To appear, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), CA, June 2010.

Visual Recognition and Detection Under Bounded Computational Resources

Sudheendra Vijayanarasimhan* and Ashish Kapoor

University of Texas at Austin, Microsoft Research Redmond

svnaras@cs.utexas.edu, akapoor@microsoft.com

Abstract

Visual recognition and detection are computationally intensive tasks and current research efforts primarily focus on solving them without considering the computational capability of the devices they run on. In this paper we explore the challenge of deriving methods that consider constraints on computation, appropriately schedule the next best computation to perform and finally have the capability of producing reasonable results at any time when a solution is required. We specifically derive an approach for the task of object category localization and classification in cluttered, natural scenes that can not only produce anytime results but also utilize the principle of value-of-information in order to provide the most recognition bang for the computational buck. Experiments on two standard object detection challenges show that the proposed framework can triage computation effectively and attain state-of-the-art results when allowed to run till completion. Additionally, the real benefit of the proposed framework is highlighted in the experiments where we demonstrate that the method can provide reasonable recognition results even if the procedure needs to terminate before completion.

1. Introduction

Visual recognition and detection are computational intensive tasks and traditionally the focus in computer vision has been on solving the problems regardless of the available computational resources. However, with recent emergence of ubiquitous computing devices such as mobile phones, laptops and netbooks, the available computational power indeed becomes a primary consideration in building systems that work in real-time. Instead of building separate algorithms that would work under different computational resources, we propose to derive recognition methods that consider such constraints on computation and appropriately triage future actions in order to provide the most recognition power for the available computational resources. Additionally, we also seek to explore if we can build methods that can provide *anytime* solutions, that is, provide a reasonable hypothesis even if they are stopped before completion.

In this paper, we particularly focus on a novel object classification/detection scenario that considers the computational resources and can scale across different computational platforms by making efficient use of resources. The problem of accurately localizing instances of a category in a novel image is a computationally intensive task because it requires matching complex object models to the observations in the image. There has been much recent work on efficient object localization and detection ([3, 20, 21, 12, 18, 27, 5, 7]). While such methods have provided state-of-the-art results on several tasks, we note these algorithms either need to run to completion to provide an answer or they cannot provide a useful solution. Consequently, in scenarios when there are bounded computational resources such *passive* methods will not degrade gracefully.

To address this challenge, we propose a method that reasons about available computational resources and triages appropriate actions to take such that best possible recognition can be performed under a specific computational budget. In particular, the proposed framework collects evidence by determining 1) the best image regions to look at and 2) the best features that should be extracted from that region. The determination of these regions and the features are guided by the principle of *value-of-information* (VOI), where the algorithm selects regions and features that promise the most evidence at the least cost. Thus, our approach *actively* selects both the feature location and feature type, unlike standard object recognition methods which are passive in the feature acquisition process.

We note that our approach is closely related to *Anytime* algorithms, a class of algorithms that degrade gracefully with the available time. While such anytime methods are gaining interest in the machine learning community, no such object recognition or detection method currently exists.

There are several contributions we make while solving the problem. First, unlike standard object detection methods, our approach is able to return a partial answer, whose quality depends on the amount of computation allotted for the algorithm. Second, we propose a novel *grid* model that

*Work done during internship at MSR, Redmond divides an object into a grid of parts and enables computation of VOI for image regions as well as different kinds of local-features. This computation of VOI can be used to determine not only the best regions in image to analyze, but also the best local features to extract from that region. Third, we derive an efficient and robust localization method that given the current evidence uses a Hough transform based voting scheme in conjunction with mean-shift clustering to estimate the modes in the hypothesis space.

We experimentally demonstrate that the proposed approach makes efficient use of computational resources by actively seeking the next computation to perform by considering trade-offs between the computational cost and the predicted gain in evidence. When compared against a passive baseline the active approach is able to detect objects with far fewer features and less computation time on two challenging object detection datasets (ETHZ shape [13] and the INRIA horses [11] dataset).

2. Related Work

Several object detection approaches have been proposed in the literature including sliding window based methods, voting schemes based on the Hough transform and biologically inspired models. Sliding window based methods are well suited for rigid objects and have been extremely successful at detecting categories such as faces [27], pedestrians [5, 23], cars [25] and a range of other objects [7]. However, for real-time detection problems, the daunting computational burden of the exhaustive search means that such methods are restricted to using primitive features which can be evaluated in constant time. While typically coarse to fine techniques such as cascades [27] are used to speed up detection, such techniques provide no guarantees.

Hough based voting techniques [20, 21, 24, 12] avoid the exhaustive search by instead searching the hypothesis space of votes from local features. However, they ignore the computational cost of extracting local features and could potentially concentrate computational resources on uninformative regions in an image. While a number of interest point and saliency operators [22, 17] have also been proposed to target informative image regions, such techniques are independent of the object recognition model. On the other hand, our approach constantly targets feature extraction on regions that are informative to the current set of hypotheses.

Recently, active vision based approaches have been proposed in the literature [15, 3, 1]. Inspired by biological vision, [15, 3] use a foveal camera to provide high resolution images near the fovea and actively schedule eye fixations while [1] provide theoretical results for actively selecting a finite set of optimal viewpoints. Our approach requires no such special hardware and instead relies on feature based evidence to schedule feature extraction.

3. Approach

The main goal of this approach is to enable efficient recognition and detection under bounded computational resources. The primary operations that are repeated in a recognition task consists of first selecting location (in scale-space) and then extracting features to provide evidence for the presence of an object. Once enough evidence has been collected, the location hypotheses can be obtained.

Simultaneous object classification and detection determines both the category an object belongs to (O) and its extent in an image x, typically represented by the four corners of a bounding box. We call (O, x) an object hypothesis during detection. We assume that the detection algorithm has at its disposal a set of procedures to extract many different kinds of features, each potentially taking different amounts of time. During detection, our algorithm sequences the next computational task by determining 1) what location and scale to focus on and 2) what specific descriptors need to be extracted. Further, the procedure also updates the object type and location hypothesis in an online manner, hence can provide results even if it needs to be terminated prematurely.

In the following, we first introduce a novel grid based object detection and classification model (Section 3.1) for both scoring and obtaining an object hypothesis. Then, we show how to estimate the parameters of the model using a set of training images in Section 3.1.2. Finally, we develop an active selection strategy for the model based on VOI which aims at maximizing the scores of possible detections at every time step during detection (Section 3.2).

3.1. Grid Model

Lets consider a test image I where we are interested in finding the parameters x that describe the extent (bounding box) of an object class O. Inspired by several successful part-based models ([20, 8, 9, 21, 19, 5]), we divide the object into parts by constructing a grid of size $N = m * n$ as shown in Figure 1. Each grid represents the extent of a semantic part of an object similar to how a bounding box represents the extent of an object. For example, the top left grid in a Giraffe captures the extent of the part “head” as long as the overall pose of the object does not change.

However, as seen in Figure 1, not all grids belong to the object of interest (some grids do not contain any part of the Giraffe). Even among grids that contain some part of a Giraffe, certain parts like the head or body might be more discriminative / detectable than others like the legs. Therefore, we associate a weight parameter w_i with each grid-part g_i that reflects how important a part is for each object for the purpose of detection. Note that g_i appears at a fixed location with respect to the object's extent x. The appearance of the part is then modeled based on evidence from local features collected within the part.

The motivation for this model is directly tied to the main goal of our approach which is to be able to compute the VOI of image regions during detection. Since each grid-

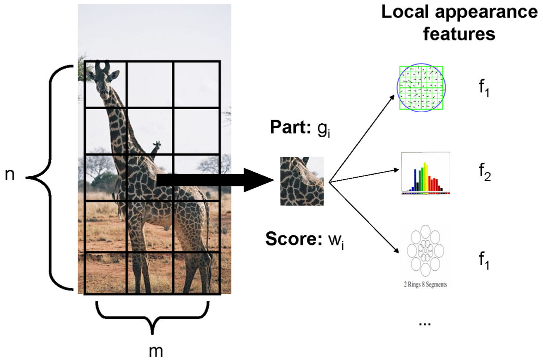


Figure 1. Our grid based model.

part always appears at a particular location with respect to the object's extent x, the weight w_i of the part provides a measure of the importance of analyzing the corresponding image region. Additionally, grid based approaches have proven successful in various recognition and detection tasks ([19, 5, 26]) as they capture the spatial relationship among the different parts effectively and provide flexibility in the actual appearance of each grid.

Formally, we obtain the score of the hypothesis (O, x) as the sum of the scores of each part ($g_i^{(O,x)}$) weighted by the probability that the part is present at the given location, based on the available features. Let F_I denote the set of features extracted from I, then we can write the score as:

$$S(O, x|F_I) \quad = \quad \sum_{i=1}^{N} w_i \;\; p(g_i^{(O,x)}|F_I) \qquad (1)$$

Here w_i is the importance weight of the part with respect to the hypothesized location x of the object O. The term $p(g_i^{(O,x)}|F_I)$ measures the evidence provided by the set of features F_I for the grid-part. It can be interpreted as the probability that the part is present as evidenced by F_I, the set of features. This term is modeled as a maximum over the probability of the part given each individual feature in F_I: $p(g_i^{(O,x)}|F_I) = \max_{f \in F_I} \; p(g_i^{(O,x)}|f, l)$, where f is a feature in the set F_I and l denotes its location.

We use the max function instead of other operators (e.g. average) as we expect every part to be defined best by a single feature type (texture for Giraffe's body, shape for the head). Further, an average might aggravate any ambiguity that already exists in the detection. Note that the two terms in the equation to compute the score consider (1) how useful a particular location is for detection and (2) the feature that provides the best evidence in detecting that part.

We estimate $p(g_i^{(O,x)}|f, l)$ using a simple nearest neighbor interpolation technique. In particular, we consider a database of features F_O for every object O that is constructed from a set of training images. The particular feature f is first matched to F_O in order to recover the set of nearest neighbors (denoted as $N(f) \in F_O$). The required conditional probability is then modeled as the weighted sum of the probabilities of its nearest neighbors:

$$p(g_i^{(O,x)}|f, l) \quad = \sum_{h \in N(f)} q_i^h \; p(h|f) \qquad (2)$$

where h is a feature in the database F_O and $q_i^h = p(g_i^{(O,x)}|h, l)$ refers to the conditional probability of part presence given the features. This term is a model parameter that needs to be estimated from the training data for every feature h and every grid part g_i. And,

$$p(h|f) = \begin{cases} \frac{1}{Z} e^{-\frac{||h-f||_2}{\gamma}} & \text{if } ||h - f||_2 < \epsilon, \\ 0 \quad \text{otherwise}, \end{cases} \qquad (3)$$

Note that we have replaced $p(h|f, l)$ with $p(h|f)$ since we match features independent of their locations.

We adopt this technique because with a large enough database the proposed method can approximate the actual conditional probability closely. Further, the whole operation can be performed fairly efficiently since we can do fast nearest neighbor lookup using approximate nearest neighbor techniques such as approximate KD-tree [2], LSH [4] and RBV [14]. We use the approach of [14] for this step as it provides significant memory advantages over the others.

Till now, we have defined the score of a detection hypothesis (O, x) based on features F_I using the parameters w_i, q_i^h for all the grid parts g_i and features h in the database. In the following, we first show how to obtain the detection hypothesis (O, x) using features F_I and then show how the rest of the model parameters can be estimated from a set of training images.

3.1.1 Determining Detection Hypotheses using Hough Voting

A clear advantage of part-based approaches over sliding window based methods is that by parameterizing the object hypothesis and allowing each local part to vote for a point in hypothesis space we can obtain globally consistent hypothesis as modes in the voting space in an efficient manner.

Hence, given a set of features F_I, we obtain object hypotheses (O, x) by matching the features in the training database to cast votes, similar to the approach of [20]. However, unlike [20] in our approach each feature casts votes for parts of objects. Further, since each grid-part appears at a fixed location with respect to the object, the features also vote for the object's extent indirectly.

The voting space is parameterized by the co-ordinates of the four corners of the object bounding box and we store these values with respect to the position of every feature in the training image using ground truth bounding box information. Thus, given a feature $f \in F_I$, we obtain its nearest neighbors, $h \in N(f)$, from the training database, F_O, and cast a vote for the corresponding x with a confidence

$q_i^h p(h|f)$ from Equation 2. We then perform Mean-Shift clustering over the hypothesis space with a bandwidth parameterized by the height and width of the bounding boxes to obtain a set of globally consistent object hypotheses.

The above search procedure can be interpreted as a Parzen window probability density estimation for the object's parameters. Finally, the score of a hypothesis (O, x) is obtained as given in Section 3.1, using all the features whose votes converged to (O, x) during mean-shift.

3.1.2 Estimating Model Parameters

We now describe how to estimate the model parameters (w_i, q_i^h), given a set of training images with ground truth bounding box information. The term q_i^h, where h is a feature in the training database F_O, can be interpreted as the probability that part i of object O is present inside the bounding box parameterized by x, given that feature h occurs at location l. Assuming that the probability $q_i^h = p(g_i^{O,x}|h, l)$ is zero whenever l is outside the grid g_i and is independent of the location l otherwise, we use the following simple way of estimating this quantity:

$$q_i^h = p(g_i^{(O,x)}|h) \propto \frac{p(h|g_i^{(O,x)})}{p(h)} \tag{4}$$

In other words, we count the number of times h occurs within the ith grid of object O and divide it by the total number of times h occurs in the training database. However, each feature occurs only once in the training database and technically this would provide a single count. Hence, we assume that the probabilities are smooth over the manifold of features and use a small number of nearest neighbors of h while performing the counts.

Once we have estimated q_i^h for all the features h in the training database we run our detection algorithm summarized in Figure 2 using uniform weights ($w_i = 1, \forall i$) on all the training images (positive and negative) to obtain a set of hypotheses. We then select a small number of high-scoring negative hypotheses along with all the positive ones and learn the weights w_i in a max-margin discriminative framework similar to the approach of [21]. This represents the training phase for our approach.

3.2. Active Selection

We assume that there are M types of features that can be extracted from the image and a feature f takes C_f units of time to obtain. If we start with a small sample of features F_I, which produce a set of candidate hypothesis, $H = \{(O1, x1), (O2, x2)...\}$, then, at every iteration, our active strategy chooses a feature type t and a location $l = (x, y, s)$, following which we extract the feature and add it to the feature pool. We can then update our candidate hypothesis set based on the newly added feature. This process is repeated until either all features have been exhausted or the allotted time for detection has been exceeded.

To this end we define an active selection criterion for image regions as well as the different kinds of local features available, based on the decision-theoretic principles of the value of information (VOI) [16]. In particular, we measure the VOI of a feature f of type t and an image location l as the ratio of the predicted gain in the scores of the current object hypotheses due to the feature to the computational cost of the feature. Our active selection function thus aims to greedily increase the score of every candidate hypothesis before the allotted time is used up. Formally,

$$VOI(f, l) \quad = \quad \frac{\Delta S(O, x|f, l)}{C_f} \tag{5}$$

Note that another possible selection criterion is the difference between the gain and the cost, however, this requires the cost and the gain to be in the same currency. Instead we define the VOI as the ratio, which intuitively seeks to maximize the gains per unit cost. The numerator of the above equation represents the improvement in $S(O, x)$ that we expect once we obtain feature f at location l, with $S(O, x)$ as defined in Equation 1. Or, in other words,

$$\Delta S(O, x|f, l) \quad = \quad S(O, x|F \cup (f, l)) - S(O, x|F)$$

Computing the VOI of all the image regions is non-trivial as we have very little information on most of the image regions. Furthermore, computing the VOI for a single location l is problematic because of the large variations in the positions, scales of a category's instance across different images. However, note that in our grid model features affect the score of a hypothesis based on the evidence they provide for a grid part (Equation 1). Therefore, we instead consider only the locations within the different parts (grid) of the current set of object hypotheses; i.e. l corresponds to some part g_i of hypothesis (O, x). Substituting the expression for $S(O, x|F)$ from Equation 1 and denoting $m_i^F = \max_{f \in F} p(g_i^{(O,x)}|f, l)$, we obtain,

$$\begin{aligned} \Delta S(O, x|f, g_i^{(O,x)}) \quad &= \quad w_i \ \left(m_i^{F \cup f} - m_i^F \right) \\ &= \quad w_i \ \max\left(0, \left(p(g_i^{(O,x)}|f, l) - m_i^F\right)\right), \end{aligned}$$

We obtain the above equation by noting that when feature f occurs inside the grid $g_i^{(O,x)}$ it can provide evidence for only the grid-part $g_i^{(O,x)}$. Therefore, the only two terms in Equation 1 affected by f are w_i and $p(g_i^{(O,x)}|F)$. The second term is the maximum over all the features F and therefore it takes the value of the new feature f if it is larger than the current value and the current value otherwise.

Interestingly, the above equation contains two terms, where one (w_i) depends purely on the grid "location" i while the other depends on the "feature" f. Since w_i has already been obtained using the max-margin framework described in Section 3.1.2 the only unknown term is

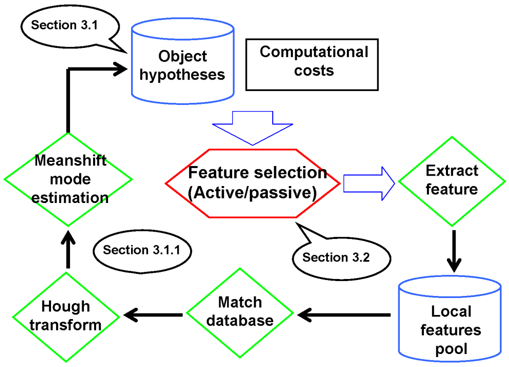


Figure 2. A summary of our algorithm.

$p(g_i^{(O,x)}|f)$. This term depends on the feature f which is yet to be extracted.

However, since we are only trying to determine the type of feature to extract, we instead estimate the expected value of the term $p(g_i^{(O,x)}|f)$ for every feature type t. We do this by considering all the features in the training database that are of type t and obtain the average value of the term. The feature type with the largest value can be interpreted as the one that is expected to provide the best evidence for object part g_i. For example, for the "body" of a giraffe, texture-based features could provide the best evidence and therefore, could have a higher value for $p(g_i^{(O,x)}|f)$.

We can now calculate the VOI of all the grid locations of the current set of hypotheses for every feature type using Equation 5. The location and feature type that is expected to provide the biggest gain in the object scores at the lowest cost is the one with the largest VOI.

Once we obtain the best grid location g_i^* and feature type t^* we sample a small number of locations l within the grid represented by g_i to extract features. The final issue is the scale at which features need to be extracted. This is simply obtained based on the ratio of the height of the bounding box represented by (O, x) to the normalized height of the ground truth bounding boxes in the training examples.

Summary. Figure 2 shows a summary of our approach for detecting and scoring object hypotheses on a test image with active/passive feature selection. The training stage for our approach consists of estimating the model parameters (w_i, q_i^h) (Section 3.1.2). The terms q_i^h are estimated by counting the number of times feature h occurs inside grid i. The weights on the grid parts w_i are then obtained using a maximum-margin classifier for which the positive and negative training examples are obtained by running our detection algorithm given in Figure 2 with uniform weights.

4. Results

We conducted experiments to 1) evaluate the strength of our grid based detection model compared to other techniques, 2) show the advantages of the proposed active selection scheme against the passive techniques, and 3) demonstrate that our active selection strategy makes efficient use of computational resources.

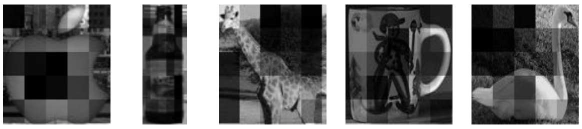


Figure 3. The grid weights learnt for each category in the ETHZ shape dataset.

Datasets: We use two challenging object detection datasets namely, the ETHZ shape dataset and the INRIA horses dataset in order to compare against several state-of-the-art hough based detection approaches [21, 24, 11, 10]. The ETHZ shape dataset contains 255 images for five shape-based classes (applelogos, bottles, giraffes, mugs and swans). The INRIA horses dataset contains 170 images with one or more side-views of horses and 170 images without the category. In both the datasets, objects occur in highly cluttered natural scenes with large variations in both scale and appearance, and sometimes contain multiple objects per image. We use the same training and testing setup used by [10] on both datasets for fair comparisons.

Implementation Details: Parameter learning for the grid model is performed by first scaling all the ground truth bounding boxes to a fixed height (100 pixels in our experiments) while preserving the aspect ratio. Then the points are uniformly sampled along the edges (using a Canny edge detector) at a small set of fixed scales and multiple type features are extracted to construct the feature database (refer Table 2 for a list of features used).

We obtain the computational costs for every type of feature by running the algorithm on the training images and averaging the time taken to update our detection hypotheses which includes time to extract the feature and perform a nearest neighbor lookup. The costs reported in the table were computed using on 2.2 Ghz dual core machine.

For the parameters of the approximate nearest neighbor method, RBV, we follow the heuristics mentioned in their paper [14]. Nearest neighbor lookup parameters γ, ϵ are set based on the average value of the distance of every features in the training database to a small number of features (500). We used a grid model of size $5 * 5$ in all experiments to obtain parts of a reasonable size of 20x20 pixels. While learning the grid weights using the max-margin formulation we used a small value of the $C = 1$ to allow for generalization.

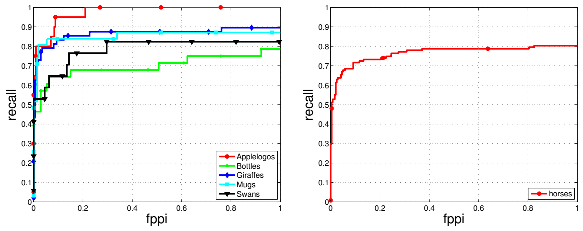
4.1. Detection

In this section, we provide overall detection results when our approach is allowed to use all the features and run to completion. These results lets us compare the method with the state-of-the-art techniques and can be considered as upper-bounds for the active selection procedure. For these comparisons, we use SIFT features sampled in all four color channels of an image (gray, red, green, blue) along the edges and a small set of scales.

Figure 3 shows the grid weights w_i superimposed on example training images for the five categories in the ETHZ

Category	Our Approach				State-of-the-art			
	Average Precision	Recall @ FPPI			Voting based (FPPI = 1.0)		With verification (FPPI = 0.3/0.4)	
		0.3	0.4	1.0	PMK rank [24]	M^{2}HT [21]	[24]	[21]
Applelogos	86.5	100.00	100.00	**100.00**	80.0	85.0	95.0/95.0	95.0/95.0
Bottles	60.5	67.9	67.9	78.6	**89.3**	67.0	89.3/89.3	92.9/96.4
Giraffes	78.3	87.5	87.5	**89.6**	80.9	55.0	70.5/75.4	89.6/89.6
Mugs	79.5	83.9	87.1	**87.1**	74.2	55.0	87.3/90.3	93.6/96.7
Swans	62.8	82.4	82.4	**82.4**	68.6	42.5	94.1/94.1	88.2/88.2
Average		84.3	85.0	**87.5**	78.6	60.9	87.2/88.8	91.9/93.2

Table 1. Comparisons to the state-of-the-art on the ETHZ dataset. Our grid based model produces state-of-the-art results among hough based approaches and is comparable to methods that search exhaustively.



(a) ETHZ dataset

(b) INRIA dataset

Figure 4. Detection rate vs FPPI for all the cateogires in the ETHZ and INRIA horses dataset.

shape dataset. Bright regions have high weights while dark regions have low weights. We note that despite the fact the ground truth bounding boxes for training images include background regions, the learned weights for the object regions are substantially high. The learned weights are reasonable for all the categories and do reflect the parts of an object that are the most discriminative. For example, the grid containing the leaf has the highest weight for applelogos, while the head seems to be the most discriminative part for the categories giraffes and swans, and the handle has the largest weight in the case of mugs. Finally we also note that in the case of bottles our model learns almost uniform weights which could either mean that none of the parts are more discriminative than the other or that the features used do not capture the parts correctly.

Figure 4 (a) and (b) show detection rate plotted against the FPPI (false positives per image) on the ETHZ shape and the INRIA horses datasets. We use the PASCAL criterion of 50% overlap to score a detection. For the ETHZ data we see that except for bottles, our approach performs quite well with 100% recall on applelogos and close to 90% on the other categories. The poor precision on bottles is understandable given that it doesn't have a distinctive shape (two parallel lines could be miscontrued to be a bottle) and also dividing the small width of bottles into five parts could be detrimental to performance. The results on the INRIA dataset are comparable to the state-of-the-art voting based methods. At an FPPI of 1.0 our approach has a recall of 80.3% while [10] report a value of 80.7%. However, at a lower FPPI of 0.1 our approach has a higher recall of 71% compared to 66% reported by [10]. Using an additional verification classifier [21] report a detection rate of 85.47%.

Table 1 further compares the results of our approach against the state-of-the-art detection methods on ETHZ dataset. We divide the table into methods that are based on Hough voting alone (our approach, PMK rank [24] and M^{2}HT [21]) and methods that use an additional verification stage which exhaustively searches around the set of candidate locations returned by voting based approaches. Among voting based approaches our grid model produces the best recall at an FPPI of 1.0 for all the categories except bottles. There is also a significant difference in the average value of recall at 1.0 FPPI between our approach and the next best voting based approach (87.5 vs. 78.6).

We also report the recalls at 0.3/0.4 FPPI to compare against methods that use an additional verification stage and we find that our approach produces comparable results for most of the categories (in fact better results for applelogos) even though it does not resort to an exhaustive search. The difference in numbers against verification based methods could be attributed to the highly non-linear kernels used by these approaches as opposed to our linear classifier.

4.2. Active Selection

In the previous section, we provided results for our detection method when it was allowed to utilize all the features extracted from an image, which in some cases was over 10,000 features and took upto 10-20 secs for a detection. This was mainly to provide an upper bound for our active selection results and to compare against other similar approaches. In this section, we provide results for our active selection strategy and show that we can achieve similar results by sampling far fewer number of features.

For experiments in this section we use the Daisy descriptors P64_T1a_S2_17, P18_T2_S2_9 of [28] in addition to SIFT (Table 2). The Daisy descriptor is constructed by first mapping the input image patch using a transformation block {T1 - gradient, T2 - gradient, T3 - steerable filter, etc.}, followed by spatial binning of the transformed block {S1 - square grid, S2 - polar arrangement, etc. } to obtain N linearly summed vectors (N = 9, 17). The low dimensional P18 produces very coarse features while P64 produces highly specific features and both have been used successfully for various image/patch matching problems [28]. We sample these features only along the edges of the image.

We compare our active selection strategy against a passive baseline which chooses a random position in the image space and a random feature type to extract at every iteration. We first sample a small percentage of edge points (five in our experiments) and generate an initial set of hy-

Feature	Channel	Dim	Computation time (ms)
SIFT	R, G, B, Gray	128	0.21
P64_T1a_S2_17	Gray	68	1.2
P18_T2_S2_9	Gray	36	0.09

Table 2. Attributes of the features used in the experiments.

potheses. Then, we run each selection strategy iteratively updating the hypotheses as features get added until a fixed amount of time has lapsed (1 sec in our case).

In Figure 5 we show some qualitative results comparing the first 1000 points selected by our active selection approach and the passive selection baseline. The first row contains example images from every category in the ETHZ shape, the second row and third rows show the points selected by the active and random selection strategies, respectively. Bright dots denote selected feature points.

We see that while random selection uniformly samples all the edge points in the image our active selection strategy selects most of the features around the target of interest while sparsely sampling the other points. In addition, our approach targets the object part that provides the best gain in score, for example, “handle” for “mug”, “head”, “body” for “giraffe”. By constantly targeting feature computation around points that provide better gains in the detection scores our approach is able to make efficient use of computational resources.

In Figure 7(left) we show a split of the selected points across the three different features that are available for extraction for an example image. The first and second rows show the points selected by active and random approaches, respectively, while the columns denote the type of feature selected. While random selection obtains the same number of features from all three types active selection samples fewer features of type P64 which takes more time to compute (Table 2) and more features of type P18 which takes much lesser time to compute.

We also generated detection cuves by plotting the accuracy of the two approaches with increasing time and averaging results over five different random initializations. We use the average precision as a measure of accuracy since it provides a complete summary over all the points in the precision-recall curve and it is accepted in the literature [6] as one of the main evaluation criterion for detection.

Figure 6 shows the accuracies of the two selection strategies (active, random) with increasing time for all the categories in the ETHZ shape dataset and the INRIA dataset. We see that for most of the categories our active selection strategy has a steeper slope. This implies that our approach detects objects in images faster than the random selection baseline. In addition, compared to the upper bound reported in Table 1 which uses a large number of features, our active selection approach is able to reach similar accuracies at lower computational cost (less than 1 sec or 500-1000 features). This illustrates the advantage of active vision, where feature extraction is actively guided by the current model.

So far, we have shown active selection results when a single object detector is run on an image. However, our active selection approach is not restricted to this setting; it can also be applied in a multi-class scenario where the selected features are shared by more than one object detector. This is because the selection criterion is maximized over all objects O in addition to locations and feature types.

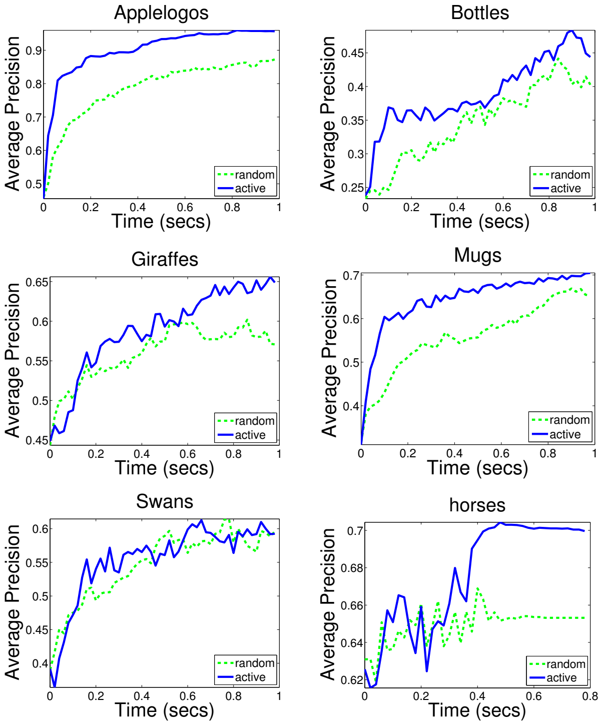


Figure 6. Results for active and random selection of gray image features among SIFT, P64_T1a_S2_17 and P18_T2_S2_9.

We run all five detectors on the features selected by active and random selection approaches. We report the multi-class classification accuracy where the class of an image is determined as that of the detector which returns the largest detection score for that image. Figure 7(right) shows the multi-class accuracy for the two approaches with increasing time. Our approach does better than random sampling in this setting too. Although we would like to investigate this further, this preliminary result further illustrates the advantages of active feature selection.

5. Conclusions

In this paper we presented a methodology that considers computational resources and requirements in a visual recognition and detection task. The proposed method utilizes the decision-theoretic principle of value-of-information in order to guide the feature acquisition process by predicting both the location and the type of feature to be extracted. Empirical comparisons on two challenging object detection datasets showed that actively predicting the next feature to acquire during detection makes more efficient use of computational resources and therefore it can localize objects faster than passive feature selection. In future, we plan to

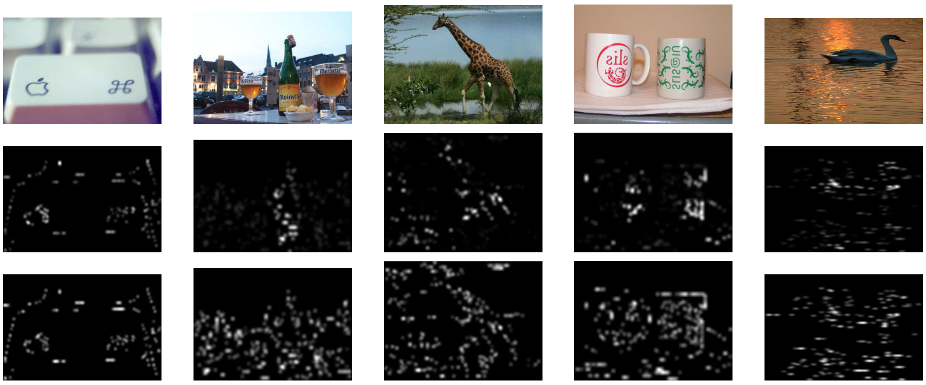
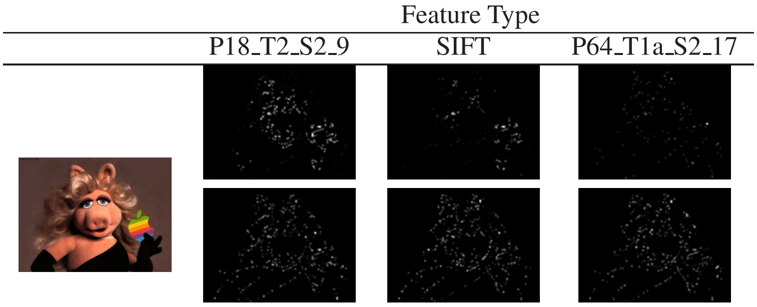


Figure 5. Example images showing the locations of the first 1000 features selected by active (**Second row**) and random (**Third row**) approaches. Locations where features were queried are denoted by white dots. Our approach samples by obtaining features that provide better gains in the scores is able to target feature computation around the target objects in the test image unlike random selection.



Time (secs)	Accuracy (%)	
	Active	Random
0.3	70.0	70.0
0.4	75.0	70.0
0.6	76.1	74.1
0.8	80.0	74.9
1.0	79.0	76.0

Multi-class classification accuracies (ETHZ)

Figure 7. **Left:** Features selected from among SIFT, P64_T1a_S2_17 and P18_T2_S2_9 by active and random methods. Our approach selects features that are faster to compute in order to utilize the alloted time to the maximum **Right:** Multi-class classification accuracies (ETHZ). Our approach selects locations and features to compute by maximizing the VOI over all five objects and produces better multi-class classification accuracies at lower costs.

investigate extending the approach to non-linear classifiers and articulated objects.

References

[1] A. Andreopoulos and J. K. Tsotsos. A Theory of Active Object Localization. In *ICCV*, 2009.

[2] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu. An Optimal Algorithm for Approximate Nearest Neighbor Searching Fixed Dimensions. *J. ACM*, 1998.

[3] J. Butko, Nicholas and R. Movellan, Javier. Optimal Scanning for Faster Object Detection. In *CVPR*, 2009.

[4] M. S. Charikar. Similarity Estimation Techniques from Rounding Algorithms. In *STOC*, pages 380–388, New York, NY, USA, 2002. ACM.

[5] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *In CVPR*, pages 886–893, 2005.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results.

[7] P. Felzenszwalb, D. McAllester, and D. Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. In *CVPR*, 2008.

[8] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial Structures for Object Recognition. *Int. J. Comput. Vision*, 61(1):55–79, 2005.

[9] R. Fergus, P. Perona, and A. Zisserman. Object Class Recognition by Unsupervised Scale-Invariant Learning. In *In CVPR*, pages 264–271, 2003.

[10] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid. Groups of Adjacent Contour Segments for Object Detection. *In PAMI*, 2008.

[11] V. Ferrari, F. Jurie, and C. Schmid. Accurate Object Detection with Deformable Shape Models Learnt from Images. In *CVPR*, 2007.

[12] V. Ferrari, F. Jurie, and C. Schmid. From Images to Shape Models for Object Detection. *International Journal of Computer Vision*, 2009.

[13] V. Ferrari, T. Tuytelaars, and L. V. Gool. Object Detection by Contour Segment Networks. In *CVPR*, 2007.

[14] J. Goldstein, J. Plat, and C. Burges. Redundant Bit Vectors for Quickly Searching High-Dimensional Regions. In *Deterministic and Statistical Methods in Machine Learning*. 2005.

[15] S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Messner, G. Bradski, P. Baumstarck, S. Chung, and A. Y. Ng. Peripheral-Foveal Vision for Real-time Object Recognition and Tracking in Video. In *IJCAI'07:*, 2007.

[16] E. Horvitz and S. Zilberstein. Computational Tradeoffs under Bounded Resources. *Artif. Intell.*, 126(1-2), 2001.

[17] T. Kadir, A. Zisserman, and M. Brady. An Affine Invariant Salient Region Detector. In *ECCV*, 2004.

[18] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond Sliding Windows: Object Localization by Efficient Subwindow Search. In *CVPR*, 2008.

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

[20] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *In ECCV workshop on statistical learning in computer vision*, pages 17–32, 2004.

[21] S. Maji and J. Malik. Object Detection using a Max-Margin Hough Transform. In *CVPR*, 2009.

[22] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A Comparison of Affine Region Detectors. *International Journal of Computer Vision*, 2005.

[23] S. Munder and D. M. Gavrila. An Experimental Study on Pedestrian Classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(11):1863–1868, 2006.

[24] B. Ommer and J. Malik. Multi-Scale Object Detection by Clustering Lines. In *ICCV*, 2009.

[25] C. Papageorgiou and T. Poggio. A Trainable System for Object Detection. *Int. J. Comput. Vision*, 38(1):15–33, 2000.

[26] T. Quack, V. Ferrari, B. Leibe, and L. V. Gool. Efficient Mining of Frequent and Distinctive Feature Configurations. In *ICCV*, 2007.

[27] P. Viola and M. Jones. Robust Real-Time Face Detection. *IJCV 2004*.

[28] S. Winder, G. Hua, and M. Brown. Picking the Best Daisy. In *CVPR*, 2009.