2011 Sixth International Conference on Image and Graphics

Visual Object Localization in Image Collections

Yanyun Qu， *Han Liu*
Computer Science Department, Xiamen University
Xiamen, 361005, P.R.China

***Abstract*—**The research of object localization is active in the field of visual object category. In this paper, we focus on object localization in a given special category dataset. We propose to exploit the context aware category discovery for object localization without any labeled examples. Firstly, the image is segmented based on a multiple segmentation algorithm. Secondly, these generated regions are clustered by spectral clustering method to find the category pattern based on the context of the dataset and the saliency. Thirdly, the object is localized based on the weakly supervised learning algorithm. To justify the effectiveness of the proposed method, the detection precision is employed to evaluate the performance of our approach. The experimental results demonstrate that our approach is promising in object localization with unsupervised learning method.

Keywords- *Multiple segmentation, Multiple instance learning, Object localization, Image labeling*

1. INTRODUCTION

Object localization is one of the tasks of visual object category, which aim to find the object instance in an image. The task is challenge in the real world scene, because the object may be in variety of appearance for the scale transformation, illumination, viewpoint transformation and occlusion and so on. There are many works to solve the problem. Generally, object localization can be divided into two categories: the first type of methods needs labeled images as training dataset [8,9,10,17], and the second type of methods does not need any labeled images. There are two important factors for the former: how to design a classifier, and how to select the search scheme. Dalal [10] designed an object appearance model and employed it to search the candidate regions in an image based on the sliding windows scheme. Lazebnik [16] designed a classifier combining both the spatial consistency between the object and its neighbor. Lampert [8] proposed to locate the object by an efficient subwindow search (ESS) which was based on the branch and bound algorithm. Fulkerson[1] assigned a category label to every superpixel, and trained a multi-class SVM classifier based on the SIFT descriptor for each superpixel. Qu et al [17] proposed

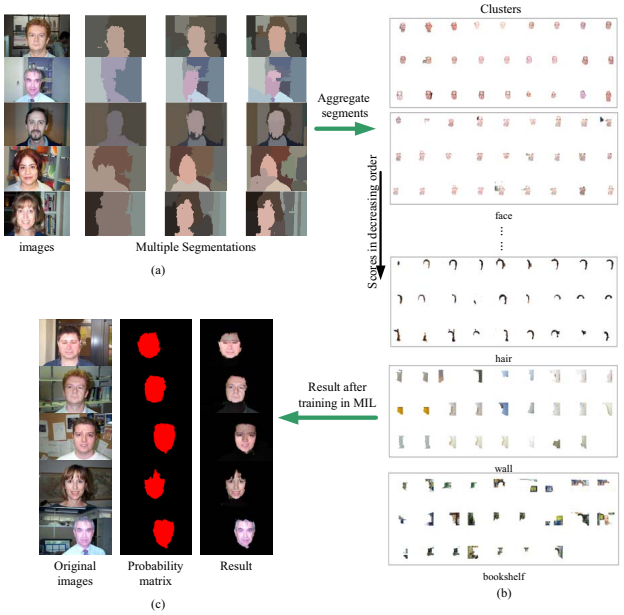


Figure 1. The framework of context aware object localization. a) Multiple segmentation, b) grouping the segments and compute the confidence value, c) object localization by MIL.

two stage object localization method. In the first stage, ESS was used to detect the candidate regions coarsely, and then in the second stage, the sliding windows searching classifier with HOG feature was used to locate the object finely. The second type of works is based on image segmentation. Tighe [6] proposed to use the scene-level matching with global image descriptors to label the image. In details, his approach found the similar superpixels among the matching images, and used the Markov Random Field optimization to incorporate the neighborhood context. Then they computed the probability of the category of each superpixel in the original image according to the amount of similar superpixels in the labeled dataset. This approach required no training data, while some dataset statistics computation was required. Galleguillos [4] proposed weakly supervised object localization, where Multiple Instance Learning (MIL) was employed to locate the object followed by multiple stable segmentations for the image. This method

978-0-7695-4541-7/11 $26.00 © 2011 IEEE
DOI 10.1109/ICIG.2011.123

required an assumption that at least one segment should contain the interesting object.

In the second type of methods, the aforementioned works only consider the context in a single image, but have not considered the context among the dataset. In this paper, we propose to exploit the context among the dataset to discover and locate the object. The framework of the proposed approach is shown in Fig 1. First, multiple segmentation algorithm is employed to segment all the images in the given dataset. Second, each segment is represented by the bag of words [19,20]. Third, the most confidence clusters are selected as the candidate object category according to their saliency, tightness, and volume. Fourth, the confidence value for every segment is calculated which results in the positive instance or negative instance. At last, MIL [18] is employed to locate the object. The main contribution of the proposed work is to exploit the context of the dataset and the saliency of the object.

The rest of the paper is organized as follows. In Section 2, we introduce the Multiple Instance Learning. In Section 3 we detail the implementation of the proposed approach. In Section 4, experimental results are given to justify the effectiveness of the proposed method. The conclusions are given in Section 5.

2. MULTIPLE INSTANCE LEARNING

Multiple Instance Learning (MIL) have been employed for training object classifiers with weakly supervised data [7,13,18,21]. MIL trains a discriminative binary classifier predicting the class of the sample, under the assumption that each positive training data set contains at least one true-positive instance, while negative training data set contain none. Let $\{(X_1, y_1),(X_2, y_2),\cdots,(X_n, y_n)\}$ denote the training data, where $X_i = \{x_1, x_2, \cdots, x_{im}\}$ is a bag and $y_i \in \{0,1\}$ is a bag label. The bag label is defined as $y_i = \max_j(y_{ij})$, where y_{ij} is the instance label, which is unknown during training. In this paper, we use the MIL based on SVM [18], which is named MI-SVM. The main problem is to how to define the margin so that the SVM can be adapted to the MIL problem. As the negative bags have all negative samples, so the margin is defined as the regular case for them. For positive instances, MI-SVM defines the margin of a bag as the maximum distance between the hyperplane and all of its instances. Therefore the problem can be formulated:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i,j}\xi_{ij},$$

$$\text{s.t.} \quad w \bullet x + b \le -1, \forall y = 0,$$

$$\max_j(w \bullet x + b > 1 - \xi_{ij}), \forall y = 1$$

We follow Andrews' work [18] to solve the optimization problem, which adopts a simple heuristic algorithm to find the solution.

3. IMPLEMENTATION DETAILS

Given a special category dataset, e.g. a Face dataset, where each image contains at least one face, two assumptions are made based on the observations: 1) the object pattern is the shared property of the special category dataset; 2) the object pattern is the focus of photographer, and the object is usually put in the salience location of an image. Take the Face dataset as an example, the shared property is that every image in the Face dataset has at least one face pattern, and the human face is usually put in the salience position of an image.

3.1. Segment Representation

In the first stage, the Normalized Cut algorithm [3] is employed to segment an image at multiple times. In our experiment, an image is segmented at least three times according to different numbers of segments. Take a face image as an example, we obtain 3, 8, 13 segments corresponding to the three level segmentations. The total segments obtained from an image amount to between 20 and 40 in general.

In the second stage, we represent the segments based on the bag-of-words model. In each segment, we use dense SIFT to describe each point, which is described with four different scales whose radiuses are 16,24,32,40 in our experiment, and we obtain four SIFT features at each point. After that, we cluster the local features by K-means method, and the clustering centers are regarded as the visual words, and then we quantize the local features and represent the segment by a normalized histogram of visual words. The flow chart of segment representation is shown in Fig 2. Because the number of the visual words affects the representation performance of the segment, and the more visual words make the quantization error smaller, we consider the tradeoff between the computation complexity and the quantized performance and use a

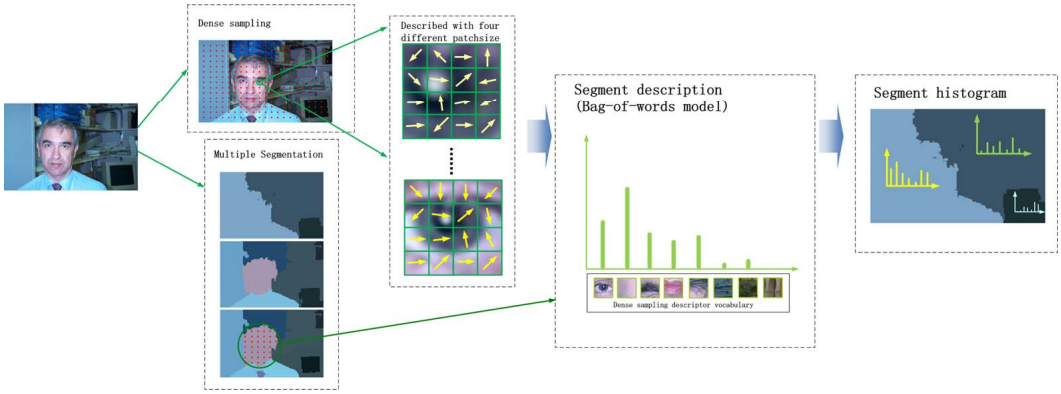


Figure 2. The flow chart of segment representation

visual dictionary with 600 visual words to quantize the SIFT vectors in our experiments.

3.2. Spectral clustering for segment mining

In the third stage, we mine the segments by spectral clustering to discover the object category. We build a weighted directed graph $G=(V,E)$, where the nodes of the graph are the segments, and the weight on each edge $w(i,j)$ is the function of the similarity between node i and node j. In our experiment, we define the weight as the Euclidean distance,

$$w_{ij}=\left\|v(i)-v(j)\right\|_2$$

where $v(i)$ is the histogram of visual words for the i th segment. We transform the clustering problem to graph cut problem. And we adopt the recursive two-way partition, as shown in Fig 3. The root node is the set of all segments, and it is partitioned into two parts by Graph cut method [22], and each part is the node of the binary tree. And then the part where the amount of segments is larger than the threshold is recursively partitioned until the amount of the leaf node is smaller than the threshold. The leaf nodes are regard as the clustering group.

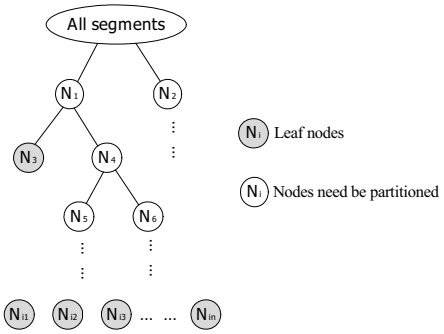


Figure 3. Two-way partition for segments

3.3. Computing the confidence value for each cluster

The main difficulty is how to discover the object category in the clustering groups. In order to solve the problem, we design a measurement method to compute the confidence value of the clustering group. We consider four factors of the clustering group: the volume of the cluster set, the saliency of the object appearance, the spatial saliency of the image, and the tightness of the cluster, which are detailed in the following.

1) Volume We define the volume of the clustering group as the number of segments contained in the group. We denote the volume of the i th clustering group c_i by $T(c_i)$. As we supposed, the object category has the common target pattern, and every image contain at least one object, so the object occur more frequently in the object category, and the segments of the object are more aggregated together. Furthermore, the object group contains more segments, and the volume of the object group is larger.

2) Appearance Saliency Although the foreground object in a single image may have little saliency, in a set of the objects it usually has high saliency. Considering the assumption that the object pattern is shared in the object category set, we think it rational that a clustering group with large volume and high saliency belongs to the foreground. So we assign the clustering group with a high weight value. In order to compute the saliency of an image, we consider the color attribute of the image in our experiments. Followed Anchanta's work [15], the image is first filtered by Gaussian filter, and the blur image I_G is obtained. Then the mean of the filtered image, denoted by I_μ, is computed. The saliency map is computed as $H(x,y)=\left|I_\mu-I_G\right|$, and all the computation is

processed in LAB color space. $H(c_i)$ denotes the saliency value of the cluster c_i. The appearance saliency value of a segment is the average of the saliency value for each pixel of the segment. And the appearance saliency value of a cluster set is the average of the saliency values of all the segments in the cluster.

3) Spatial Saliency In our experience, when we take photo of an object, we usually put it in the spatial salience location of the photo. Therefore, the object is usually in the salience location of an image. According to the assumption, we design a Gaussian function to compute the spatial saliency for each pixel of the image. The center of an image is put the largest weight and the weights of other pixels are computed according to the Gaussian function. We compute the weight of a segment and the weight of the clustering center respectively. The weight of a segment is the amount of the weights of the pixels in a segment and the weight of the i th clustering group $G(c_i)$ is the average of all the weights of the segments in the i th cluster group.

4) Tightness As we supposed before, the object category has the common property. Therefore, we think the clustering group contained the object pattern has smaller divergence. We use the deviation of the clustering group as the measurement,

$$D(c_i) = \sum_{j=1}^{K} \sigma_i(w_j) \tag{1}$$

where $\sigma_i(\cdot)$ denotes the standard deviation of cluster i, and $K = 600$ which is the number of the words contained in the visual dictionary.

We also normalize the four values corresponding to a clustering group respectively. For the volume value, we compute the probability of the i th clustering group c_i as:

$$Pr(c_i \mid T) = \frac{\sum_{k=1}^{i} T(c_k)}{N_T} \tag{2}$$

where N_T is the amount of all the volumes of the T clustering volume. The same normalized operation is done on the values of appearance saliency, spatial saliency and the tightness, and we denote them by $Pr(c_i \mid H)$, $Pr(c_i \mid G)$ and $Pr(c_i \mid D)$.

Now we can compute the confidence value for each cluster set by integrate the four probability values as:

$$\begin{aligned} S(c_i) = {} & \mu\, Pr(c_i \mid T) + \omega\, Pr(c_i \mid H) \\ & + \eta\, Pr(c_i \mid G) + \varphi\, Pr(c_i \mid D) \end{aligned} \tag{3}$$

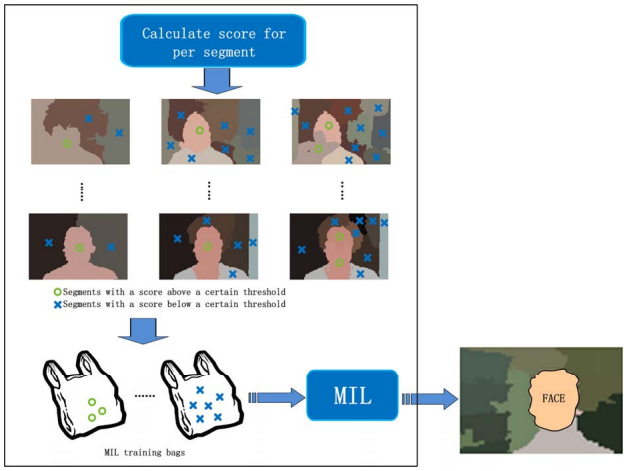


Figure 4. Weakly supervised localization

where $\mu, \omega, \eta, \varphi$ are the weights of confidence value. We choose $\mu = 0.4, \omega = 0.2, \eta = 0.2, \varphi = 0.2$ in our experiment. After that, the confidence values of the cluster sets are further sorted in increasing order $(c_1, c_2, \cdots, c_n)$. The confidence value of each segment which measures the probability of the object could get from its corresponding cluster.

3.4. Weakly supervised localization

The confidence value of a clustering group implies the probability of the object category, and the confidence value of a segment implies the probability that the segment belongs to the object pattern. In an image, the segments whose confidence values are below the certain threshold are labeled as the negative instances which form the negative bag. The remaining instances beyond the threshold could contain the object. Note that there may be not any segments beyond the threshold in an image; in this case we select the segments whose confidence values are the first three highest scores as the candidate positive instances and the others as the negative instances. Each instance is represented by the histogram of visual words. The MIL classifier is trained on the positive bags and negative bags. The weakly supervised localization algorithm is shown in Fig 4. In testing stage, we compute the confidence value for every segment of an image. Considering a pixel may belong to different segments, we compute all the confidence values of the segments which contain the same pixel, and average them to get the probability of the pixel, and we obtain the probability matrix as shown in the middle column of Fig 5. We multiply the probability matrix and the intensity of the image, and obtain the object location, as shown in the last column of Fig 5.

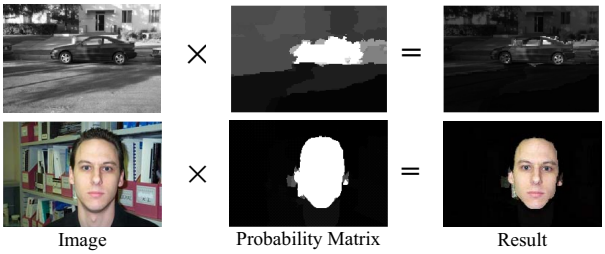


Figure 5. Using the probability matrix to locate the object.

4. EXPERIMENTAL RESULTS

In this section, we implement our approach on the four object categories from the Caltech101 database--faces (435), carsides (123), and motorbikes (798), airplanes (800). Each class set is split into two separate sets of images. One set is for training and the other set is for testing.

In order to evaluate the performance of our approach, we use the detection precision as the evaluation criterion, which is defined as $P = \frac{N_c}{N}$, where N_c is the number of the correct labeled regions and N is the total number of regions with the ground truth labels. The correct labeled region satisfies $\frac{Area(R_d \bigcap R_{GT})}{Area(R_d \bigcup R_{GT})} \geq 0.5$, where R_d is the detected region, and R_{GT} is the ground truth region.

We first show the results of our approach in object localization, as shown in Fig 6. In each group there are four images, and the upper is the original image. The middle left is the probability matrix under the MIL classification. In this figure, the lighter the pixel is, the more confident the region belongs to the object. The middle right is the result generated by the proposed method. The ground truth label is shown in the last row and signed by the red contour curve. These results show that our localization results are close to the ground truth region.

Now we evaluate the performance of our approach in term of the detection precision. We compare four methods of object localization: our method, ILIL[17], ESS[8], HOG which is a sliding window searching method with HOG classifier. Our method is an unsupervised learning method which does not need the labeled training data. And the other three methods are supervised methods which need the well labeled training data. The comparison results are shown in Table 1. We find that our approach's precision is close to the prior work in the present four category datasets, though it is an unsupervised learning method.

Table 1. Comparison of the detection precision.

	Face	Motorbike	Airplane	Carside
Data size Train/Test	218 / 217	335 / 463	336 / 464	50 / 73
Ours	.963	.883	.804	.781
ILIL	.867	.932	.770	.879
ESS	.319	.914	.215	.208
HOG	.743	.790	.743	.939

5. CONCLUSIONS

In this paper, we propose to localize the object in a given category dataset by exploiting the context of the dataset. We also design a measurement to compute the confidence value for a segment, which implies the positive instance or the negative instance. In this method, the object is located based on the MIL algorithm. The experimental results and comparison with existing methods demonstrate the effectiveness of the proposed approach. In our future works, we will further leverage this automatic localization for related applications, such as object retrieval or recognition.

ACKNOWLEDGMENT

The research work was supported by the National Basic Research Program of China under Grant No. 2007CB311005, the Fundamental Research Funds for the Central Universities under Grant No.2010121067, and National Defense Basic Scientific Research program of China under Grant No.B1420110155.

REFERENCES

[1] B. Fulkerson, A. Vedaldi, S. Soatto, "Class Segmentation and Object Localization with Superpixel Neighborhoods," In Proc ICCV, 2009.

[2] B. Russell, A. Efros, J. Sivic, W. Freeman, A. Zisserman, "Using Multiple Segmentations to Discover Objects and their Extent in Image Collections," In Proc CVPR, 2006.

[3] J. Shi, J. Malik, "Normalized Cuts and Image Segmentation," *CVPR*, 1997, pp. 731–743.

[4] C. Galleguillos, B. Babenko, A. Rabinovich, S. Belongie, "Weakly Supervised Object Localization with Stable Segmentations," In Proc ECCV, 2008.

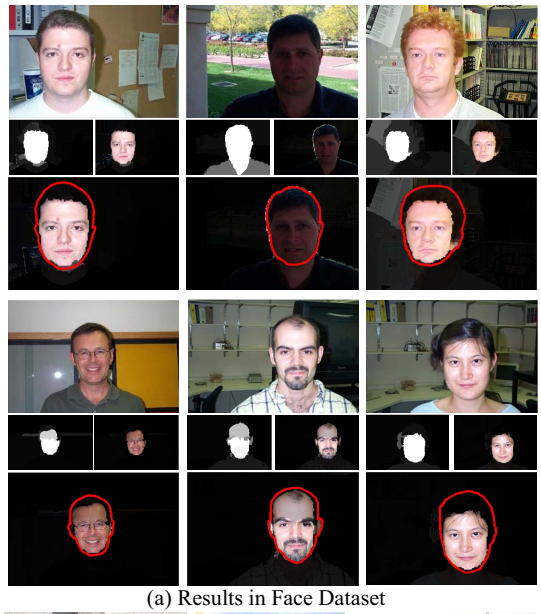
[5] D. Lowe, "Object recognition from local scale-invariant features," In Proc ICCV, 1999.

[6] J. Tighe, S. Lazebnik, "SuperParsing: Scralable Nonparametric Image Parsing with SuperPixels," In Proc ECCV, 2010.

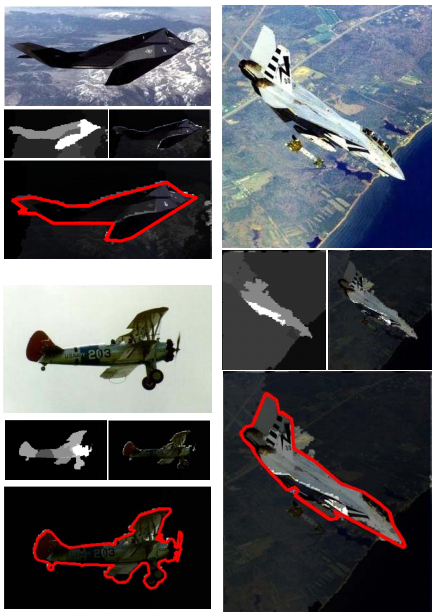
[7] J. Wang, J.D. Zucker, "Solving the multiple-instance problem: a lazy learning approach," 17th International Conference on Machine Learning, 2000, pp. 1119-1125.

[8] C.H. Lampert, M.B. Blaschko. T. Hofmann, "Efficient Subwindow Search: A Branch and Bound Framework

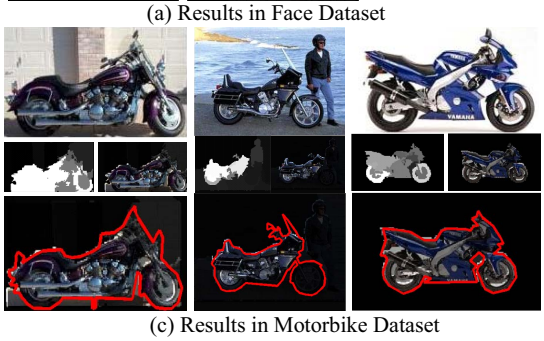
for Object Localization," IEEE Pattern Analysis and Machine Learning 31(12), 2009, pp. 2129–2142.
[9] M. Blaschko, C. Lampert, "Learning to localize objects with structured output regression," In Proc ECCV, 2008.
[10] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," In Proc CVPR, 2005.
[11] T. Ojala, M. Pietikäinen, T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with Local Binary Patterns," IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 2002, pp. 971-987.
[12] T. Ojala, M. Pietikäinen, T. Mäenpää, "A generalized Local Binary Pattern operator for multiresolution gray scale and rotation invariant texture classification," In Proc Second International Conference on Advances in Pattern Recognition, Rio de Janeiro, Brazil, 2001.
[13] O. Maron, T. Lozano Pérez, "A framework for multiple-instance learning," In Proc of the 1997 Conference on Advances in Neural Information Processing Systems 10, 1998, pp. 570-576.
[14] P. Carbonetto, N. de Freitas, K. Barnard, "A Statistical Model for General Contextual Object Recognition," In Proc ECCV, 2004.
[15] R. Achanta, S. Hemami, F. Estrada, S. Susstrun, "Frequency-tuned Salient Region Detection," In Proc CVPR, 2009.
[16] S. Lazebnik, C. Schmid, J. Ponce, "Beyond bag of features: Spatial pyramid matching for recognizing natural scene categories," In Proc CVPR, 2006.
[17] Y. Qu, C. Chen, D. Wu, Y. Xie, "Image labeling via incremental model learning," In Proc ICIP, 2010.
[18] S. Andrews, I. Tsochantaridis, T. Hofmann, "Support vector machines for multiple-instance learning," In Neural Information Processing Systems, 2003.
[19] Feifei Li, Perona Perona, "A Bayesian Heirarcical Model for Learning Natural Scene Categories," In Proc CVPR, 2005.
[20] J. Sivic, B. Russell, A. Efros, A. Zisserman, "Freeman W. Discovering object categories in image collections," In Proc ICCV, 2005.
[21] Q. Zhang, S.A. Goldman, "EM-DD: An improved multiple-instance learning technique," In Neural Information Processing Systems 14, 2001.
[22] J. Shi, J. Malik, "Normalized Cuts and Image Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000.



(a) Results in Face Dataset



(b) Results in Airplane Dataset



(c) Results in Motorbike Dataset

(d) Results in Carsides Dataset

Figure 6. Exemplar localization results in Caltech 101