Automatic Discrimination of Color Retinal Images using the Bag of Words Approach

Ibrahim Sadek, Désiré Sidibé, F Meriaudeau

Automatic Discrimination of Color Retinal Images using the Bag of Words Approach

I. Sadek [a], D. Sidibé [a], F. Meriaudeau [a]

[a] University of Burgundy, Le2i Laboratory, 12 rue de la fonderie, 71200 Le Creusot, France.

ABSTRACT

Diabetic retinopathy (DR) and age related macular degeneration (ARMD) are among the major causes of visual impairment all over the world. DR is mainly characterized by small red spots, namely microaneurysms and bright lesions, specifically exudates. However, ARMD is mainly identified by tiny yellow or white deposits called drusen. Since exudates might be the only visible signs of the early diabetic retinopathy, there is an increase demand for automatic diagnosis of retinopathy. Exudates and drusen may share similar appearances; as a result discriminating between them plays a key role in improving screening performance. In this research, we investigative the role of bag of words approach in the automatic diagnosis of retinopathy diabetes. Initially, the color retinal images are preprocessed in order to reduce the intra and inter patient variability. Subsequently, SURF (Speeded up Robust Features), HOG (Histogram of Oriented Gradients), and LBP (Local Binary Patterns) descriptors are extracted. We proposed to use single-based and multiple-based methods to construct the visual dictionary by combining the histogram of word occurrences from each dictionary and building a single histogram. Finally, this histogram representation is fed into a support vector machine with linear kernel for classification. The introduced approach is evaluated for automatic diagnosis of normal and abnormal color retinal images with bright lesions such as drusen and exudates. This approach has been implemented on 430 color retinal images, including six publicly available datasets, in addition to one local dataset. The mean accuracies achieved are 97.2% and 99.77% for single-based and multiple-based dictionaries respectively.

Keywords: Diabetic retinopathy, Bag of words, Exudates, Drusen, SVM

1. INTRODUCTION

According to the world health organization (WHO) diabetes mellitus (DM) is a lifelong disorder which takes place either when the pancreas doesn't produce sufficient insulin (type 1 diabetes) or when the body cannot effectively benefit the insulin it produces (type 2 diabetes). Insulin is a hormone developed by beta cells in the pancreas that regulates the level of blood sugar. Hyperglycemia, or increased blood sugar level causes serious damage to body's system, including diabetic retinopathy. The most important reasons of diabetes are increasing age, overweight, and sedentary lifestyle. Studies show that almost all patients with type 1 diabetes and more than 60% of patients with type 2 diabetes evolve retinopathy during the first two decades of disease.[1] The prevalence of diabetes is estimated to increase from 2.8% to 4.4% in the time span of $2000 - 2030$. The total number of people with diabetes is projected to increase from 171 million in 2000 to 360 million in 2030.[2] Diabetic patients can prevent severe visual loss by attending regular diabetic eye screening programs and receiving optimal treatments.[3] Diabetic retinopathy (DR) and age related macular degeneration (ARMD) are among the leading causes of visual impairment worldwide. DR occurs most frequently in adult aged $(20 - 74)$ years, and it is characterized by the presence of red lesions (microaneurysms) and bright lesions (exudates) which appear as small white or yellowish white deposits with sharp margins and variable shapes located in the outer layer of the retina, their detection is essential for diabetic retinopathy screening systems. ARMD usually affects people over 50 years of age. It is caused by damage to the macula, the small sensitive area of the retina that gives central vision (seeing fine details and colors), and categorized by drusen, tiny yellow or white deposits in a retina layer

Further author information: (Send correspondence to Ibrahim S.)
Ibrahim S.: E-mail: ibrahimsadek87@gmail.com
Désiré S.: E-mail: dro-desire.sidibe@u-bourgogne.fr
Fabrice M.: E-mail: fabrice.meriaudeau@u-bourgogne.fr

called Bruch's membrane. The severity of ARMD can be categorized into three classes: early, intermediate, and advanced. In some patients, bright lesions such as retinal exudates can be the only manifestations of early diabetic retinopathy. Thus, computer aided detection (CAD) systems have been proposed in order to detect exudates. However, these bright lesions must be identified from drusen because they share common characteristics.[4] This represents a challenge for readers or CAD based screening systems designed for DR diagnosis. Consequently, developing a CAD system for reading and analyzing retinal images decreases observational unintentional failure and the false negative rates of ophthalmologists interpreting these images. The aim of this research work is to design a system that will be able to identify normal, drusen, and exudates in color retinal images using the bag of words approach (BOW), because there are a few approaches in the literature designed for this purpose.

2. RELATED WORK

In literature, a wide variety of CAD systems to detect retinal features and lesions involve three main steps. The first step is the preprocessing in order to compensate for great variability between and within retinal images, where the green channel is considered the most preferable choice because it provides a maximum contrast between different retinal lesions and structures. The second step is to extract candidate lesions, in some approaches feature selection may be performed in order to remove redundant features. The last step is to classify candidate lesions into normal or abnormal. Grinsven et al.[5] have proposed to use the BOW approach to retrieve and classify images with bright lesions, namely drusen and exudates. However, this approach needs a prior knowledge about the location of the optic disk and macula. Pires et al.[6] have also used the BOW in order to identify images with bright lesions such as hard exudates, cotton wool spots, and drusen, in addition to images with red lesions like hemorrhages and microaneurysms. Nevertheless, the proposed strategy requires manual annotation of unhealthy regions. Deepak et al.[7] have developed a strategy for bright lesion detection using a visual saliency based framework. This method relies on accurate detection of drusen and exudates that is considered as a challenge to any CAD system. We present a novel approach by adapting single-based and multiple-based dictionaries for identifying normal images and abnormal images with bright lesions.

3. METHODOLOGY

Basically, the proposed method depends on the bag of words (BOW) approach to automatically discriminate between normal, drusen, and exudates in color retinal fundus images. In this approach, the images are pre-processed. Subsequently, SURF as well as HOG and LBP features are extracted from local regions of retinal images. Then, a visual codebook is constructed using a K-means clustering algorithm. The cluster's centers are considered as visual words within the codebook. Each individual feature in the image is quantized to the nearest word in the codebook. The whole image is substituted by a global histogram which counts the frequencies of each word in the codebook. The resulting histogram size is the same as the number of words in the codebook and also the number of clusters achieved by the clustering algorithm. The final histogram representation is fed into a linear kernel SVM for classification.

3.1 Preprocessing

The main purpose of the preprocessing step is to reduce the inter and intra patient variability. According to the paper introduced by Cree et al.[8] the background-less fundus image has normally distributed colors. Thus, the image can be represented by the scalar mean μ and standard deviation σ throughout the entire image. If these two parameters are calculated for a reference image, it is possible to equalize the colors of the new image to the reference one in a more effective manner than simple histogram equalization.[9] In this work, the mean μ and std σ are empirically chosen for all datasets instead of computing them from a reference image. Furthermore, the preprocessing is applied only to the green channel rather than the three planes of the RGB color space. All images are resized to a height of 512 pixels, while maintaining the aspect ratio because of their large sizes. The

description of the process for a single color plane is explained as follows:

$$\begin{aligned}
\mu_{ref} &= 0.5 \\
\sigma_{ref} &= 0.1 \\
I_{out} &= I_{in} - \text{medianFilter}(I_{in}) \\
\mu_{out} &= mean(I_{out}) \\
\sigma_{out} &= std(I_{out}) \\
I^1_{out} &= (I_{out} - \mu_{out}) \div \sigma_{out} \\
I^2_{out} &= (I^1_{out} \times \sigma_{ref}) + \mu_{ref}
\end{aligned} \tag{1}$$

The background image is estimated by a median filter, whose size is approximately $\frac{1}{30}$ the height of the fundus image. I_{in} is the image to be equalized, I_{out} is the background-less image, and I^2_{out} is the equalized image. Fig. 1 shows an example for normal image equalization as well as drusen image equalization. Although the two images (Fig. 1 (a) and (b)) have different ethnic backgrounds and quality levels, the resultant (Fig. 1 (c) and (d)) images have very similar colors.

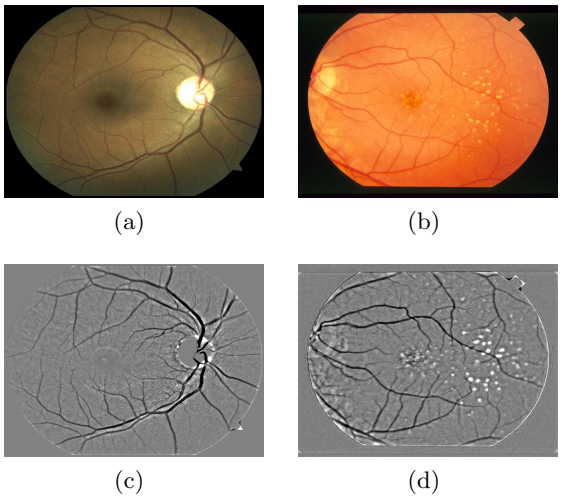

Fig. 1 (a) Normal image, (b) drusen image, (c) equalized normal image, and (d) equalized drusen image.

3.2 Feature Extraction

In this approach, SURF, HOG, and LBP features are extracted from the three channels of the RGB color space. Typically, the dimension of the SURF descriptors per image are 64 × number of interest points. Two strategies are adapted such as ordinary SURF and dense SURF (DSURF). In the first, SURF descriptors are extracted from all RGB color channels, then they are horizontally concatenated to get a feature matrix of a size 64 × total number of interest points extracted from the three channels. In the second, SURF descriptors are extracted from a dense grid uniformly distributed throughout the image i.e. SURF descriptors are computed on 16 × 16 pixel patches (non overlapping) with a spacing of 16 pixels. As opposite to ordinary SURF, for each patch we get a feature vector of a dimension 64, then for each image channel we get a feature matrix of a size 64 × number of patches. Finally, each image constitutes a feature matrix of a size 192 × number of patches by vertically concatenating each feature matrix. The implementation of SURF is done using mat-lab built in function. The HOG descriptors are obtained as similar to,[10] due to its lower dimension and discriminative capacity. Each image channel is divided into fixed number of blocks with a size of 32 × 32 pixels, then each block is subdivided into 4 cells (each cell is 16 × 16 pixels), as a result each block contributes to a feature histogram of a dimension 31. For each channel, the histograms are vertically concatenated forming a feature matrix of a size 31 × number of blocks. In total, the three channels will constitute a feature matrix of a size 93 × number of blocks. The null

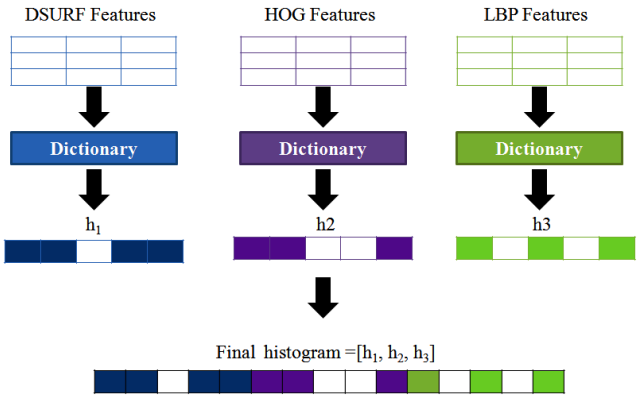

Fig. 2 Multiple based dictionary example. The set of features represent a single image in the training dataset.

descriptors originating from the black area surrounding the fundus image are not taken into account. The LBP descriptors are extracted from local patches of a size 32 × 32 pixels similar to the HOG. However, for each patch LBP features are computed using a 3 × 3 moving window centered at each pixel within the patch. The uniform local binary patterns are selected because of its lower dimension and reducing the number of codes inflicted by high frequency noise. The size of the feature matrix per image is 58 × number of patches, so in total we get a feature matrix of a size 174 × number of patches. HOG and LBP descriptors are implemented using the VLFeat open source library.[11]

3.3 Codebook Generation

The visual dictionary is generated using two different criteria such as; single-based criterion and multiple-based criterion. In the former, a single dictionary D_i is independently constructed from a pool of features, i.e. DSURF, SURF, HOG, or LBP using K-means clustering algorithm. Then, each image feature is quantized to its nearest visual word in every dictionary D_i. These visual words are combined into individual histogram h_i for each dictionary and the system performance is assessed accordingly. In the latter, similar steps are followed. However, the individual histograms are horizontally concatenated into a single histogram based on[12] i.e. $h = [h_1, h_2, \ldots, h_N]$. Fig. 2 shows an example of the multiple-based dictionary.

3.4 Classification

The classification's problem has been carried out using LIBSVM.[13] The data is separated into training and testing sets, where each example in the training set contains a class label and a unique histogram counting the frequencies of each visual word. Based on the training data, the objective of the SVM is to produce a model which is able to estimate the target values of the test data given only the test data histograms. Assume a training set of instance label pairs (x_i, y_i), $i = 1, 2, \ldots, l$ where $x_i \in \mathbb{R}^n$ and $y \in \{1, -1\}^l$ such that $y = +1$ for positive samples and $y = -1$ for negative samples, the SVM requires solution of the following Lagrange optimization problem:

$$\begin{aligned} &\min_{w,b,\xi} \ \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \\ &\text{subject to} \quad y_i \left(w^T \phi(x) + b \right) \geq 1 - \xi_i \end{aligned} \tag{2}$$

The training vectors x_i are mapped into a higher dimensional space by a kernel function $\phi(x)$. The SVM finds a linear hyperplane, which maximizes the margin $(\frac{1}{2} w^T w)$ in this higher dimensional space. $C > 0$ is the penalty parameter of the error term. This parameter is referred to as *bestc*, which should be tuned carefully during the training phase since it significantly affects the classifier performance. There are different kernel functions available i.e. linear, polynomial, radial basis function, and sigmoid. However, in our case we consider the linear one since it provides us with the best classification results. The linear kernel function is defined as $K(x_i, x_j) = x_i^T x_j$.

4. DATASET

In this research, we have used 430 images from six publicly available datasets as follows: STARE[*], DRIVE[†], DRIDB[‡], HEI-MED[§], MESSIDOR[¶], and HRF[‖], in addition to one private dataset obtained from the Oak Ridge National Laboratory, USA (ORNL). For the MESSIDOR dataset, images are taken at three different clinical sites. The distribution of these datasets is described as shown in Table. 1. We have employed 81 normal images,

Table. 1 Data distribution of Set A and Set B. MES1: MESSIDOR site 1, MES2: MESSIDOR site 2, and MES3: MESSIDOR site 3.

SETA			
	Normal	Drusen	Exudates
ORNL	18	30	10
HEI-MED	...	...	13
STARE	...	12	...
HRF	7	...	...
DRIDB	5	...	...
DRIVE	10	...	...
MSE1	...	...	115
MSE2	...	...	...
MSE3	...	...	...
Number of images	40	42	138

SETB			
	Normal	Drusen	Exudates
ORNL	18	31	10
HEI-MED	...	...	13
STARE	...	12	...
HRF	8	...	...
DRIDB	5	...	...
DRIVE	10	...	...
MSE1	...	...	...
MSE2	...	...	63
MSE3	...	...	40
Number of images	41	43	126

85 drusen images, and 264 exudate images obtained from (ORNL, HRF, DRIDB, and DRIVE), (ORNL and STARE), and (ORNL, HEI-MED, and MESSIDOR) respectively. The images are divided into two sets; Set A and Set B. Set A contains 220 images acquired from ORNL, HEI-MED, HRF, DRIVE, DRIDB, and only one clinical site of MESSIDOR named MES1 whereas Set B constitutes 210 images obtained from ORNL, HEI-MED, HRF, DRIVE, DRIDB, and two clinical sites of MESSIDOR named MES1 and MES1. The idea is to use Set A as a training set, then measure the system performance based on Set B and vice-versa. In this way, we can assess how well the system behaves when the test set contains different images than the ones included in the training set. This is usually called cross dataset testing. That means the proposed system (selected features, dictionary: single or multiple, and number of visual words) should be discriminative enough to classify the date present in the Set A based on Set B, and also the data present in Set B based on Set A. The system performance

[*] see (http://www.ces.clemson.edu/~ahoover/stare/)
[†] see (http://www.isi.uu.nl/Research/Databases/DRIVE/)
[‡] see (http://www.fer.unizg.hr/ipg/resources/image_database)
[§] see (http://vibot.u-bourgogne.fr/luca/heimed.php)
[¶] kindly provided by the Messidor program partners (see http://messidor.crihan.fr)
[‖] see (http://www5.cs.fau.de/research/data/fundus-images/)

is assessed using the accuracy measurement that is calculated as follows:

$$\text{Accuracy} = \frac{\text{Total \# of correctly classified images}}{\text{Total \# of images}} \% \tag{3}$$

5. RESULTS AND DISCUSSION

The classifier's parameter i.e. the value of C, which is referred to as *bestc* is computed by carrying out a class classification with 10 fold cross validation. Since K-means clustering algorithm (hard assignment) is employed, different values of K are used such as K = $[10, 20, 30, 40, 50, 60, 70, 80, 90, 100]$ in order to achieve satisfactory classification results. Two experiments are performed. The first experiment is to use Set B as a training set and Set A as a test set, while the second experiment is to use Set A as a training set and Set B as a test set.

5.1 Experiment 1

Regarding the single-based dictionary, the highest accuracy 98.63% is obtained using DSURF descriptors at K=70, subsequently HOG, SURF, and LBP achieve accuracies of 97.27% at K=100, 85.91% at K=90, and 91.36% at K=80 respectively. Neither SURF nor LBP descriptors provide satisfactory results as expected. On the contrary, HOG gives approximately similar results to DSURF 97.27% at K=100. Since we don't apply a preprocessing step to remove the optic disk, it might be confusing for the SURF or LBP descriptors to discriminate between normal and exudate images as the intensity characteristics of the optic disk is very similar to the exudate lesions. On the other hand, multiple-based dictionary approach overcomes the single-based dictionary. At K=100, an accuracy of 99.54% is obtained. In fact for all values of K, multiple-based dictionary approach achieves higher results than the single-based one, except at K=70 DSURF descriptors result 98.63% is slightly better than multiple-based 97.72%. Fig. 3 shows the resultant accuracy for all descriptors versus different values of visual words.

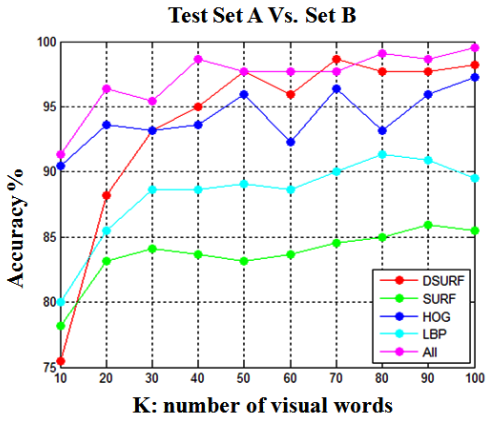

Fig. 3 Accuracy Vs. visual words K for a single and multiple based dictionary (test Set A Vs. Set B) . All: multiple based dictionary approach, DSURF, HOG, and LBP descriptors.

5.2 Experiment 2

With respect to the single-based dictionary, the HOG descriptors achieve the highest accuracy 97.14% at K=100, after that DSURF, SURF, and HOG achieve accuracies of 90.95% at K=50, 85.23% at K=80, and 84.76% at K=70 respectively. SURF and LBP descriptors attain relatively similar results as before. We can notice that the DSURF descriptors don't attain similar performance in both experiments owing to the sharp decrease in accuracy from 98.63% to 90.95%. However, the HOG descriptors achieve satisfactory results with the two experiments which implies the discriminative capacity of these descriptors. Once more, the multiple-based dictionary approach overcomes the single-based dictionary as at K=100 a 100% accuracy is achieved. Furthermore, for all visual words, it achieves higher results than the single-based method as shown in Fig. 4. So far, we can conclude that

the multiple-based approach achieves significant results in both conditions, which indicates the importance of integrating several descriptors in the task of diabetic retinopathy diagnosis. As we discussed in section 4 the proposed approach should be able to discriminate the data present in Set A based on Set B and vice-versa, the multiple-based approach managed to accomplish this task with satisfying results such as 99.54%, 100% for the first and second experiment respectively and a mean accuracy of 99.77%.

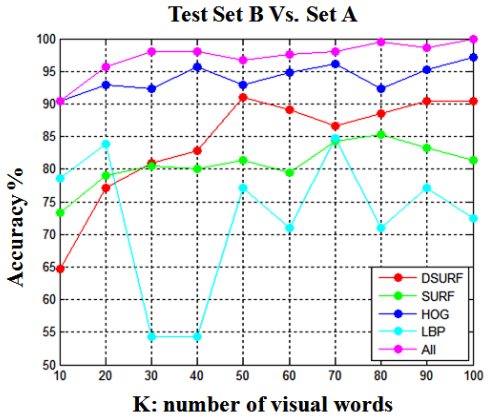

Fig. 4 Accuracy Vs. visual words K for a single and multiple based dictionary (test Set B Vs. Set A). All: multiple based dictionary approach, DSURF, HOG, and LBP descriptors.

6. CONCLUSION AND FUTURE WORK

In this work, a bag of words approach was employed in order to discriminate between normal fundus images and abnormal fundus images with bright lesions, specifically drusen and exudates. We have proposed to use single-based and multiple-based dictionaries. In the first, a single dictionary is constructed from DSURF, SURF, HOG, or LBP descriptors, after that a histogram of word occurrences is generated for each image and the system performance is assessed accordingly. In the second, the image gets a histogram from each dictionary. Then, all histograms are horizontally concatenated to form a single histogram, where each feature gets N entries in the histogram, one from each dictionary. The two schemes are evaluated on different datasets. We achieved a mean accuracy of 97.2% with respect to the single-based dictionary, while our best accuracy is obtained using the multiple-based dictionary with a mean accuracy of 99.77% which reflects the discriminative capacity of this approach. To conclude, the bag of words approach can play a significant role in the classification of normal fundus images and abnormal fundus images with bright lesions. It can also help physicians in the early diagnosis of diabetic retinopathy as exudates might be the only sign of diabetic retinopathy. In the future, we would increase the size of the datasets and perform more experiments, introduce a preprocessing step to localize and segment the optic disk, and extend the proposed approach to deal with more challenging spot lesions, namely microaneurysms.

REFERENCES

1. D. S. Fong, L. Aiello, T. W. Gardner, G. L. King, G. Blankenship, J. D. Cavallerano, F. L. Ferris, and R. Klein, "Retinopathy in diabetes," *Diabetes Care* **27**(suppl 1), pp. 84–87, 2004.
2. S. Wild, G. Roglic, A. Green, R. Sicree, and H. King, "Global prevalence of diabetes: Estimates for the year 2000 and projections for 2030," *Diabetes Care* **27**(5), pp. 1047–1053, 2004.
3. G. Yen and W.-F. Leong, "A sorting system for hierarchical grading of diabetic fundus images: A preliminary study," *Information Technology in Biomedicine, IEEE Transactions on* **12**, pp. 118–130, Jan 2008.
4. M. Niemeijer, B. van Ginneken, S. R. Russell, M. S. A. Suttorp-Schulten, and M. D. Abrmoff, "Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis," *Investigative Ophthalmology & Visual Science* **48**(5), pp. 2260–2267, 2007.

5. M. van Grinsven, A. Chakravarty, J. Sivaswamy, T. Theelen, B. van Ginneken, and C. Sanchez, "A bag of words approach for discriminating between retinal images containing exudates or drusen," in *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pp. 1444–1447, April 2013.
6. R. Pires, H. Jelinek, J. Wainer, S. Goldenstein, E. Valle, and A. Rocha, "Assessing the need for referral in automatic diabetic retinopathy detection," *Biomedical Engineering, IEEE Transactions on* **60**, pp. 3391–3398, Dec 2013.
7. Ujjwal, K. Deepak, A. Chakravarty, and J. Sivaswamy, "Visual saliency based bright lesion detection and discrimination in retinal images," in *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pp. 1436–1439, April 2013.
8. M. J. Cree, E. Gamble, and D. Cornforth, "Colour normalisation to reduce inter-patient and intra-patient variability in microaneurysm detection in colour retinal images," *Workshop on Digital Image Computing* , pp. 163–169, 2005.
9. L. Giancardo, F. Meriaudeau, T. P. Karnowski, Y. Li, S. Garg, K. W. T. Jr., and E. Chaum, "Exudate-based diabetic macular edema detection in fundus images using publicly available datasets," *Medical Image Analysis* **16**(1), pp. 216 – 226, 2012.
10. P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **32**, pp. 1627–1645, Sept 2010.
11. A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms." http://www.vlfeat.org/, 2008.
12. M. Aly, M. Munich, and P. Perona, "Using more visual words in bag of words large scale image search," tech. rep., Caltech, USA, 2011.
13. C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology* **2**, pp. 27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.