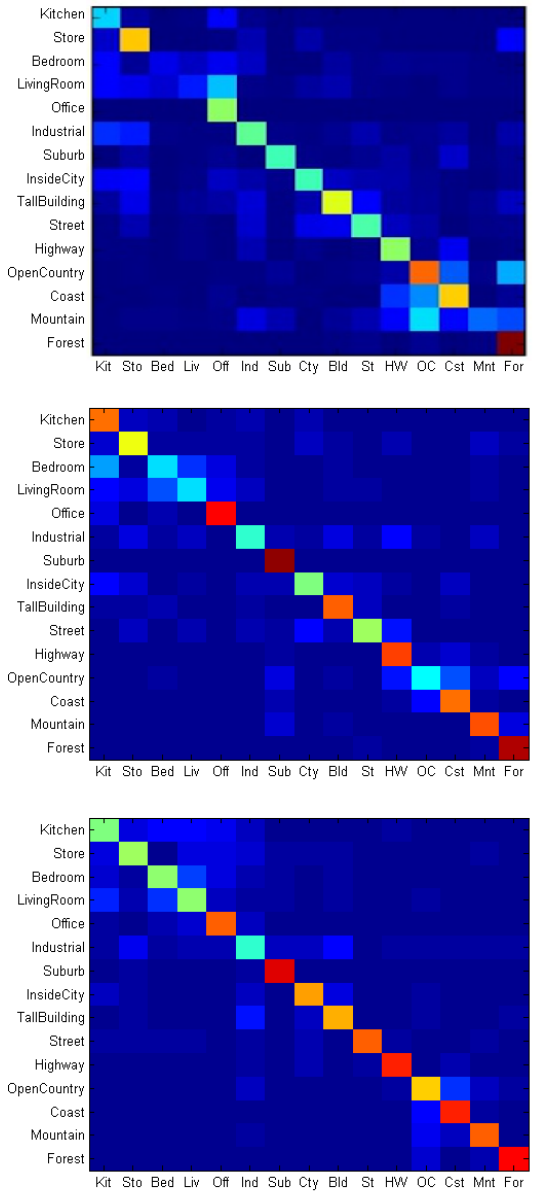

Figure 5. Results of scene recognition with SVM classifier among (a) CoSIFT (first image)(a) Bags of SIFT (second image) and (b) GIST descriptor (third image).

sensitive to noisy data and outliers. And the ranking among the results of different features also changed. Now the combination of four co-occurrence features beats the performance of that of a single feature.

Comparison on scene recognition with other attempts

Since the scene classification utilizing co-occurrence features and SVM classifier still does not provide a satisfying result. I try to implement bag of features and GIST descriptor to improve the classification accuracy. Both of which are famous in scene classification.

Bag of words models [12][14] are a popular technique for image classification inspired by models used in natural

language processing. The model ignores or downplays word arrangement (spatial information in the image) and classifies based on a histogram of the frequency of visual words. The visual word "vocabulary" is established by clustering a large corpus of local features.

To extract bags of SIFT, one has to first build the vocabularies starting by sampling the SIFT descriptors from every image. Then cluster all the descriptors with K-means then return the cluster centers as vocabularies. After building the vocabularies, implementing the KD-tree algorithm can get the bags of SIFT.

On the other hand, GIST descriptor was developed to provide a holistic descriptor that provides a simpler representation. It is usually computed over the entire image as a global image descriptor, unlike SIFT is a localized image patch descriptor. GIST summarizes the gradient information (scales and orientations) for different parts of an image, which gives a rough description of the scene.

Given an input image, a GIST descriptor is computed by first convolve the image with 32 Gabor filters at 4 scales, 8 orientations, producing 32 feature maps of the same size of the input image. In addition, divide each feature map into 16 regions (by a 4 x 4 grid), and then average the feature values within each region. Last, concatenate the 16 averaged values of all 32 feature maps, resulting in a 512 (16 x 32) GIST descriptor.

In this comparison, I choose linear SVM as classifier instead of AdaBoost classifier based on the discovery of the better performance in scene recognition upon SVM classifier. I implement bags of SIFT with linear SVM classifier and get accomplished the accuracy of 64.9%. GIST descriptor with linear SVM classifier gives an even better performance of 68.5%, while CoSIFT with linear SVM classifier only accomplished the accuracy of 29.4%, and the combination feature of all four co-occurrence features gives the accuracy of 55.8%. The classification results of the above mentioned three features type are shown in figure 5. From the results, one can conclude that, the co-occurrence features are not mature enough for the implementation of scene recognition. Since there are more variables in scenes than in a human's posture, where in the contrary, the postures of a person are mostly standing.

Conclusion

In this paper, there are first critiques of three papers I've selected including "Learning Query and Images Similarities with Ranking Canonical Correlation Analysis" by Yao et al, "Mining And-Or Graphs for Graph Matching and Object Discovery" by Zhang et al, and "Object Detection Using Generalization and Efficiency Balanced Co-occurrence Features" by Ren et al. Later on, I implemented some concept from Ren et al's paper, and constructed features including CoLBP, CoHOG, CoSIFT, and CoSURF. The results show that utilizing the combination of Co-occurrence features with AdaBoost classifier in pedestrian detection is rather effective. On the other hand, for scene recognition, comparing to other descriptor such as bags of features and gist descriptor, co-occurrence features are not the preferable features even with the combination of AdaBoost classifier. However, with the combination of SVM classifier, it does give a relatively good result.

For future experiments, among pedestrian detection and scene recognition, the prediction outcome might be further optimized with the GEM framework that Ren et al introduce.

References

[1] Y. Ting, M. Tao, and N, Chong-Wah. Learning Query and Images Similarities with Ranking Canonical Correlation Analysis. In ICCV, 2015

[2] Z. Quanshi, N. W. Ying and Z. Song-Chun. Mining And-Or Graphs for Graph Matching and Object Discovery. In ICCV, 2015

[3] R. Haoyu and L. Ze-Nian. Object Detection Using Generalization and Efficiency Balanced Co-occurrence Features. In ICCV, 2015

[4] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In CVPR, 2001.

[5] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. Annals of statistics, 1998.

[6] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In CVPR, 2005.

[7] S. Zhang, C. Bauckhage, and A. Cremers. Informed Haar-like Features Improve Pedestrian Detection. In CVPR, 2013.

[8] T. Ojala, M. Pietikäinen and T. Mäenpää. Multiresolution Gray-scale and Rotation Invariant Rexture Classification with Local Binary Patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7) : 971-987, 2002.

[9] D. Lowe. Distinctive Image Features from Scale Invariant Keypoints. International Journal of Computer Vision, 60, 2 (2004), pp. 91-110

[10] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346-359, 2008

[11] A. J. Smola, B. Schölkopf, A Tutorial on Support Vector Regression, Statistics and Computing archive Volume 14 Issue 3, August 2004, p. 199-222.

[12] S. Lazebnik, C. Schmid, and J. Ponce1. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In CVPR, 2006

[13] P. Viola, and M. J. Jones, Robust Real-Time Face Detection. International Journal of Computer Vision 57(2), 137–154, 2004

[14] K. Chatfield, V. Lempitsky, A. Vedaldi, A. Zisserman. The Devil is In the Details: An Evaluation of Recent Feature Encoding Methods. In BMVC, 2011