A general boosting-based framework for active object recognition

Zhaoyin Jia
zj32@cornell.edu
Yao-Jen Chang
yc682@cornell.edu
Tsuhan Chen
tsuhan@ece.cornell.edu

Electrical and Computer Engineering Department
Cornell University
Ithaca, NY, USA

Abstract

We propose a novel general framework with a boosting algorithm to achieve active object classification by view selection. The proposed framework actively decides the next best view for the recognition task. It evaluates different information sources for top hypotheses, generates a voting matrix for candidate views and the view selection is achieved by picking up the one with the maximum votes. Three different sources - similarity based on Implicit Shape Model, prior for model, and prior for views - are presented in the paper. Moreover, we convert view selection itself into a classification problem, and propose a boosting algorithm that is able to combine the previous sources. Experiments show that our algorithm produces a better strategy compared to the other baseline methods.

1 Introduction

In recent years the problem of object recognition have been extensively studied and the performance is constantly improved [6][8][13][14][15] [17]. Other than recognition in a single image, researchers also focus on multi-view recognition and utilize different relations [19][21] [22]. Under many circumstances users will have the access to control the vision system, such as a guided robot. In that case not only can we acquire multiple views, but also are able to *actively* control the system to pick a certain angle, moving the camera to a more discriminative view point and disambiguating the confusion between potential object hypotheses. This is where the context of active view recognition appears.

Take Fig.1 as an example, from 0^o this toaster is hard to identify even for human. However the image at 135^o becomes a good view point to distinguish the correct hypothesis from wrong ones. This image has a typical appearance for a toaster with buttons and a little knob. It is the intuition underlying the active object recognition. Unlike taking images at randome, active recognition selects an view point that can solve the current ambiguity with the object. It will help the recognition model identify the object with a higher accuracy at early steps, or require fewer images to achieve the same confidence.

In this paper we propose a new framework that can adaptively make use of different criteria. This generalized framework takes the current top K hypotheses, evaluates different

(a) 0^o (b) 45^o (c) 90^o (d) 135^o

Figure 1: Some view angles are easier to identify the object while the others are hard, e.g. 0^o is not a good view for identify this toaster. However, at 135^o, it becomes much easier. In this paper, we propose a strategy to selecting the views to maximize the recognition performance

sources and gathers the votes for the next candidate view. Three sources are presented: similarity on Implicit Shape Model, prior for model and prior for view. We convert the view selection into a classification problem, and propose a boosting algorithm that can utilize all these sources to obtain a better active recognition performance.

The remainder of the paper is organized as follows: in Sec.2 we introduce the previous work and emphasize the contribution made by this article. In Sec.3 we briefly describe the recognition method and the settings in this problem. In Sec.4 we propose our framework for view selection and present three different sources in order: similarity on Implicit Shape Model, prior for model and prior for view. In Sec.5 a boosting algorithm is proposed to learn the optimal combination of different sources. Finally in Sec.6 the experiments are shown and analyzed.

2 Related work

Active vision has been discussed in some previous papers while several different criteria for various tasks have been proposed and implemented [1][2][3][4] [6][9][10] [11][12][16] [18][20]. In[6][16], 3D environment is attentively searched and an object recognition algorithm is combined. For the task of view selection only, the ideal method would be modeling the object in a continuous view-angle setting, and optimizing globally over infinite steps ahead. However the modeling complexity and NP-hard computation prevent us from doing that, therefore many researchers alter to an approximate solution by quantizing the model in fix view-angle intervals and finding the myopia solution for view selection, which still gives a good result comparing to random selection.[3][4][11]

In [3][4], the authors presented a view selection algorithm based on Mutual Information. The algorithm tries to select those views with less uncertainty in the training set, and the uncertainty for one hypothesis is modeled as entropy. The view selection is achieved by maximizing the information gain across all the hypotheses and possible actions. [11] tried to measure the similarity between two candidate hypotheses by Jeffrey Divergence. The idea is that the optimal view should compose an image that is not similar to any other hypotheses, and the algorithm selects the view by maximizing this criterion.

One major difference and contribution of our paper is that we try to solve the problem by making use of the vision feature and putting it into our generalized framework. We propose a similarity on actual visual words of the training model. We believe this source can provide a direct and better understanding in the concept of similarity. In [1] authors propose an appearance based descriptor for selecting optimal views in multi-views. It should be noted that our

approach is still significantly different from that, for we propose a measurement of similarity directly linked to the recognition model and firstly apply it for the active recognition task.

Another novelty of our framework would be its ability to take different sources as criteria and assemble them through boosting. The input sources could be the appearance similarity, entropy sampled before or the prior probability. This framework works as a generalization format, and the boosting algorithm can give a better view-selection strategy by combining these sources.

3 Image category and pose classification

We can denote different view angles as $v, v+d, v+2d, \ldots, v+(n-1)d$ and form an image sequence, where v is the starting view angle, d is the degree interval, and n is the total number of views. In the following paper, the concept of selecting tth view means selects the tth image in this image sequence. Once the tth view is selected for recognition, it is then no longer a candidate view for the next step since it will provide no more information.

The multi-view recognition problem could be considered as a combination of several single-view recognition tasks [17][23]. We use Implicit Shape Model for category classification on single image, similar to the procedure conducted in [13][23]. We model different views of a category separately, which not only makes the recognition performance more accurate, but also enables us to identify the pose the object. Multi-view classification is usually achieved by linking the single view responses together and sums/multiply them up [3][11]. For instance, suppose the initial image has belief $P(O_{c_i,v_1})$ in one hypothesis, where c_i and v_1 mean the predicted category and pose respectively. Then after selecting tth image in the view sequence, the multi-view recognition belief P_m is calculated by multiply the response of the corresponding model: $P_m = P(O_{c_i,v_1}) \times P(O_{c_i,v_1+(t-1)\times d})$. And the predicted category/pose is the maximum hypothesis response.

4 Framework for active view selection

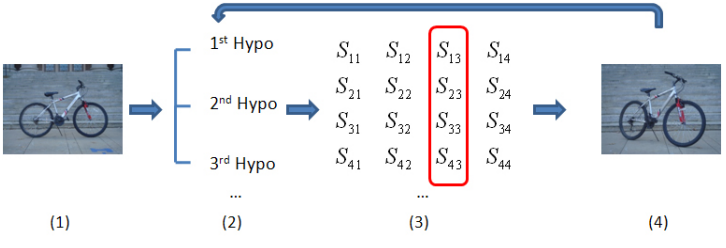


Figure 2: The flow-chart of the framework: (1) input image. (2) perform recognition on the image, and select the top K hypothesis. (3) generate the view matrix S based on the hypotheses (4) select and move to the next view using S.

The proposed framework is described in a flow chart in Fig. 2: given one input image, firstly classification on this image is performed, which will give a belief on each hypothesis. Then these beliefs are sorted and the top likely K hypotheses are picked up. A view matrix S is generated with respect to these hypotheses. S contains the votes for all the candidate views

and the next view is decided by finding the view with the highest voting. Each column in matrix S corresponds to one candidate view, and the number of row is related to the number of top hypotheses K kept. Also S becomes the feature space in the learning step. Firstly we exploit several criteria to generate this matrix.

The basic idea of generating the view matrix S is that the view selection should be related to the information that training set provides, as well as what the current *top* hypotheses are. If all the potential hypotheses are kept, then we only rely on the information from the training phase. This is not always the preferable case, because if the hypotheses are not the top ones, the confusion between them are less important. There usually exist a large number of hypotheses, and many of them will only contribute a small belief value for classification, but summing them up will unnecessarily affect the decision for view selection. Therefore choosing a clever K will improve the result since only top K hypotheses are kept to disambiguate. We propose a boosting algorithm to learning the weigh for the top K hypothesis, balancing the knowledge provided by the training set and the current testing image.

4.1 Similarity

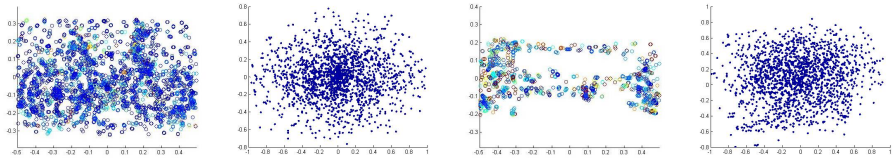
To separate two models, the next view should correspond to models that are least similar with each other. That is the intuition of using similarity for view selection. We measure the similarity based on the recognition model we use: the Implicit Shape Model[13]. This model is a spatial collection of the words $\{w_i\}$ in the codebook with different priors and weights, as shown in Fig.3 (a) and (c). During training it builds a model by recording the visual words at their occurrence location $\lambda = (\lambda_x, \lambda_y)$. In the inference step, a feature e collected at location l of a testing image is firstly mapped to the codebook $\{w_i\}$, and cast a vote on the ISM $O_{c,v}$ (category c and view angle v) centering at location $\lambda + l$. The votes are collected in a voting space and object detection is achieved by finding the maximum. Such comprehension enables us to measure the similarity between two Implicit Shape Models by casting one model to another and examining the voting space. For two ISMs, O_1 and O_2, we firstly measure the one-way similarity $Simi'(O_1, O_2)$ by treating the word $w_{1,i}$ in O_1 as a feature e in the test image, and calculate the voting of this to all the words $w_{2,j}$ in the model O_2.

$$p(O_2, \lambda | w_{1,i}, l) = \sum_j p(O_2, \lambda | w_{2,j}, l) \cdot p(O_2 | w_{2,j}) p(w_{2,j} | w_{1,i}) \quad (1)$$

The term $p(w_{2,j}|w_{1,i})$ measures the similarity between words $w_{2,j}$ and $w_{1,i}$. Fig.3 (c) and (d) show examples of the voting spaces when we cast the model O_1(bicycle, view angle 90^o) to a similar model (itself) and to a different model (stapler, view angle 90^o). It shows that similar models result in a more congregated voting space in the center. Since the models are built in a way that location λ is normalized, we evaluate the votes at the central region and sum them as the measurement for similarity. We represent the central region as a circle with a pre-set radius rd. Therefore the one-way similarity $Simi'(O_1, O_2)$ becomes:

$$Simi'(O_1, O_2) = \sum_{i, |\lambda| < rd} p(O_2, \lambda | w_{1,i}, l) \quad (2)$$

Furthermore the similarity should be irrelevant of the order, i.e. $Simi(O_1, O_2) = Simi(O_2, O_2)$. This is achieved by adding two one-way similarities together: $Simi(O_1, O_2) = Simi'(O_1, O_2) + Simi'(O_2, O_1)$



(a) (b) (c) (d)

Figure 3: (a) and (b): the Implicit Shape Models of bicycle(a) and stapler(b) from view angle 90^o. Color represents different weights for visual words. (c) and (d): the voting space when measuring the similarity of the similar models (c) and distinct ones (d).

4.1.1 View matrix from similarity

The view matrix S is acquired from similarity as follows. We pick up the top K hypotheses and choose two of them: O_{c_1,v_1}, O_{c_2,v_2}. For hypothesis O_{c_1,v_1}, if it is true, the following images in the sequence (not yet acquired) would be $seq_1 : O_{c_1,v_1+d}, O_{c_1,v_1+2d}, O_{c_1,v_1+3d} \ldots$. Similarly for O_{c_2,v_2} the followings would be $seq_2 : O_{c_2,v_2+d}, O_{c_2,v_2+2d}, O_{c_2,v_2+3d} \ldots$. We can align seq_1 and seq_2 correspondingly, and measure the pairwise similarity for each view: $R_{simi}(t) = Simi(O_{c_1,v_1+(t-1)d}, O_{c_2,v_2+(t-1)d})$ where t is the index of a candidate view. The view with the minimum similarity is calculated $\tilde{t} = \underset{t}{\operatorname{argmin}}(R_{simi})$ and one vote is cast for $\tilde{t}$. We repeat this for all the pairs of top K hypotheses, and record the votes for each candidate view in the view matrix S, resulting in C_K^2 rows. We make the voting discrete: the corresponding column $S_{i,t}$ is set to 1 if $t = \tilde{t}$, and 0 otherwise. For the next step, we can select the view $\bar{t}$ that has the maximum votes.

$$\bar{t} = \underset{t}{\operatorname{argmax}} \sum_i S_{it} \tag{3}$$

In the situation when the final goal is to identify the category of the image only regardless of the pose, we modify the algorithm as follows: when picking up the top K hypotheses and generating the view matrix S, the pair of hypotheses is omitted if they predict the same category.

4.2 Prior for model

Another factor taken into consideration is whether a model itself is easy to recognize or not. Models have different performance in recognition, and preferred models are those empirically easier to recognize. Therefore it provides another source for view selection: the prior for the model. A separate validation set is used for evaluating the prior of a model. This part could be considered as a training set for view selection. We evaluate them on the implicit shape model, and sample the prior for each model $O_{c,v}$ as $Pr(O_{c,v}) = TP/(TP+FP)$, where TP and FP are the number of true positives and false positives of the model when testing the validation set.

For one hypothesis $O_{c,v}$, the prior $R_{pr}(t)$ for the candidate views t is calculated as $R_{pr}(t) = Pr(O_{c,v+(t-1)d})$. And the vote for the next view is: $\tilde{t} = \underset{t}{\operatorname{argmax}}(R_{pr})$. One hypothesis corresponds to one row S_i in view matrix S_{it}, and the column value is set to 1 if $t = \tilde{t}$ and 0

otherwise. We record the votes for all K hypotheses and form view matrix $\{S_{iu}\}$ of K rows. The next view is chosen by using Eq.3.

4.3 Prior for view

We also sample $p(t|O_{c,v})$: the optimal views to pick-up given the ground truth $O_{c,v}$, and use it as another information source for view selection. $p(t|O_{c,v})$ is calculated in the following way: given the ground truth hypothesis $O_{c,v}$ of one testing image from the validation set, the optimal view index t is selected by picking up the view t that can maximize the corresponding true hypothesis $O_{c,v+(t-1)d}$, while minimizing all the other wrong ones. Having acquired all the optimal views, the prior for view is calculated in a Bayesian probability format: $p(t|O_{c,v}) = p(t, O_{c,v})/p(O_{c,v})$.

During the view selection step, $p(t_i|O_{c,v})$, the priors for all the candidate views of one hypothesis $O_{c,v}$, can be retrieved and it generates a vote for view $\tilde{t}$ where $\tilde{t} = \underset{t}{\operatorname{argmax}}(p(t|O_{c,v}))$. Similar to the previous section, this hypothesis forms one row in the view matrix S where the $\tilde{t}$th column has value 1 and 0 elsewhere. Also the next view is selected using Eq.3.

5 A combined boosting algorithm

From the previous sections we introduce three different criteria to form the view matrix S. In this section we propose a boosting algorithm that aims to combine different sources together and transform it into a better strategy. Noticing that the number of hypotheses kept K is a parameter that balancing the knowledge between the training set and the current test, varying K may provide us richer information. Here the view selection is considered as a classification problem: given different sources as the input feature, a preferred view is classified from unwanted ones. Following this idea we propose a boosting algorithm [2] to learn the view selection scheme.

The training instances for view selection is collected by selecting the optimal view t sequentially introduced in Sec.4.3. For each training image I_m, an optimal view sequence $t_{m1}, t_{m2}, \ldots, t_{mN}$ is collected, where N is the number of candidate views available. In this sequence t_{m1} is the optimal next view, and t_{mN} is the least wanted view.

A source u(could be similarity, prior for model or prior for view) of parameter K is considered as a *weak classifier* $C_{u,K}(t|I_m)$. Using Eq.3, $C_{u,K}(t|I_m)$ gives a classification result as follows:

$$C_{u,K}(t|I_m) = \begin{cases} +1, t = \bar{t} \\ -1, t \neq \bar{t} \end{cases} \tag{4}$$

where $\bar{t}$ is the predicted view from the source, I_m is the input image and K is the number of hypotheses kept. These weak classifiers can be combined using Adaboost. By varying K, the view selection results differ a lot, especially when K is small. This is because of the disagreement view decision by the top K hypotheses. We utilize such feature to obtain a richer set of weak classifiers. Three sources of different K value are collected as the weak classifier and weighted by boosting algorithm.

Some modifications are necessary to meet the needs of this special problem. The training error of each $C_{u,K}$ is redefined as follows: for one training image I_m with the optimal sequence $\{t_{mn}\}, n = 1, \ldots, N$, the weak classifier $C_{u,K}$ classifies view $\bar{t}$ as positive. We find

the index $d(\bar{t})$ of $\bar{t}$ in the optimal sequence $\{t_{mn}\}$, and calculate training error as

$$err_{u,k}(I_m) = \begin{cases} \frac{d(\bar{t})-1}{N-1} & d(\bar{t}) \leq W \\ 1 & d(\bar{t}) > W \end{cases} \quad (5)$$

W is a given threshold. The intuition is that if the predicted view $\bar{t}$ is close to the optimal one, i.e. it is still the top W optimal view in the sequence $\{t_{mn}\}$, then we do not give a full penalty to $C_{u,K}$. We change the training error to a soft value $(d(\bar{t})-1)/(N-1)$. The training error is set to 1 if $\bar{t}$ is over the accepted index W, which indicates $C_{u,K}$ picks up a quite unwanted view. Also we consider that the instance I_m is correctly classified when $d(\bar{t}) \leq W$. The training phase of the boosting algorithm is described in algorithm.1 where R is the max

Algorithm 1 Algorithm to train the view selection

Given M training images $\{I_m, m = 1,\ldots,M\}$, their corresponding optimal view sequence $\{t_{mn}\}$, and a set of weak classifiers $\{C_{u,K}\}$ with various u and K, initialize $D_1(m) = 1/M$

for $r = 1$ to R **do**

evaluate every $C_{u,K}$ on all $\{I_m\}$ and calculate weighted error $\varepsilon_{u,k} = \sum_{m=1}^{M} err_{u,K}(I_m) D_r(m)$

Select $C_r = \underset{C_{u,K}}{\operatorname{argmin}} \varepsilon_{u,K}$. and record the training error ε_r of C_r.

Choose α_r for C_r: $\alpha_r = 0.5 \ln \frac{1-\varepsilon_r}{\varepsilon_r}$

Update $D_{r+1}(m) = \frac{D_r(m) \exp(-\alpha_r \cdot Id(err(I_m)=1))}{Z_t}$

end for

iteration number, and Z_t is a normalize term to ensure $\sum_m D_{r+1} = 1$. $Id(err(I_m) = 1)$ is a validation function, it gives 1 if $err(I_m) = 1$, which means it is wrongly classified and thus increases the weight $D_r(m)$ as I_m is a "hard instance". It gives -1 otherwise.

The algorithm outputs a final combined classifier $C_A = \sum_{r=1}^{R} \alpha_r C_r$ in the training phase. Given one testing image I', the view selection is achieved by calculating the response of C_A for each candidate view $t_i, i = 1,\ldots,N$, and pick up the one with the maximum belief: $\bar{t} = \underset{t_i}{\operatorname{argmax}} C_A(t_i|I')$

6 Experiment

To make the results easily comparable, we evaluate different algorithms and compare them on the UIUC Dataset of 3D object categories [13]. For one object from a specific height and scale, this dataset contains 8 images of different views with 45^o intervals, shown in Fig.1. The evaluation is performed on nine categories of UIUC 3D dataset (car, toaster, cell phone, bicycle, iron, stapler, shoe, head, mouse). Around 3000 images are used for training $8 \times 9 = 72$ Implicit Shape Models, and SIFT descriptor is applied as the local feature, clustered in 800 visual words. Another 2000 images are used for training the action of view selection, and 2000 images are used for final testing. We empirically set $rd = 0.2$, $R = 100$ and $W = 3$.

One way to evaluate the performance is to examine the classification accuracy along each step, and a good scheme would process a higher accuracy at early steps. The evaluations are made for two different goals: One aims to classify the pose and category for the input image.

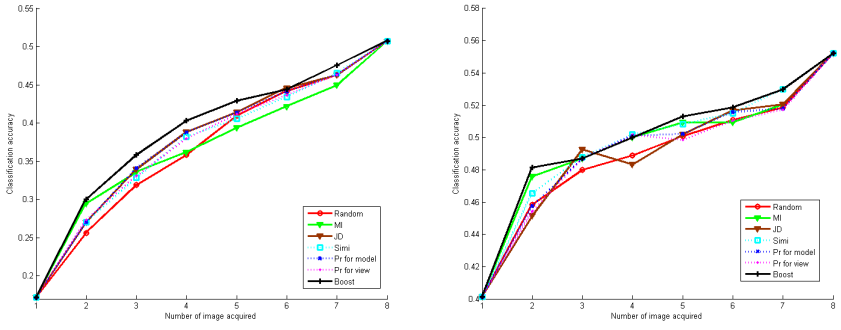


Figure 4: Performance of different schemes. (a) is the result of classification for both the pose and category of the object. (b) is the result aiming to classify the category of the object only.

Another is evaluated on classifying the category only, regardless of the pose, and the test image is regarded correctly classified once the predicted category matches the ground truth.

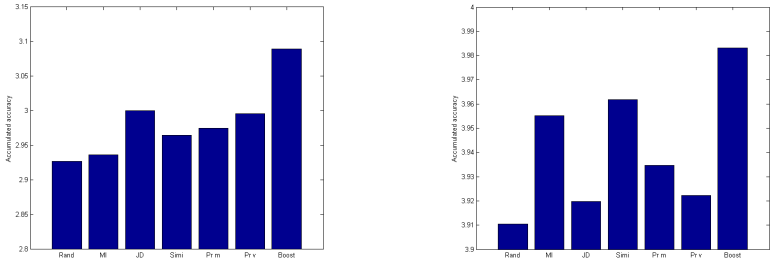
Fig.4 shows the performance of different schemes for two types of classification goal. X axis is the number of images acquired, i.e the steps; Y axis is the average classification accuracy. We compare our method to three baselines: random selection, methods proposed in [3] (MI) and [10] (JD). The red, brown, green curves are the three baseline methods respectively. The cyan, blue and magenta dotted lines are the performance of different view selection schemes based on similarity, prior for model and prior for view respectively. The performance is averaged with $K = 2, \ldots, 21$. The black line is the performance of the boosting algorithm. In each of the three sources, we gather 20 weak classifiers for three sources, and in total we have 60 classifiers to combine.

The accuracy should be the same for all the schemes in the first step and after acquiring all the views, since underlying classification method is the same. The experiment shows that the boosting algorithm achieves best performance in the midway, for it can utilize different sources with various parameter settings, and combine them into a more discriminative scheme.

Another evaluation is to calculate the accumulated accuracy, i.e the sum of the accuracy for all the steps. It represents the area under the curve in Fig.4, and a higher value leads to a better scheme for view selection. We compare the accumulated accuracy for different schemes in Fig.5. It also shows that overall the boosting algorithm achieves the highest accumulated accuracy comparing to other methods.

Fig.6 gives a view selection result from boosting algorithm for category and pose classification. Given the initial image, such as an iron from side view, the algorithms moves to the view of $90^o, 45^o, 270^o$ to the right of the initial view. One can see that these views are quite different from the initial view and thus may become more informative. The algorithm leaves 180^o to the last, which is a mirror of the initial image and most likely it cannot provide valuable information for classification.

We investigate the accumulated accuracy of the three proposed sources while changing K from 2 to 72 (maximum), shown in Fig.7. As mentioned before, the performance varies a



(a) (b)

Figure 5: Accumulated accuracy for different schemes. From left to right they represent the accumulated accuracy for: (1) random scheme, (2) Mutual information in [3], (3) Jeffrey Divergence in [11], (4) similarity (5) prior for model, (6) prior for view, and (7) boosting algorithm. (a) is for classifying both category and model. (b) is for category only

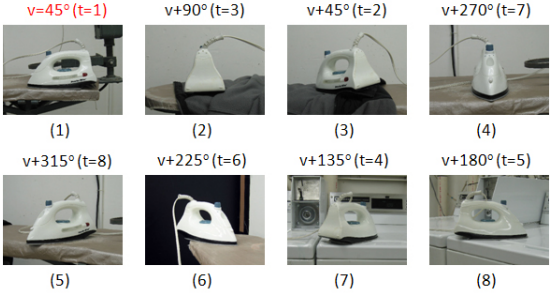


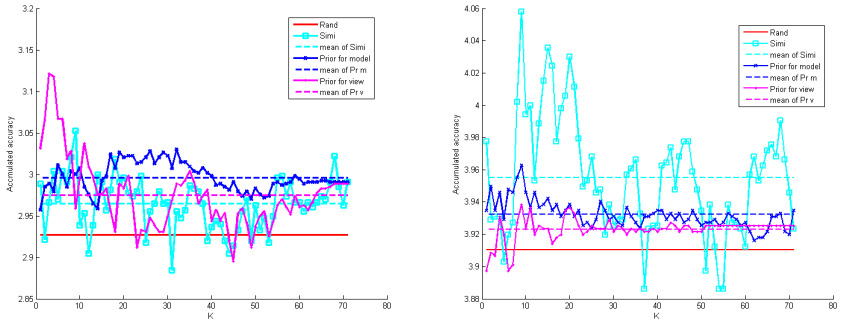
Figure 6: A result of view selection for boosting algorithm. Number at bottom notes the order of the image in the selected views.

lot with different K value. It is may because the votes for the next view from these K top hypotheses cannot agree. However in the most time, three sources provide better performance over the random method.

7 Conclusion

In the paper we present a novel framework of view selection for active recognition, propose different sources for the framework, and combine them with a boosting algorithm. Results show that each individual source can provide a better scheme than the random, and the boosting algorithm achieves the best result in recognition accuracy.

One future work would be extending the view selection in real 3D environment. With a free camera controlled by the user, it is possible to mapping feature points of the interested object in 3D location, and active view selection as well as multi-view recognition may benefit from such extra information. Another extension would be combining the view selection method with the recognition algorithm. Rather than simply multiplying the belief response from different views, those views are ambiguous should have less weights, and then after



(a) (b)

Figure 7: Accumulated accuracy of different sources with variation of K. (a) is the classification for both category and pose; (b) is for category only

acquiring all the images the classification result won't be lowered by those views.

References

[1] S. Abbasi and F. Mokhtarian. Automatic view selection in multi-view object recognition. In *ICPR*, pages Vol I: 13–16, 2000.

[2] Hermann Borotschnig, Lucas Paletta, and Axel Pinz. A comparison of probabilistic, possibilistic and evidence theoretic fusion schemes for active object recognition. *Computing*, 62(4):293–319, 1999.

[3] J. Denzler and C. M. Brown. Information theoretic sensor data selection for active object recognition and state estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):145–157, 2002.

[4] F. Farshidi, S. Sirouspour, and T. Kirubarajan. Robust sequential view planning for object recognition using multiple cameras. *Image and Vision Computing*, 27(8):1072–1082, July 2009.

[5] P. F. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, pages 1–8, 2008.

[6] Per-Erik Forssén, David Meger, Kevin Lai, Scott Helmer, James J. Little, and David G. Lowe. Informed visual search: Combining attention and object recognition. In *ICRA*, pages 935–942. IEEE, 2008.

[7] Freund and Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *JCSS: Journal of Computer and System Sciences*, 55, 1997.

[8] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *CVPR*, pages 1022–1029, 2009.

[9] Stephen Gould, Morgan Quigley, and Andrew Y. Ng. Peripheral-foveal vision for real-time object recognition and tracking in video, 2007.

[10] Zhaoyin Jia, Yao-Jen.Chang, and Tsuhan Chen. Active view selection for object and pose recognition. In *ICCV 3DRR Workshop*, pages 1–8, 2009.

[11] C. Laporte and T. Arbel. Efficient discriminant viewpoint selection for active bayesian recognition. *International Journal of Computer Vision*, 68(3):267–287, July 2006.

[12] C. Laporte, R. Brooks, and T. Arbel. A fast discriminant approach to active object recognition and pose estimation. In *ICPR*, pages III: 91–94, 2004.

[13] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. volume 77, pages 259–289, May 2008.

[14] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999.

[15] S. Maji and J. Malik. Object detection using a max-margin hough transform. In *CVPR*, pages 1038–1045, 2009.

[16] David Meger, Per-Erik Forssén, Kevin Lai, Scott Helmer, Sancho McCann, Tristram Southey, Matthew A. Baumann, James J. Little, and David G. Lowe. Curious george: An attentive semantic robot. *Robotics and Autonomous Systems*, 56(6):503–511, 2008.

[17] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In *CVPR*, pages 778–785, 2009.

[18] Lucas Paletta and Axel Pinz. Active object recognition by view integration and reinforcement learning. *Robotics and Autonomous Systems*, 31(1-2):71–86, 2000.

[19] S. Savarese and F. F. Li. 3D generic object categorization, localization and pose estimation. In *ICCV*, pages 1–8, 2007.

[20] Per Skoglar, Jonas Nygårds, and Morgan Ulvklo. Concurrent path and sensor planning for a UAV - towards an information based approach incorporating models of environment and sensor. In *IROS*, pages 2436–2442. IEEE, 2006.

[21] M. Sun, H. Su, S. Savarese, and L. Fei Fei. A multi-view probabilistic model for 3D object classes. In *CVPR*, pages 1247–1254, 2009.

[22] A. Thomas, V. Ferrar, B. Leibe, T. Tuytelaars, B. Schiel, and L. J. Van Gool. Towards multi-view object class detection. In *CVPR*, pages II: 1589–1596, 2006.

[23] Alexander Thomas, Vittorio Ferrari, Bastian Leibe, Tinne Tuytelaars, and Luc J. Van Gool. Using multi-view recognition and meta-data annotation to guide a robot's attention. *I. J. Robotic Res*, 28(8):976–998, 2009.