Loose Shape Model for Discriminative Learning of Object Categories

Margarita Osadchy and Elran Morash
Computer Science Department
University of Haifa
Mount Carmel, Haifa 31905, Israel
rita@cs.haifa.ac.il

Abstract

We consider the problem of visual categorization with minimal supervision during training. We propose a part-based model that loosely captures structural information. We represent images as a collection of parts characterized by an appearance codeword from a visual vocabulary and by a neighborhood context, organized in an ordered set of bag-of-features representations. These bags are computed in a local overlapping areas around the part. A semantic distance between images is obtained by matching parts associated with the same codeword using their context distributions. The classification is done using SVM with the kernel obtained from the proposed distance. The experiments show that our method outperforms all the classification methods from the PASCAL challenge on half of the VOC2006 categories and has the best average EER. It also outperforms the constellation model learned via boosting, as proposed by Bar-Hillel et al. on their data set, which contains more rigid objects.

1. Introduction

We consider the problem of generic visual categorization: Given an image, categorize it into one of the considered visual categories according to its semantic content. The training is performed on unsegmented images with significant clutter. No information about object location, size, or pose is available. The only information provided is the category label of the image.

Current part-based methods can be roughly divided into 1) pure appearance methods that discard all shape information, but are fairly flexible and computationally efficient[19, 17, 18, 9, 20]; 2) methods that incorporate shape either as a generative model of locations of parts [7, 6, 1] or by geometric correspondence search [2, 13]. These models are close to rigid, computationally expensive, and some require a bounding box during training [15, 23]. As was noted by many authors [7, 5, 6, 12], the geometric information is important and should not be ignored. However, because shape usually varies a lot, it is easier to discard it than model it [20].

In this work we consider a compromise between the two opposing views and propose modeling the structural information in a loose manner. This is done by augmenting a bag-of-features [3] with loose spatial information. We represent images as a collection of parts characterized by an appearance codeword from a visual vocabulary and by a neighborhood context, organized in an ordered set of bag-of-features representations. These bags are computed in four overlapping areas in the local coordinate system of each part. We call these representations *context distributions*. The semantic distance between images is obtained by matching parts associated with the same codeword using context distributions. The matching is polynomial in the number of parts. The average weight of the matching yields the distance between the images, which is small when they correspond to the same category. The proposed distance can be easily incorporated in different discriminative classifiers. In this work we convert it into a kernel and use it in the SVM classification. Our method is robust to translation, scale, and some degree of rotation; thus it can be applied to images with clutter and pose variations. The experiments show that our approach outperforms state-of-the-art appearance based algorithms and loose shape methods on a very challenging VOC2006 [4] data set, which contains different views of 10 categories of objects with a lot of clutter. We also show that the proposed loose shape approach performs much better than the constellation type models (such as [1]) on the dogs vs. animals test [1], despite the similar pose of the dogs in this data set, which allows learning the spatial relation between parts.

1.1. Related Work

It was noted recently that augmenting a bag-of-features representation with some spatial information might be more efficient than the existing approaches to shape modeling dis-

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

cussed above. [13, 22] incorporated pairwise relation between neighboring local features in a bag-of-features approach. A step forward in this direction is the work of Lazebnik et al. [14]. They propose to repeatedly divide an image into sub-regions and compute histograms of local features in each sub-region. The resulting spatial pyramids are matched using an adaptation of the pyramid matching scheme proposed by [9]. The method of Lazebnik et al. [14] incorporates the information about the spatial arrangement of features in a global coordinate system of the image (the sub-regions are fixed). Thus the representation is not invariant to geometrical transformation. [21] suggests incorporating appearance, shape, and context in a discriminative model based on textons that jointly capture local shape and texture. The method allows variation in pose, because the structural information is learned locally. Kushal et al. [12] propose a more flexible model that represents objects as a collection of "partial surface models" that obey loose local geometric constraints. This approach also allows variation in viewpoint.

2. Our Approach

Our goal is to construct a semantic distance between images that is small when the images come from the same category and large when they belong to different categories. Measuring the distance between bags-of-features obtained from full images doesn't work well if the images contain similar scenes with different types of objects. This is true especially when the objects are small – cars and bikes in an urban scene, for example. Dividing the image into sub-regions and matching bags-of-features from these sub-regions (as was done in [14]) partially solves the problem, but it doesn't allow much variation in pose. To handle more variation, the bags of features should be constructed locally around each part of the object. Such a representation can be viewed as context distribution. Next we discuss the details of the image representation and the matching procedure.

2.1. Image Representation

We build a visual vocabulary similar to [3]. The low-level features are represented using histograms of gradient directions [16] computed at interest points found by the Saliency detector [10]. The codebook is constructed using the K-means clustering algorithm. The codebooks are built for each class separately and then concatenated into a single codebook.

Following the bag-of-features approach, we classify low-level features into N types corresponding to the visual words in the codebook. Features sampled close to each other in the image usually correspond to the same codeword. This can be explained by similar gradients in the neighboring regions. Interest operators usually find a large number of patches, many of which overlap or are in proximity. As a result, features belonging to the same codeword tend to appear in clusters, as shown in Figure 1a. We replace these clusters with *parts*. Specifically, we group features that are associated with the same codeword and appear in spatial proximity in the image. The average location of the features within the group is set to be the coordinates of the part (Figure 1, a).

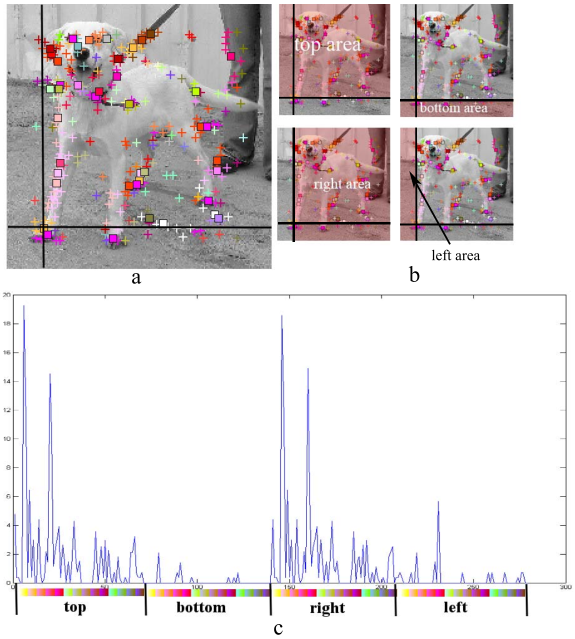

Figure 1. Construction of the context distribution for one part: a, features are shown as colored crosses and parts are shown as colored squares. The black cross lines indicate the local coordinate system of the part for which we show the construction of the context distribution; b, regions associated with the part; c, context distribution obtained by concatenation of four bags-of-features, computed in the top, bottom, right and left regions of the part. The contribution of the features to the histogram is weighted according to their distance to the origin of the local coordinate system.

We represent images as a collection of parts. Each part is characterized by the location in the image, by the appearance type corresponding to the codeword, and by the context distribution, defined as an ordered set of four bags of features constructed in the overlapping regions around the part (as shown in Section 2.2). Even though a bag-of-features is orderless representation, the context distribution captures structural information in a loose way, because it computes bags in the local neighboring regions and organizes them in a specific order.

2.2. Building Context Distributions

We place a local coordinate system at the part and divide the image around it into 4 overlapping areas: top, bottom, left and right. All the features located above the local x-axis are considered to be in the top area. All the features below the local x-axis contribute to the bottom region. The left and the right regions are set in a similar way (Figure 1). A context distribution is formed by concatenating the bags-of-features computed in the top, bottom, right, and left areas of the part and normalizing the resulting vector of size $4N$ (N is the size of the codebook) to unit sum (Figure 1).

To make the context information local, we weight the features in the histograms according to their distance to the origin of the local coordinate system. The weight is maximal inside a certain radius around the origin and decays fast outside the radius (see Section 3.1 for details). The idea behind the weighting is to assign greater weight to the features in the neighborhood that might correspond to the same object and less weight to the distant parts of the image. Since we do not know in advance the size of the object, we cannot set the radius for the weighting in advance. Thus we create image representations for a range of radiuses corresponding to different scales and apply multi-scale matching.

2.3. Calculating the Distance between Images

Consider a matrix that contains all context distributions for all parts in an image. Such a matrix will always have $4N$ columns – the length of the context distribution. The number of rows, however, will depend on the number of parts per codeword. Since images may contain a different number of parts of different types, their representation matrices will have different sizes. Thus we cannot use the Euclidian distance or any other metric for comparison.

We assume that semantically similar parts are likely to be associated with the same codeword and have similar context distributions. We propose to match parts of the same codeword using the χ^2 distance between their context distributions. For each codeword we extract the corresponding parts in both images and organize them in a bipartite graph where each side represents an image. The weights on the edges are χ^2 distances between the context distributions associated with the parts (see Figure 2). We want to match the maximum number of parts between the two images with minimum weight among all possible maximum matchings. This matching problem can be formulated as maximum bipartite matching with minimal cost and can be solved in polynomial time by applying a variation of the Hungarian algorithm [11]. Our experiments show that the number of parts per type is usually quite small (less than 6) and, with the polynomial algorithm, the matching is very fast.

The average distance between the matched parts will represent the distance between the images associated with the current codeword (Figure 2). We repeat the matching procedure for each codeword and form a vector of distances of the size of the codebook. We set the magnitude of the vector to be the semantic distance between the two images. Since not all the codewords appear in each image, the algorithm considers two special cases in the distance calculation:

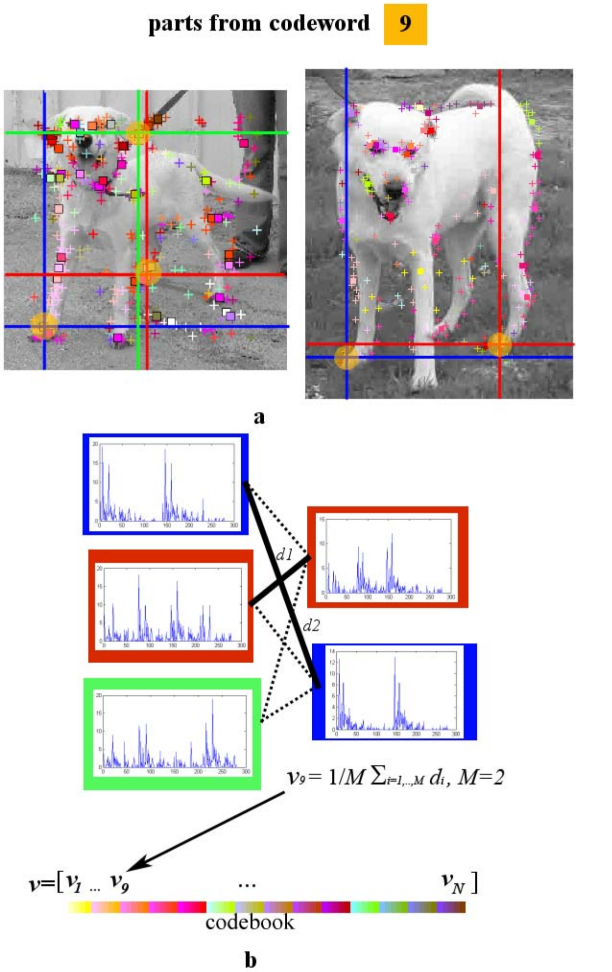
parts from codeword 9
a
d1
d2
$v_9 = 1/M \sum_{i=1,\dots,M} d_i,\ M=2$
$v=[v_1 \dots v_9$ … $v_N]$
codebook
b

Figure 2. The matching of parts corresponding to codeword 9 (shown in orange): a, three parts associated with this codeword (left-hand image), shown with blue, red, and green local coordinate systems, and two parts (right-hand image) associated with this codeword, shown with blue and red coordinate systems; b, bipartite graph with nodes corresponding to the parts from both images. The weights on the edges are χ^2 distances between the context distributions of the parts. Bold edges show the maximum match with the minimal cost. The average weight from the match represents the distance between the parts corresponding to codeword 9. Vector v contains distances for all the codewords. The magnitude of the vector measures the semantic distance between the images.

- When a certain codeword is missing in both images, the distance for this codeword is set to zero. This can happen when the codeword doesn't belong to the categories represented in the images. If the images come from the same category, the choice of zero distance is optimal. If the images belong to different categories, the choice is arbitrary.

- When only one image contains parts corresponding to a certain codeword and the other doesn't, the distance is set to a constant denoting the maximum distance.

It can be shown that the resulting distance is positive and symmetric. These properties allow it to be incorporated into various classifiers.

2.4. Classification Using Kernel SVM

The proposed image distance can be employed in different classifiers. In this paper we convert it to a kernel and apply Support Vector Machine.

Any kernel can be seen as a measure of similarity. For example, the Gaussian RBF kernel is based on the Euclidian distance. If we substitute the Euclidian distance with our new distance $D(x, y)$, we will obtain a measure of image similarity: $K_{im}(x, y) = \exp\left(-\frac{D(x,y)}{2\sigma^2}\right)$. In order to use it as a kernel in SVM, we should ensure that the similarity function is a valid kernel. Specifically, K_{im} must be symmetric positive (semi)definite.

Since $D(x, y)$ is symmetric, the matrix K_{im} is symmetric as well. Since $D(x, x) = 0$ and $D(x, y) > 0$ for $x \neq y$, thus $K_{im}(x, x) = 1$, and all off-diagonal elements of K_{im} are smaller than 1. Although these conditions are not sufficient to ensure that the matrix K_{im} is positive definite, in practice, by choosing the best σ for discrimination, we make the off-diagonal elements much smaller than 1. Thus the matrix we obtain is symmetric, diagonally dominant with diagonal elements equal to 1. These conditions imply that the matrix is positive definite [8]. We checked the kernel matrix on training sets of many different categories, and all these matrices were positive definite. Since training and test data are usually sampled from the same distribution, the sigma that we choose using training data fits the test data, and the kernel is valid.

3. Experimental Results

The following experiments show that our method, which captures structural information in a loose way, shows excellent performance on different kinds of categories, including structural objects and objects with large deformations and viewpoint variation.

3.1. Implementation Details

It is reasonable to expect that different categories will need different size codebooks. However, this requires too much tuning. Thus, in all experiments, we created codebooks with 35 codewords per class.

Spatial grouping of features into parts was done using morphological operators with seeds proportional to the size of the image. In future work we plan to implement hierarchical clustering for grouping of features.

The following weighting function was used in the construction of the context distributions:

$$f(R) = \{ \begin{array}{cc} 0.7 & d < R \\ 1/d^{1/5} & d \geqslant R, \end{array}$$

where d is the distance from the feature to the center of the local coordinate system and radius R is a parameter. The form of the weighting function was chosen empirically, although a simple step function worked almost as well. Context distributions were computed using the weighting function with four radiuses: 100, 200, 300 400 pixels.

The resulting four-level representations were matched using the implementation of the Munkres algorithm from [24]. The obtained distances were converted to the kernel shown in Section 2.4.

3.2. The VOC 2006 Dataset

We tested our method on the VOC2006 set from the PASCAL Visual Object Classes Challenge 2006 [4]. It includes images provided by Microsoft Research Cambridge and "flickr." The data contains views of bicycles, buses, cats, cars, cows, dogs, horses, motorbikes, people, and sheep, in arbitrary pose, a total of 5,304 images annotated with ten categories. All the images contain a lot of clutter. Each category is divided into training, validation, and test sets. We used the training set for codebook creation and the validation set for kernel SVM training. Using separate data for codebook construction helped reduce overfitting. The categorization tests were conducted on the test set.

The classification part of the Challenge tested twenty-one state-of-the-art visual categorization methods. The winners of the classification part of the competition (depending on the category) were different variations of bag-of-features representations that discard shape information (QMUL-HSLS [4], INRIA-NOWAK [17], XRCE [19]), and the best average performance was shown by QMUL-LSPCH[4], which applies a two-layer SVM classifier on the pyramid representations introduced in [13]. We conducted our experiments on the data used in the challenge, because it provides a benchmark on the state-of-the-art pure appearance methods and the approaches that go beyond bag-of-features (QMUL-LSPCH, [21]).

Table 1 compares the recognition performance of our approach with the winners of the challenge in each category. The results show that our method outperforms all the methods on half of the categories and is close to the best in the rest. It also achieves the best average EER.

We emphasize that our approach uses only one type of low-level features, while others combine different types of features. The codebooks used in our method are small and constructed using K-means. The codebooks proposed in the other methods ([19, 17]) are much larger and are obtained using more powerful tools. Even though our appearance

	Our method	VOC 2006 Winner
Bicycle	0.920	0.870 (QMUL-LSPCH)
Bus	0.932	0.940 (QMUL-HSLS)
Car	0.910	0.921 (INRIA-Nowak)
Cat	0.850	0.866 (QMUL-LSPCH,QMUL-HSLS)
Cow	0.875	0.863 (QMUL-LSPCH,QMUL-HSLS)
Dog	0.820	0.800 (QMUL-HSLS)
Horse	0.840	0.850 (QMUL-LSPCH)
Motorbike	0.901	0.897 (QMUL-LSPCH,QMUL-HSLS,INRIA-Nowak)
Person	0.820	0.780 (QMUL-LSPCH)
Sheep	0.865	0.882 (XRCE)
Average recognition rate	0.873	0.863(QMUL-LSPCH)

Table 1. Categorization performance, corresponding to the EER. See [4] for the description of QMUL-LSPCH, QMUL-HSLS, XRCE,and INRIA-NOWAK methods.

	Dogs vs. easy animals	Dogs vs. hard animals
Our method	0.837	0.968
Bar-Hillel at al. [1]	0.79	0.654

Table 2. Recognition rate, corresponding to the EER in dogs vs. animals experiments

model is very simple, the proposed approach shows excellent performance due to incorporation of loose shape information. In future work we plan to improve visual vocabulary, as was done in [19]; and we expect that it will further improve recognition.

3.3. Comparison to the Part-Based Shape Models

Next we test our method on more structured objects and compare the results to a constellation type of model that captures the location and scale relation between parts. The method introduced in Bar-Hillel et al. [1] learns a generative model in a discriminative way using a boosting algorithm and shows better performance than other constellation models [7].

We test our method on the database of dogs and animals presented in [1]. The dog set contains 460 images of dogs of different breeds and sizes with varying amount of background. The pose in all images is almost the same. This allows very good modeling of geometric information. We follow the experimental setup of [1]. The images of animals are divided into two sets: “hard animals” and “easy animals.” The “hard animals” set contains 460 images of animals, with 50% quadrupeds – structurally similar to dogs, such as horses, cows, bears, and elephants, and 50% other animals such as birds, rabbits, monkeys, and insects. The “easy animals” set contains 460 images of animals that are not similar to dogs. We conducted two tests (as in [1]): dogs vs. “hard animals” and dogs vs. “easy animals.” In both experiments the training set included 230 images of dogs, and 230 images of animals from the corresponding set. We used the training set for both codebook creation and kernel SVM training. The test set included the 230 remaining images of dogs and 230 images of animals from the corresponding set. Table 2 presents the EER obtained in these experiments. Even though the dogs in this data set appear in almost the same pose, our method, which allows flexibility in shape, significantly outperforms [1], which explicitly models the locations of parts.

Our results in the “hard animals” test significantly outperform [1] and are even better than the result in the “easy animals” test. This may seem unintuitive, because the animals from the “easy” set don’t look like dogs. Lower recognition performance in the “easy animals” test can be explained by overfitting due to lack of training data – up to ten images per animal.

We do not compare the performance of our method to the bag-of-features representations on Bar-Hillel’s set, because of possible differences between the implementations. However, we performed an experiment that illustrates the benefits of local bags over bags of features computed globally from the entire image. Figure 3 shows the average distances between dogs (in red) and the average distances between dogs and other animals (in blue). Our local bag approach has a much larger margin between within-class distances and between-class distances than the global approach.

In summary, the experiments show that the proposed approach outperforms the state-of-the-art pure appearance methods from ([4]), the methods that explicitly model spatial configuration of parts ([1],[7]), and intermediate models that augment bags of features with some spatial information (QMUL-LSPCH in [4], [21]).

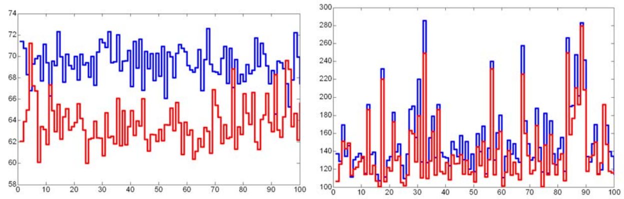

Figure 3. The red line shows the average distance between 100 dogs and the blue line shows average distance between 100 dogs and 100 other animals. The distances in the left plot were created by the proposed method, which matches local bags of features. The right plot shows distances computed between bags created from full images. Our distance is considerably better because the within-class distances are much lower than the between-class distances.

4. Conclusions

This paper presented a model that captures structural information in a loose manner by dividing the neighborhood of each part into an ordered set of regions and computing weighted bags of features in these regions. Since the local bags are organized in a specific order, the representation captures structural information. The regions are overlapping, which allows for the shifting of parts that can result from change of pose or other deformations. Our model also allows variation in scale.

We have shown that the proposed representations can be compared by matching (in polynomial time), which yields a semantic distance between images. The distance is small when the images correspond to the same category and large otherwise. The distance was converted to a kernel and used in SVM.

The experiments on objects with different levels of structural stability have demonstrated that our method showed excellent performance in all cases. The benefit of a loose shape approach can be explained by the fact that structural information exists even in very flexible objects; modeling this information in a loose way assists in recognition. Furthermore, even objects with very stable structure can still vary enough to break rigid models.

References

[1] A. Bar-Hillel, T. Hertz, and D. Weinshall. Efficient learning of relational object class models. In *ICCV*, 2005. 1, 5

[2] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondence. In *CVPR*, 2005. 1

[3] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka. Visual categorization with bags of keypoints. In *ECCV*, 2004. 1, 2

[4] M. Everingham, A. Zisserman, C. Williams, and L. VanGool. The Pascal visual object classes challenge 2006 (voc2006) results. 1, 4, 5

[5] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 28(4):594 – 611, 2006. 1

[6] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55 – 79, 2005. 1

[7] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *CVPR*, 2005. 1, 5

[8] G. H. Golub and C. F. V. Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences, 1996. 4

[9] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, 2005. 1, 2

[10] T. Kadir and J. M. Brady. Scale, saliency and image description. In *International Journal of Computer Vision*, volume 45(2), pages 83 – 105, 2001. 2

[11] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2:83–97, 1955. 3

[12] A. Kushal, C. Schmid, and J. Ponce. Flexible object models for category-level 3d object recognition. In *CVPR*, 2007. 1, 2

[13] S. Lazebnik, C. Schmid, and J. Ponce. A maximum entropy framework for part-based texture and object recognition. In *ICCV*, 2005. 1, 2, 4

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

[15] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV*, 2004. 1

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 2

[17] E. Nowak, F. Jurie, and W. Triggs. Sampling strategies for bag-of-features image classification. In *ECCV*, 2006. 1, 4

[18] A. Opelt, A. Pinz, M.Fussenegger, and P.Auer. Generic object recognition with boosting. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 28(3):416–431, 2006. 1

[19] F. Perronnin, C. Dance, G. Csurka, and M. Bressan. Adapted vocabularies for generic visual categorization. In *ECCV*, 2006. 1, 4, 5

[20] T. Serre, L. Wolf, and T. Poggio. A new biologically motivated framework for robust object recognition. In *CVPR*, 2005. 1

[21] J. Shotton, J. Winn, C.Rother, and A.Criminisi. Texton boost: Joint appearance, shape and context for multi-class object recognition and segmentation. In *ECCV*, 2006. 2, 4, 5

[22] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *ICCV*, 2005. 2

[23] S. Ullman, E. Sali, and M. Vidal-Naquet. A fragment-based approach to object representation and classification. In *Proceedings of the 4th International Workshop on Visual Form*, 2001. 1

[24] J. Weaver. http://johnweaver.zxdevelopment.com. 4