Feature Selection using Constellation Models

Heydar Maboudi Afkham, Carl Herik Ek and Stefan Calsson
Computer Vision and Active Perception Lab.
Royal Institute of Technology (KTH)
Stockholm, Sweden
{heydarma,chek,stefanc}@csc.kth.se

Abstract

In this paper we propose a new approach for learning low-dimensional image features that retains class discriminative properties while simultaneously generalising between within class variations. Our approach is based on the concept of a joint feature where several small features are combined in a spatial structure. The proposed framework automatically learns the structure of the joint constellations in a class dependent manner improving the generalisation and discrimination capabilities of the local descriptor while still retaining a low-dimensional representations. As such the framework is capable of learning from small data-sets where previous approaches becomes severely affected by the curse of dimensionality.

1. Introduction

The ideal characteristics of an image feature is such that will robustly extract interclass variation (discrimination) while simultaneously be insensitive to intraclass variations (generalisation). Further, it should be low-dimensional such that sufficient statistics can be extracted from the data that is available. One approach to satisfy both the low-dimensionality and the generalising capabilities is to extract variations on a very small scale. However, these variations will also generalise between classes. One approach to facilitate discrimination is to consider several local features in a constellation rather than single features in isolation see Fig 1. Learning such a constellation is a challenging task as it will induce a combinatorial explosion in the possible joint configurations. The traditional approach in the computer vision literature have been to avoid this problem completely by using a single feature whose support covers a much larger region or for problems where a semantic structure exists fix the structure a-priori [4]. The first contribution of this paper is a method capable of learning such features directly from data in a principled manner. We refer to such a constellation as a *joint feature*. The second contribution of the paper is a novel method for summarising sets of local feature responses. The traditional approach of summarising an unordered set of local features in computer vision is a Bag-of-Words descriptor (BoW). In this paper we propose a different approach for summarising a set of responses into a single descriptor. Similarly to a BoW approach we use a vocabulary to model the feature space but rather than using *quantity* we use a class dependent *quality* measurement of each word as a descriptor of the set. This class dependent view is an essential part of our approach as it facilitate adaptive learning of joint features in order to achieve a better balance between discrimination and generalisation in the final descriptor. In specific, our method initially considers features in isolation and then gracefully adapts features through the creation of joint structures such that the final representation can discriminate with between

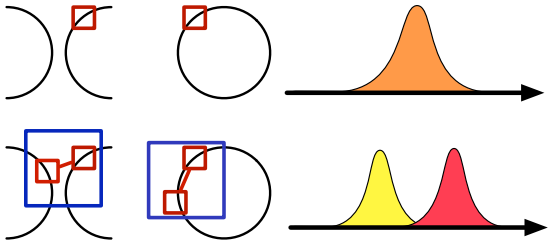

Figure 1. The above figure depicts a schematic of the approach presented in this paper. The right-most pair of images show two shapes belonging two separate "classes" that we wish to discriminate. As the shapes are generated as spatial permeations of each other the local statistics, which we have sufficient data to reliable extract, will be the same (top image middle column). In order to recover discriminative information larger spatial structures needs to encoded. Either by using a less local features (blue square) or use spatial combinations of local features. In the first case the dimensionality of the feature will explode which will require a significantly much more training data while in the later case finding the discriminative structure leads to a combinatorial problem which will be very challenging to approach. In this paper we propose a method to handle the combinatorial problem an learn low-dimensional feature that generalises well while at the same time being discriminative.

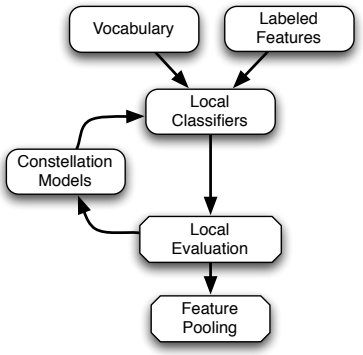

Figure 2. This figure illustrates the pipeline of the proposed descriptor in this paper. In this framework, we make use use of an already trained vocabulary and labeled local features to train a number of local classifiers which are used to summarize the image using qualitative information. In this work we will show that by gradually replacing the poor local classifiers with discriminative constellation models it is possible to obtain a richer summarization for the images.

the classes while keeping the dimensionality of the feature as low as possible.

We will now proceed to relate our approach to the current literature.

2. Related Works

Vocabulary based models are very popular for image and object description and recognition [14, 12, 8, 10]. However, these models were originally developed for text and document processing [6, 9] where the notion of vocabulary is well-defined. For images the notion of words and vocabularies are not as obvious and have the topic of much work. One of the most influential works in this area is the work by Savarese *et al.* [13] in which they a well-defined visual vocabulary is built by introducing relational spatial constraints in calculation of the vocabulary. Similarly [17] have shown that it is possible to improve the quality of the inference obtained from the words by incorporating higher order relational information. Our method shares the same goal as [11] which is encoding the relational information at the feature level rather than between the different visual words. In our work we show that collecting relational information is not required for all the words in the vocabulary and it is possible to obtain richer descriptors by just gathering this information for a small fraction them. The framework used in this paper uses a series of local classifiers based on the vocabulary to obtain qualitative information from the features. Methods such as [16, 7] share the same goal of building category-specific visual words with the difference that our method method builds them based on a standard visual vocabulary rather than redefining the notion of visual word.

To encode the relational information our work uses constellation models for locating the most informative features around each word and uses this information for the inferences. Constellation models and pictorial structures have played an important role in the development of computer vision in the past decade [18, 3, 4, 5]. These models consist of a root "part" connected to a number of flexible parts that are connected to the root feature. Due to the built-in flexibility of these models they can adapt to the intra-class variation to form better and more robust descriptors for the objects. The parts involved in the constellation models are either learnt using ground truth labeling [18] learned well aligned data using strong heuristics on their location and sizes [4]. Studies such as [2] have shown that in the later case alignment and number of components play a more important role than the flexibility of the parts. In this paper we are applying constellation models to a less controlled scenario where the root feature corresponds to the instances of a certain visual word. Where these instances appear is only governed by the local structure in the image, which can be completely different in a global context. The support features are chosen from a much larger area than compared to the size of the root feature. This will enables us to extract rich features while still retaining a low dimensional feature.

The pipeline of our method for describing images can be seen in Fig. 2. We make use of an already trained vocabulary to train single feature local classifiers and then gradually replace with more complex constellation models. In the experiments section we will show that this replacement has an significant impact on the final classification performance.

The reminder of the paper is structured as follows: in §3 we describe first our qualitative descriptor used for summarising a set of local descriptors and then how joint features can be learned. We then proceed to experimentally evaluate the performance of the approach in §4 both in terms of quantitative and qualitative experiments. Finally §5 concludes the paper.

3. Methodology

In this work we introduce a different approach toward the visual vocabularies. The information is gathered by our method used for describing regions relies on finding the most representative features for each word. The goal here is to show, by using constellation models and joint features it is possible to increase the performance of the vocabulary based methods *without changing the vocabulary itself*.

For better understanding of the methodology and showing how constellation models are used for improving the inference, this section is divided into four subsections. In §3.1 we describe how qualitative information is gathered based on single features from the region. This descriptor consists of a large number of local classifiers assigned to different visual words and provides a ground which enables us to em-

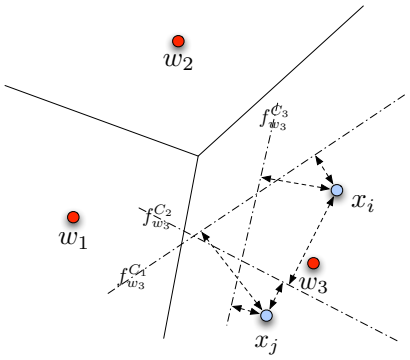

Figure 3. This figure shows how the differences between the features assigned to a visual word can be highlighted by employing class specific classifiers in that region.

ploy the constellation models. §3.2 discusses the conditions in which constellation models can be beneficial for improving the performance of the proposed descriptor. The result of this process is the selection of a number of visual words as root features for constellation models. §3.3 describes the details of how these models are trained and used. Finally §3.4 discusses some details of the methodology.

3.1. Qualitative Descriptor

The fundamental principle underpinning a bag-of-words approach is that the elements of the dictionary $\mathbf{D}$ capture the local structures of the image. Here the goal is to measure the quality of these structures with respect to different target classes in a discriminative manner and use this information to describe the image. In other words the question being asked in this paper is "How representative of the word is the feature?" rather than "How often a word is seen?". In this work we exploit the differences between the features assigned to a certain visual word based on the object class they have appeared on. While this difference is usually ignored by quantitative approaches, we show how by exploiting this information it is possible to obtain a richer summarization.

To measure the quality of the features assigned to the different visual words lets assume that $\{(x_1, y_1), ..., (x_n, y_n)\}$ is a set of labeled local features extracted from an image dataset with $y_i \in \{C_1, ..., C_M\}$ and $\mathbf{D}$ is an already trained vocabulary with N words. The goal here is to train class specific classifiers, f_w^C, for the features assigned to the word w. These classifiers are trained by selecting the features assigned to the word w and creating a binary labeling where features with $y_i = C$ are assigned to the positive and others to the negative set. Each classifier is formulated as

$$f_w^C = \arg\min_f \frac{1}{N} \sum_x \mathcal{L}(x, \bar{y}^C; f) + \lambda |f|^2. \qquad (1)$$

Here the x is chosen only from the features with $l(x) = w$, $\bar{y}^C$ represents the binary labeling of these features with respect to class C and $\mathcal{L}(x, \bar{y}^C; f)$ is a given loss function. In figure 3 we can see that the two features x_i and x_j are both labeled as w_3 have a different behavior with respect to the hyperplanes $f_{w_3}^{C_1}$, $f_{w_3}^{C_2}$ and $f_{w_3}^{C_3}$ which encode class properties in this section of the space. To estimate the quality of a feature (the likelihood of belonging to class C while assigned to the word w), we use the logistic function

$$P_w^C(x) = \frac{1}{1 + \exp(-a(x^T f_w^C))}. \qquad (2)$$

For any set of features extracted from an image we wish to build a descriptor based on their visual word quality or certainty. Each word in $\mathbf{D}$ captures a certain structure on the image. Therefore, the role of $P_w^C(x_i)$ function, Eq. 2, is to measure the quality the discovered structures assigned to the word w with respect to class C. This is a one dimensional measurement corresponding to the models confidence. To that end it is possible construct a (N, M) dimensional descriptor D, with N being the size of the vocabulary and M the number of classes. Each dimension of this vector corresponds to responses associated with a certain word (w_n) with respect to a certain class (C_m). The question here is how one can summarize these values into a number that can capture the qualitative properties of features seen in the image. Here we analyze the *max descriptor* defined as

$$D_{max}[i] = max\left\{P_{w_n}^{C_m}(x) : x \in I, l(x) = w_n\right\}, \qquad (3)$$

which focuses on pooling the features with the most confident rather than relying on the quantitative properties of their labeling. This can also be seen as a feature selection problem, where the highest likelihood features are used for describing the image.

3.2. Analysis of local classifiers

The role of f_w^C is to determine if the features assigned to the word w belong to class C or not (binary classification). Is possible to construct local classifiers such that the descriptor D_{max} (Eq. 3) can perfectly distinguish different object classes? To answer, we take a look at the behaviour of the obtained local classifiers in the previous section. Each local classifier f_w^C can be scored by calculating its empirical loss on the training set given by,

$$\mathcal{L}_{emp}(f_w^C) = \frac{1}{N} \sum_x \mathcal{L}(x, \bar{y}^C; f_w^C). \qquad (4)$$

The value of the empirical loss can be used as heuristic notion for evaluating the behaviour of the local classifiers. The classifiers with less miss-classification tend to have a lower empirical loss than the ones with high miss-classification. In other terms, the lower the empirical loss the more accurate the classifier f_w^C is in separating the data. Classifiers

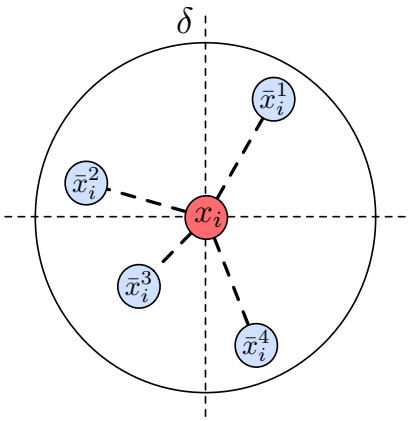

Figure 4. *Feature architecture* defines how joint features are constructed. To create a joint feature the spatial region surrounding each feature x_i is divided into four quadrants and each support feature is selected from one of the quadrants.

with high empirical loss tend to make more noisy decisions on the data which makes the resulting descriptor D_{max} more noisy. In our observations a low empirical loss can happen in three different cases. For a binary classifier f_w^C, the value of $\mathcal{L}_{emp}(f_w^C)$ is low if one of the following conditions is met: **(a)** The word w has very distinctive properties for class C which is resulting in a strong classifier, **(b)** The word w is only frequent on the positive data or **(c)** it is only frequent on the negative data. The main remaining type of words are the ones that are frequent on both positive and negative regions but the feature is not distinctive enough for construction of a strong classifier. To improve the quality of the local classifiers not much can be done for the ones with low empirical loss, since they are either very strong or we lack sufficient data for training. We now focus on how a constellation model can be used for improving the accuracy of the classifiers with high empirical loss.

3.3. Joint Features

As mentioned, a property of features assigned to a word, w with high empirical loss is that it is frequent on both positive and negative regions and therefore is not discriminative. Since all the instances of w from the positive set share the property that they have appeared on the same object class it is possible to couple instances of w with more distinctive features of that object class to build a richer joint feature and use that in the summarization D_{max} (Eq. 3). In this work we will treat the joint features as constellation models.

Let's assume that $\{(x_i, \bar{y}_i^C)\}_{i=1}^N$ are the features assigned to w, which has a high empirical loss with respect to class C. Here the goal is to find a series of local features $\bar{x}_i^1, \ldots, \bar{x}_i^n$ in the support region of each x_i such that the concatenated vectors $\{([x_i, \bar{x}_i^1, \ldots, \bar{x}_i^n], \bar{y}_i^C)\}_{i=1}^N$ are linearly separable according to the binary labeling. To formulate this selection let

$$\mathbf{F}_{x_i} = \{[x_i, \bar{x}_i^1, \ldots, \bar{x}_i^n] : \bar{x}_i^j \in N_\delta(x_i)\}, \tag{5}$$

be the set of all possible joint features centered at x_i where n features are chosen from a spatial neighborhood with size δ of the feature x_i on the image. This set will be referred as the feature cloud of x_i. In this work the features in the neighbouring of x_i is partitioned into four quadrants and each of the four support features is selected from a different quadrant. The visualization of this structure can be seen in Fig. 4. For a given linear decision boundary β we define the function,

$$\Phi(\mathbf{F}_{x_i}, \beta) = \underset{\phi \in F_{x_i}}{\arg\max} \{\phi^T \beta\}. \tag{6}$$

The role of this function is to select a joint feature within the feature cloud $\mathbf{F}_{x_i}$, which best represents the decision boundary β. Using this definition each decision boundary imposes a different feature selection and changes the original classification problem into $\{(\Phi(\mathbf{F}_{x_i}, \beta), \bar{y}_i^C)\}_{i=1}^N$. With this change the feature selection problem is reduced to finding the decision boundary β such that its corresponding joint features are linearly separable with respect the binary labeling. As it can be seen in this formulation the data changes with the changes of β. While this makes the problem non-convex but gradually updating β using Alg. 1 will always increases the performance on the training set. It is easily noticed that the feature vectors obtained using Eq. 6 given a decision boundary β, will not necessarily have β as their optimal decision boundary. Here the role of the optimization is to find a decision boundary that its corresponding data (Eq. 6) reproduces the same decision boundary. In this work we are solving this optimization problem as a two step gradient descend approach. This process is shown in Alg. 1, where β is initialized as a ones vector which corresponds to making a feature selection from which every dimension of the composite feature is treated equally. In this algorithm the convergence of β is controlled by the learning rates $\eta_{(k_1,k_2)}$ which are set experimentally.

3.4. Effect of the loss function

So far the discussion was based on a generic loss function $\mathcal{L}(x, y; \beta)$. In this section we discuss the effect on the output with respect to different loss functions. In a more precise manner we will compare the Squared Error Loss ($|y - x^T\beta|^2$) with the Hinge Loss ($max(0, 1 - y.x^T\beta)$). These loss functions can be viewed as the extremes with respect to how the data points are used for finding the optimal decision boundary. In a binary classification problem each of the loss functions impose a different strategy on the optimizer for finding an optimal decision boundary. Squared

Algorithm 1 Feature Selection

Input: $\mathcal{C} = \{(\mathbf{F}_{x_i}, y_i)\}_{n=1}^{N}$
Output: Decision boundary β

Select initial β
for $k_1 \leftarrow 1$ **to** $niter1$ **do**
  $\bar{x}_i \leftarrow \Phi(\mathbf{F}_{\mathbf{x_i}}, \beta)$, $\forall i$; (Eq. 6)
  for $k_2 \leftarrow 1$ **to** $niter2$ **do**
    $\eta_{(k_1,k_2)} \leftarrow$ Learning rate of this stage
    $\nabla f(\beta) \leftarrow \frac{1}{N} \sum_x \nabla \mathcal{L}(\bar{x}_i, \bar{y}_i; \beta) + \lambda \nabla |\beta|^2$
    $\beta \leftarrow \beta - \eta_{(k_1,k_2)} \nabla f(\beta)$
  end for
end for
return β

Error Loss (SEL) uses all the data points for finding the optimal decision boundary. Due to this the decision boundary obtained using this loss function encodes both how the two classes are separated and how the data is distributed on each side of the boundary. On the other hand Hinge Loss (HL) focuses on only a small fraction of data points for obtaining the decision boundary. While this reduction of the data leads to a faster convergence, the distribution of the points will no longer be encoded within the decision boundary. As will be shown in the results section (§4) the models trained using SEL in Eq. 1 provide a better image summarization than the ones trained using HL. Unfortunately due to computational limitations, it is only feasible to use HL in the Alg. 1 to search the joint feature space.

4. Experiment Setting and Results

The experimental focus of this paper is on two aspects of the proposed method. First, we show how the summarization presented in §3.1 performs against the standard BoW model and how the choice of the loss function affects the quality of the summarization. Second, we evaluate how the quality of the summarization is improved when composite feature selection is employed. The experiments are conducted on the MSRCv2 dataset [15]. Although this dataset is relatively small compared to other datasets, it is considered as a challenging and difficult dataset due to its high intra-class variation. The main focus of the experiments is to show how by incorporating features that gracefully adapt to intra class variations by adding joint features, it is possible to provide better description of the image regions while still avoiding an explosion in dimensionality thus running the risk of over fitting. In this work we have followed the experiments setup used in [17, 11] in which nine of fifteen classes are chosen ({*cow, airplanes, faces, cars, bikes, books, signs, sheep, chairs*}) with each class containing 30 images. The focus of these experiments is to summarize the whole image into one vector and predict the class labeling of the images based on this vector. For each experiment, the images of each class were randomly divided into 15 training and 15 testing images and no background was removed from the images. The random sampling of training and testing images were repeated 5 times to eliminate the train and test partitioning and in all experiments SIFT features are densely sampled at every 5 pixels from multiple scales. In this work we will show how by employing a class dependent summarization and feature selection it is possible to recover a robust description for such a small and challenging dataset.

The visual vocabulary plays an important role in our method and compares our approach with previously published methods. The main difference between this method and other vocabulary based methods is how the vocabulary is used to summarize the image. To compare the performance of bag-of-words histogram with the proposed descriptor, visual vocabularies with different sizes {*50, 100, 200, 300, 400, 500, 1000, 1500, 2000*} were computed over the training subset using standard k-means algorithm. The same vocabulary was shared between all methods.

To efficiently search for joint features, we rely on a predefined search structure. This structure can be seen as the "feature architecture" as it defines how composite features are constructed. While there are many different ways to define this architecture, we focus on a simple constellation model with four support features. For a root feature x_i its neighbouring features are partitioned into four quadrants and each of the support features is selected from a different quadrant. While the size of each single feature is 16×16 pixels the neighbourhood size of the constellation δ, is chosen to be 60 pixels. A visualization of this architecture can be seen in Fig. 4. Even though we are only presenting the results with four support features, we would like to note that our framework will work with any architectures.

4.1. Experiments on MSRCv2 [15]

In this section we present the experiments by comparing the performance of the method described in Sec 3.1 compares against the standard BoW descriptor. To conduct this experiment we train a linear SVM classifier [1] on the obtained image descriptors $D_{\max}$ (Eq. 3). The results of this experiment can be seen in Fig. 5(Left). To provide a fair comparison the BoW model was trained using both *linear* and *RBF* kernels and as it can be seen in Tab. 1 their performance is comparable to the previously published results on this dataset. Having this as a baseline we can see the linear SVM models trained on top of $D_{\max}$ descriptors out perform not only the baseline but also previously published results by a large margin. By taking another look at Fig. 5(Left) we can see that the summarization based on Square Error Loss (SEL) is providing a significantly better performance than the ones with Hinge Loss (HL). To verify

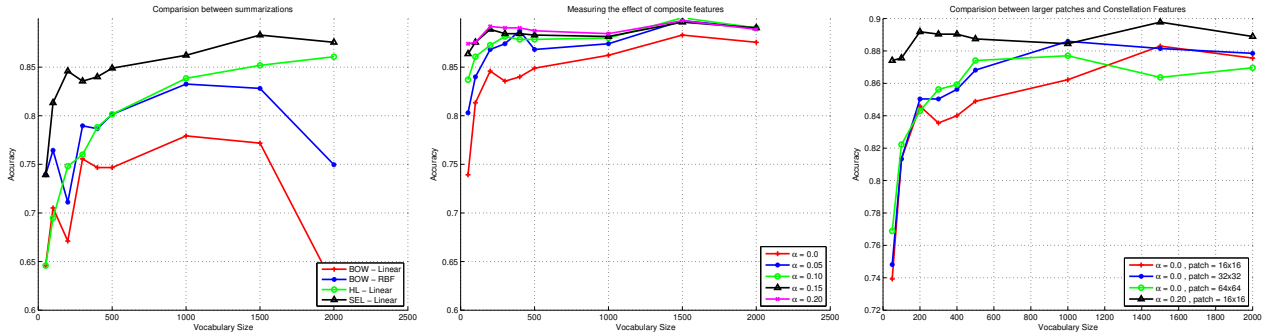

Figure 5. **(Right)** This plot compares the performance of $D_{\max}$ descriptor trained using both Square Error Loss (SEL) and Hinge Loss (HL) with Bag-Of-Words (BoW) descriptor. It can be clearly seen that linear SVM classifier on $D_{\max}$ is out performing that BoW trained using both linear and rbf SVM classifier. **(Middle)** This plot shows how gradually replacing the local classifiers with joint classifiers in the summarization $D_{\max}$ improves the over all performance of the descriptor in all vocabulary sizes. Here α represent the fraction of the classifiers that are replaced by constellation models. **(Left)** To show the effect of discriminative feature selection this figure compares the performance of $D_{\max}$ with $\alpha = 0.2$ with vocabularies built with spatially larger SIFT features. It is clear to see how discriminative feature selection is outperforming larger SIFT features that cover the support region of the constellation models.

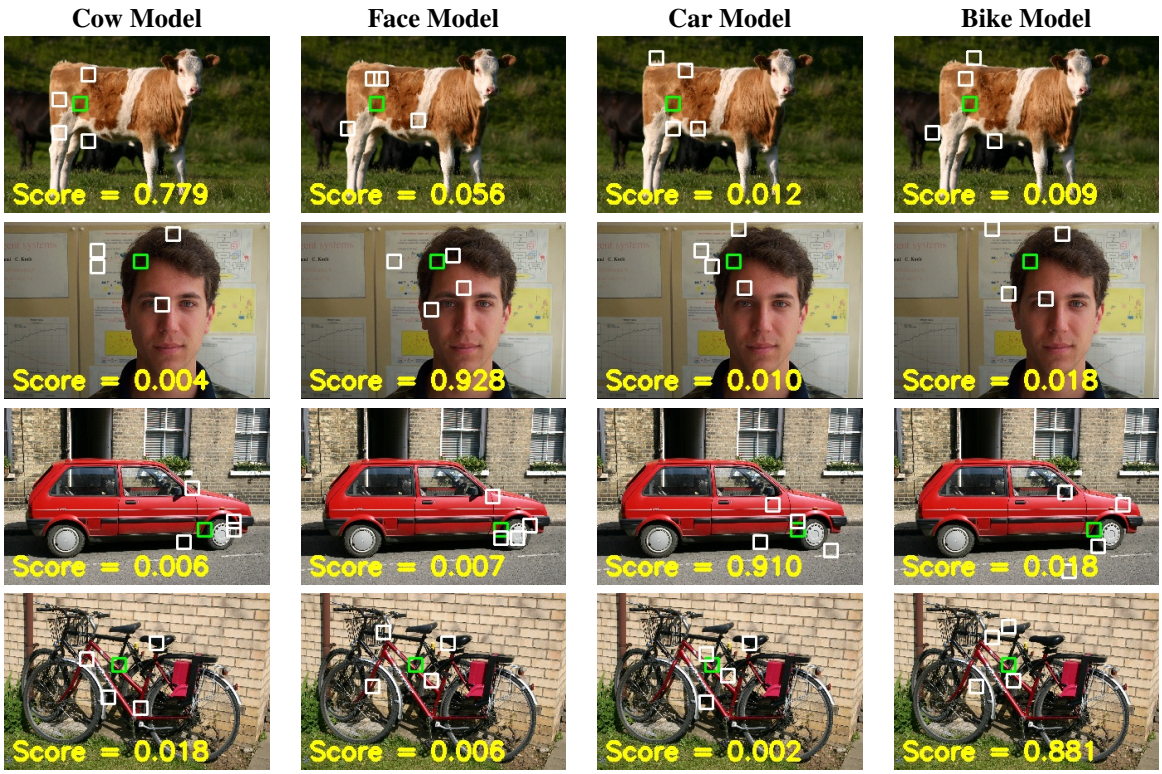

Figure 6. **(Best viewed in digital format)** The feature selection is optimized for each object class separately and the score represents how good the located features fit the model. In this figure a word w (green patch) is selected which was a candidate for joint feature selection in all object classes. **Each column visualises the feature selection** (white patches) **done for a different object class.** Based on the score of the feature selection it can be seen how sensitive the method is to the features that exist on the object class.

whether this is due to how these loss functions score the data points and not the difference in the decision boundaries, we have calculated the average accuracy of these models for both loss functions and the accuracies were too close to result in such a difference in the over all recognition performance.

After seeing how this summarization preforms against other methods, we ask the question whether the extra in-

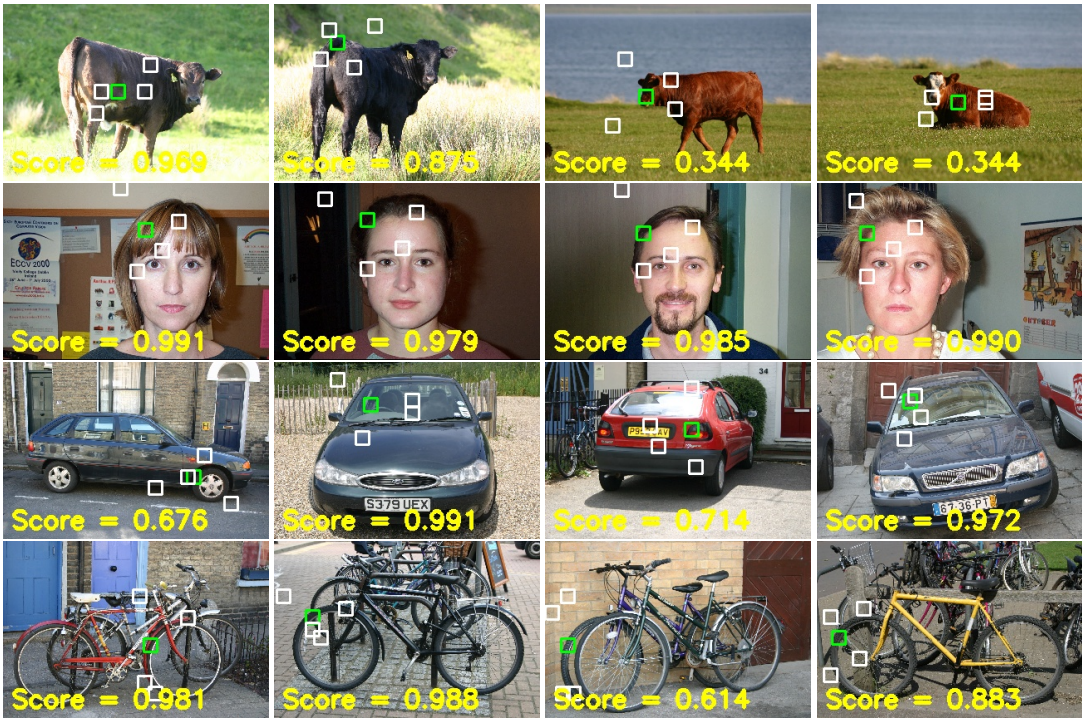

Figure 7. **(Best viewed in digital format)** This figure shows how the feature selection is consistent across each object class. In this figure the same word as Fig. 6 is selected on different instances of each object class and it can be seen there is a large consistency in feature selection done for each class. It should be mentioned consistency in feature selection does not necessarily imply that the root should also lay in a globally similar context. Since a visual word can appear on many different global structures, it is the role of the optimizer to select support features that are discriminative and is shared between these global structures.

Method	Acc %
2^{nd} order spatial [17]	$78.3 \pm 2.6\%$
10^{th} order spatial [17]	$80.4 \pm 2.5\%$
QPC [11]	$81.8 \pm 3.4\%$
LPC [11]	$83.9 \pm 2.9\%$
BOW *Linear*	$78.0 \pm 5.0\%$
BOW *RBF*	$83.2 \pm 4.0\%$
$D_{\max}$ - HL	$86.0 \pm 4.0\%$
$D_{\max}$ - SEL	$88.3 \pm 3.6\%$
$D_{\max}$ - ($\alpha = 0.10$)	$90.0 \pm 3.2\%$

Table 1. Comparison between the classification rates obtained by the proposed method and the previously published methods on MSRCv2 dataset.

formation encoded in representing some of the classifiers as constellation models will affect the classification performance or not. To test this hypothesis we use the summarizations based on Square Error Loss (the winning summarization) and gradually change the bad classifiers (§3.2) into constellation models. Here we are introducing a parameter α which represents the fraction of the words presented as constellation models. The results of this experiment can be seen in Fig. 5(Middle). Here we can see how the overall performance of the method improves as the value of α increases. This improvement is more noticeable for small sized vocabularies as their performance is pushed toward the performance of large vocabularies with the increase of α. In short this experiment shows that the overall performance is converging toward a performance independent of the size of the vocabulary. Importantly this means that we can reduce the dimensionality of the descriptor by using a smaller vocabulary compared to the other methods thus circumventing the curse of dimensionality.

Since the constellation features collect information on a broader region on the image one might argue that having larger base features will show a similar behaviour. To examine this we compare the average performance of $D_{\max}$ descriptor with single features of different sizes over all the vocabulary sizes with descriptors using composite features. The results of this comparison are seen in Fig. 5(Left). As it can be seen even though the performance of the single feature descriptor increases with the size of the base feature but still they are out performed by the summarization employing joint features.

4.2. Visualization of Constellation Features

Local features such as SIFT, capture very local edge configurations and textural information of the image. While the information captured by these local features is in no sense related to class semantics, the aim of this experiment is to show how class semantic information can be encoded within the joint local features. With each constellation model imposing a different feature selection on the image, the expectation is to see that the feature selection best fits the class it is optimized for. The quality of these feature selections is measured according to the discussions in §3.1 and §3.3. To continue the discussions we randomly select a word w from a vocabulary of size *500* from the words that has been a candidate for joint feature selection in several classes. The green patch shown in Fig. 6 represent the instance of this word found on different object classes. As it can be seen in none of the classes this patch contains very discriminative information. The white patches on each image are the support features found on the image according to the architecture shown in Fig. 4. Here each column represents the feature selection imposed by each object model and it can be seen how the relative score changes from the object class they were trained on, to the other object classes. For example in the first column the feature selection is done based on the constellation learnt over the *cow* class and it can be seen how the other classes are not well represented by this feature selection. The reason for this large difference in the scores is the fact that even though the best candidate is selected from negative regions, it fails to provide the proper textural, edge configuration and the composition that exists on the positive object class. Another important property of the feature selection which should be noticed is how consistent is their selection across their native object class. To show this Fig. 7 shows the feature selection on the different instances of the object using the same previously selected visual word. It is interesting to notice that the method shows a large consistency in selection of the local features while the over all configuration can lay in a completely different context. This behaviour is expected since there is no control on where a visual word appears on the object and it is interesting to see how the optimization process described in §3.3 finds local features that are shared and are stable across this contextual variation.

5. Conclusion

In this paper we proposed a method capable of extracting flexible, class dependant local joint features. These features capture both the intra-class and discriminative variations while still being low-dimensional making them applicable to setting when training data is scarce. Our approach makes use of the "quality" of a local feature when describing an image. This not only provides a better summarization of the images but also allows for automatically learning joint features. We have shown that the propsed method is capable of using a much smaller vocabulary while still retaining discriminative information with respect to object classes compared to traditional BoW methods. The proposed method significantly outperforms the base line algorithms on a very challenging data-set.

References

[1] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011. 5

[2] S. K. Divvala, A. A. Efros, and M. Hebert. How important are 'deformable parts' in the deformable parts model? In *ECCV*, 2012. 2

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010. 2

[4] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61, 2005. 1, 2

[5] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *ECCV*, 2004. 2

[6] T. Joachims. Text categorization with suport vector machines: Learning with many relevant features. In *ECML*, 1998. 2

[7] D. Larlus and F. Jurie. Latent mixture vocabularies for object categorization and segmentation. *JIVC*, 2009. 2

[8] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 2

[9] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. 2

[10] M. Marszałek and C. Schmid. Spatial weighting for bag-of-features. In *CVPR*, 2006. 2

[11] N. Morioka and S. Satoh. Building compact local pairwise codebook with joint feature space clustering. In *Proceedings of the 11th European conference on Computer vision: Part I*, ECCV, 2010. 2, 5, 7

[12] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006. 2

[13] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlatons. In *CVPR*, 2006. 2

[14] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *ICCV*, 2005. 2

[15] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, 2005. 5

[16] L. Yang, R. Jin, R. Sukthankar, and F. Jurie. Unifying discriminative visual codebook generation with classifier training for object category reorganization, 2008. 2

[17] Y. Zhang and T. Chen. Efficient kernels for identifying unbounded-order spatial features. In *CVPR*, 2009. 2, 5, 7

[18] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012. 2