

A statistically Selected Part-Based Probabilistic Model for Object Recognition

Zhipeng Zhao , Ahmed Elgammal
Computer Science Department, Rutgers University
110 Frelinghuysen Road, Piscataway, NJ 08854-8019, U.S.A
{zhipeng,elgammal}@cs.rutgers.edu

Abstract. In an object recognition task where an image is represented as a constellation of image patches, often many patches correspond to the cluttered background. If such patches are used for object class recognition, they will adversely affect the recognition rate. In this paper, we present a statistical method for selecting the image patches which characterize the target object class and are capable of discriminating between the positive images containing the target objects and the complementary negative images. This statistical method select those images patches from the positive images which, when used individually, have the power of discriminating between the positive and negative images in the evaluation data. Another contribution of this paper is the part-based probabilistic method for object recognition. This Bayesian approach uses a common reference frame instead of reference patch to avoid the possible occlusion problem. We also explore different feature representation using PCA an 2D PCA. The experiment demonstrates our approach has outperformed most of the other known methods on a popular benchmark data set while approaching the best known results.

Key words: Computer vision, Pattern representation and modeling, Object detection, Class recognition, Feature selection

1 Introduction

Object detection and class recognition is a classical fundamental problem in computer vision which has witnessed much research. This problem has two critical components: the representation of the images (image features) and recognizing the object class using this representation which requires learning models of objects that relate the object geometry to image representation. Both the representation problem, which attempts to extract features which capture the essence of the object, and the following classification problem are active areas of research and have been widely studied from various perspectives. The methods for recognition stage can be broadly divided into three categories: the 3D model-based methods, the appearance template search-based methods, and the part-based methods. 3D model-based methods ([20]) are successful when we can describe accurate geometric models for the object. Appearance based matching approaches are based on searching the image at different locations and different scales for best match for object ‘template’ where the object template can be learned from training data and act as a local classifier [18, 15]. Such approaches are highly successful in modeling objects with wide within-class appearance variations such as in face detection [18,

15] but they are limited when the within-class geometric variations are large, such as detecting a motorbike.

In contrast, object recognition based on dense local “invariant” image features have shown a lot of success recently [8, 11, 14, 19, 1, 3, 6, 16, 7] for objects with large within-class variability in shape and appearance. In such approaches objects are modeled as a collection of parts or local features and the recognition is based on inferring object class based on similarity in parts’ appearance and their spatial arrangement. Typically, such approaches find interest points using some operator such as [9] and then extract local image descriptors around such interest points. Several local image descriptors have been suggested and evaluated, such as Lowe’s scale invariant features (SIFT) feature [11], entropy-based scale invariant features [9, 6] and other local features which exhibit affine invariance such as [2, 17, 13]. Other approaches that model objects using local features include graph-based approaches such as [5]. In this paper, we adopt a part-based method with a common reference frame. We also experiment with both PCA and 2D PCA [21] for image patch representation.

An important subtask in object recognition lies at the interface between feature extraction and their use for recognition. It involves deciding which extracted features are most suitable for improving recognition rate [19], because the initial set of features is large, and often features are redundant or correspond to clutter in the image. Finding such actual object features reduces the dimensionality of the problem and is essential to learn a representative object model to enhance the recognition performance. Weber *et al.* [19] suggested the use of clustering to find common object parts and to reject background clutter from the positive training data. In such approach large clusters are retained as they are likely to contain parts coming from the object. Similar approach has been used in [10]. However, there is no guarantee that large cluster will just contain only object parts. Since the success of recognition is based on using many local features, such local features (parts) typically correspond to low level feature rather than actual high level object parts. In this paper we introduce a statistical approach to select discriminative object parts out of a pool of parts extracted from the training images.

Contributions: The contribution of this paper is threefold. Firstly, we introduce a probabilistic Bayesian approach for recognition where object model does not need a reference part [6]. Instead object parts are related to a common reference frame. Secondly, we propose a novel approach for unsupervised selection of discriminative parts that finds features that best discriminate the positive and negative examples. Finally, we investigate PCA and 2D PCA for image patch representation in our experiment and did a comparison.

The organization of this paper is as follows. Section 2 describes our part-based probabilistic model, the recognition method and 2D PCA representation for image patch. Section 3 explains our statistical method for image patch selection. section 4 presents the results of applying the proposed methods on a benchmark dataset. Section 5 is the conclusion.

2 Part-based probabilistic model

We model an object class as a constellation of image patches from the object, which is similar in spirit to [19], but we also model their relative locations to a common reference frame. In doing this, we avoid the problem of not detecting the landmark patch. We assume objects from the same class should always have the same set of image patches detected and these image patches are similar both in their appearance and their relative location to the reference frame. The recognition of an object in an image will be a high probability event of detecting similar image patches sharing a common reference frame. In our work, we use the centroid as the reference frame and use the image patches simultaneously to build a probabilistic model for the object class and the centroid.

2.1 Model structure

The model structure is best explained by first considering recognition. Using m observed image patches v_k , ($k = 1, \dots, m$), from an image V , the problem of estimating the probability $P(O, C|V)$ of object class O and its centroid C given V can be formulated as (assuming independence between the patches and using Bayes' rule):

$$P(O, C|V) = \frac{P(V|O, C)P(O, C)}{P(V)} = P(O, C) \prod_{k=1}^m \frac{P(v_k|O, C)}{P(v_k)} \quad (1)$$

We wish to approximate the probability $P(v_k|O, C)$ as a mixture of Gaussian model using the observed patches from the training data. We simplify this by clustering all the patches selected from the training data into n clusters, A_i , $i = 1, \dots, n$ according to their appearance and spatial information, which is the 2D offset to the centroid C . We can decompose $P(v_k|O, C)$ as

$$P(v_k|O, C) = \sum_{i=1}^n P(v_k|A_i)P(A_i|O, C) = \frac{\sum_{i=1}^n P(v_k|A_i)P(O, C|A_i)P(A_i)}{P(O, C)} \quad (2)$$

Substituting (2) in (1), we get

$$P(O, C|V) \propto \prod_{k=1}^m \frac{\sum_{i=1}^n P(v_k|A_i)P(O, C|A_i)P(A_i)}{P(v_k)} \quad (3)$$

While performing recognition, $P(v_k)$ can be ignored. Assuming that $P(C)$ and $P(O)$ are independent, we have

$$P(O, C|V) \propto \prod_{k=1}^m \sum_{i=1}^n P(v_k|A_i)P(O|A_i)P(C|A_i)P(A_i) \quad (4)$$

2.2 learning

The task of learning is to estimate each term in (4) from the training data. We concatenate the image patches' appearance and spatial vectors as features in the image patches clustering process. Since the resulting clusters contain similar features, we can assume image patches from one cluster will follow normal distribution in both appearance and spatial subspaces. By calculating the sample mean and sample covariance matrix of the subspaces of these clusters, we can approximate the probability of v_k and C for each cluster $A_i, i = 1, \dots, n$. We use μ_i^v and μ_i^c to denote the sample means for v_k and C , respectively, and Σ_i^v and Σ_i^c to denote the sample covariances for v_k and C , respectively. Then for cluster A_i we have $P(v_k|A_i) \sim \mathcal{N}(v_k|\mu_i^v, \Sigma_i^v)$ and $P(C|A_i) \sim \mathcal{N}(C|\mu_i^c, \Sigma_i^c)$.

The rest of the terms in (4), can be approximated using the statistics from each of the cluster $A_i, i = 1, \dots, n$. If the Cluster A_i has n_i points of which n_{ij} belong to Class O_j , we can estimate the following: $P(A_i) = n_i / \sum_{i=1}^n n_i$ and $P(O_j|A_i) = n_{ij}/n_i$ ¹.

2.3 Recognition

Recognition proceeds by first detecting and selecting image patches, and then evaluating the probability of the event of detecting object features sharing a common reference frame, as described in section 2.1. By calculating the probability and comparing it to a threshold, the presence or the absence of the object in the image may be determined.

Equation 4 can be interpreted as a probabilistic voting where each patch gives a weighted vote for the object class and centroid given its similarity to each of the clusters. This formulation extends to handle scale variations by considering each pair of patches instead of each individual patch.

2.4 Image feature representation

The image patch feature concatenated from appearance and spatial information could be a high dimension vector. Usually PCA is applied to reduce the dimension while retaining much of the information. Recently Yang [21] has proposed 2D PCA for image representation. This method can easily evaluate the covariance matrix accurately to calculate the eigen vectors and also take less time. In this paper, we have experimented with both approaches and did a comparison.

3 Statistical image patch selection

In an object recognition task where an image is represented as a constellation of image patches, often many patches correspond to the cluttered background. If such patches are used to build the model for object class recognition, they will adversely affect the recognition rate. In this section, we proposed a statistical method for selecting those images patches from the positive images which, when used individually, have the power of discriminating between the positive and negative images in the evaluation data.

¹ It must be remarked that this model extends to modeling multiple object classes directly, however, since our problem consists of only one class, we have $P(O_j|A_i) = 1$.

We formulate the image patch selection problem in a statistical framework by selecting those images patches from the positive images which consistently appear in multiple instances of the positive images but only rarely appear in the negative images (barring some hypothetical and pathological cases). Intuitively, if an individual image patch from a positive image performs well in recognizing the images of the target object, a combination of a number of such image patches is likely to enhance the overall performance. This is because the individual classifiers, although weak, can synergistically guide the combined classifier in producing statistically better results.

Our approach is different from the Boosting method [16]. Boosting is originally a way of combining classifiers and its use as feature selection is an overkill. In contrast, our statistical method does not boost the previous stage but filters out the over-represented and undesirable clusters of patches corresponding to background. In spirit, our approach is similar to [4]. We formalize this intuitive statistical idea in the following straightforward yet effective method for selecting the characteristic image patches.

We select an image patch $v \in V^+$ from the positive images V^+ in the training data if it is able to discriminate between the positive and negative images in the evaluation data, $V_e = \{V_e^+, V_e^-\}$ with a certain accuracy. A complete description of this method requires describing the classification method using a single image patch and the accuracy threshold. For classifying an image $\mathcal{V} \in V_e$ in the evaluation set, using a single image patch $v \in V^+$, we first calculate the distance, $D(\mathcal{V}, v) = \min_{\nu \in \mathcal{V}} d(\nu, v)$, between \mathcal{V} and v defined as the euclidean distance between v and the closest image patch from \mathcal{V} . For classifying the images in the evaluation data, we use a threshold, t on distance $D(\mathcal{V}, v)$; if $D(\mathcal{V}, v) < t$, the image \mathcal{V} is predicted to contain the target object, otherwise not. Accordingly we can associate an error function, $\mathcal{E}r(\mathcal{V}, v, t)$ (defined below 5), which assumes a value 1 if and only if the classifier makes the mistake .

$$\mathcal{E}r(\mathcal{V}, v, t) = \begin{cases} 0, & \text{if } (D(\mathcal{V}, v) < t \wedge \mathcal{V} \in V_e^+) \vee \\ & (D(\mathcal{V}, v) \geq t \wedge \mathcal{V} \in V_e^-) \\ 1, & \text{otherwise} \end{cases} \quad (5)$$

Clearly, the performance depends on the parameter t , so we find an optimal circular region of radius t_v around v which minimizes the error rate of the classifier on the evaluation data. Finally, only those image patches from the positive images are selected which have recognition rate above a threshold, θ . A description of this algorithm, in the form of a pseudocode, is given in Algorithm 3.1. This algorithm takes the positive image patches V^+ , patches from the evaluation data V_e , and the threshold θ as input and outputs $\hat{H} \subseteq V^+$, the subset of selected image patches.

Algorithm 3.1: SELECT PATCHES, $\widehat{H}(V^+, V_e, \theta)$

```

 $\widehat{H} \leftarrow \emptyset;$ 
for each  $v \in V^+$ 
  for each  $\mathcal{V} \in V_e$ 
    do  $\left\{ \begin{array}{l} D(\mathcal{V}, v) = \min_{\nu \in \mathcal{V}} d(\nu, v); \\ t_v \leftarrow \arg \min_{t \in \mathbb{R}^+} \sum_{\mathcal{V} \in V_e} \mathcal{E}r(\mathcal{V}, v, t) \end{array} \right.$ 
  do  $\left\{ \begin{array}{l} err \leftarrow \frac{1}{|V_e|} \sum_{\mathcal{V} \in V_e} \mathcal{E}r(\mathcal{V}, v, t_v) \\ \text{if } (err < \theta) \\ \text{then } \left\{ \widehat{H} \leftarrow \widehat{H} \cup \{v\} \right. \end{array} \right.$ 

```

4 Experiment

4.1 Data Set

The experiment was carried out using Caltech database ². This database contains four classes of objects: motorbikes, airplanes, faces, car rear end which have to be distinguished from image in the background data set, also available in the database. Each object class is represented by 450 different instances of the target object, which were randomly and evenly split into training and testing images. Of the 225 positive images set aside for selecting the characteristic image patches, 175 were used as the training images and the remaining 50 were spared to be used as evaluation data. In addition, the evaluation data also consisted of 50 negative images from the background.

4.2 Image patch detection and the intensity representation

We use region based detector [9] for detecting informative image patches. We perform normalization for intensity and rescaled the image patches to 11×11 pixels, thus representing them as a 121 dimension intensity vectors. Then we tried with both PCA and 2D PCA on these vectors to get a more compact 18 dimension intensity representation.

4.3 Experimental Setting

We extracted 100 image patches for each of the 175 training images, and 100 evaluation images. Following this, we applied the statistical image patch selection method for removing the image patches from the background. In this process, we built simple classifier from each image patch in the training images and selected the one which led to a classifier with classification error rate less than 24%, an empirically calculated value. Figure 1 shows results from the image patches selection, which removes a significant number of patches corresponding to background.

² <http://www.vision.caltech.edu/html-files/archive.html>

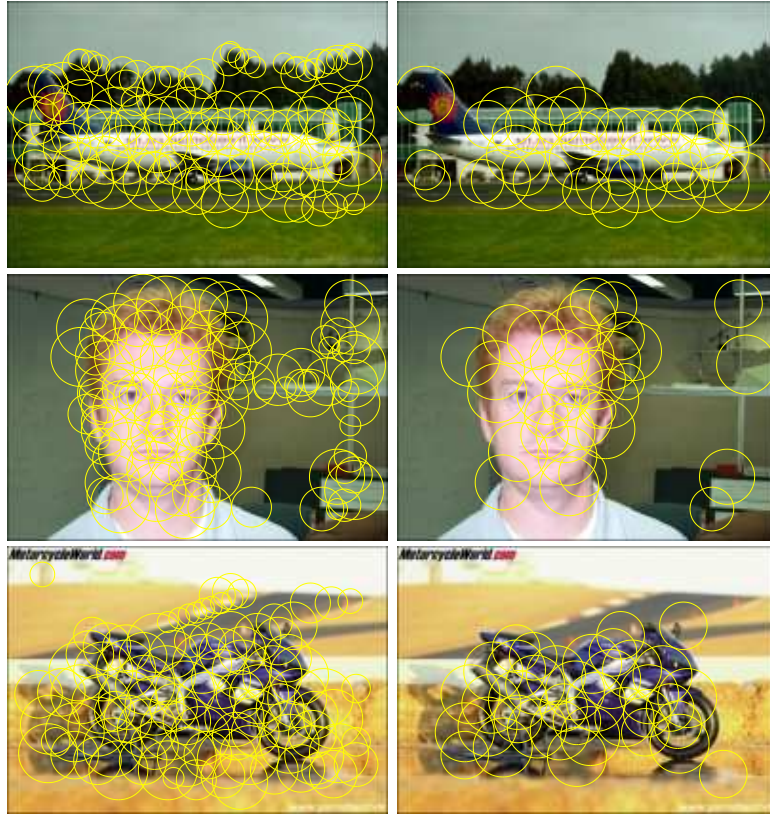


Fig. 1. Image patch selection. The image patches are shown using a yellow circle on the images. The left column shows the image patches extracted by Kadir & Brady's feature detector. The right column shows image patches selected by the statistical method.

After the image patch selection process, we computed the centroid for each object in the image. We used 2-D offset between the image patch and the object centroid as the spatial feature for the image patch and concatenated it with the intensity feature vector as the feature representation for each image patch. We then used k-means algorithm for clustering them into 70 clusters (this number was empirically chosen) and calculated the mean and covariance for them.

4.4 Experimental Results

In the testing phase, we used Kadir & Brady's[9] feature detector for extracting the image patches. Then we calculated the probability of the centroid of a possible object in the image as an indicator of its presence.

Figure 2 shows the computationally estimated frame for the object along with the image patches which contributed towards estimating this frame. Observe that the esti-

mated frame was mainly voted by the image patches located on the object. It also shows some examples of misclassification. There are two major reasons for such misclassification. The first is the presence of multiple target objects in the image, as shown in the airplane example. In this scenario, there is no centroid which gets a strong probability estimation from the matched parts. The second is poor illumination conditions which seriously limits the number of initial image patches extracted from the object, as illustrated by the face example.

We compared our result to the state of the art results from [6] and [12]. Table 1 summarizes the recognition accuracy at the equal ROC points (point at which the true positive rate equals one minus the false positive rate) of our different approach: no part selection with PCA, part selection with PCA, part selection with 2D PCA and results from other recent methods. This shows that the result from 2D PCA representation is similar that from PCA and our approach are comparable to other recent methods reporting equal ROC performance using this data set.

Dataset	No selection	statistical method	statistical method	Fergus	Opelt
	with PCA	with 2D PCA	with PCA	[6]	[12]
Airplane	54.2	95.8	94.4	90.2	88.9
Motorbike	67.8	93.7	94.9	92.5	92.2
Face	62.7	97.3	98.4	96.4	93.5
Car (rear)	65.6	98.0	96.7	90.3	n/a

Table 1. Equal ROC performance of our different approaches and other recent methods.

5 Conclusion

We have presented a statistical method for selecting informative image patches for patch-based object detection and class recognition. The experiments show our approach surpasses the performance of many existing methods. Although this method has been demonstrated in the context of image patch selection, it is a general method suitable for selecting a subset of features in other applications. A natural extension of this method is by integrating the auxiliary information regarding spatial arrangement between image patches; one way for doing this currently under investigation. In future, we intend to further develop and disseminate this framework as a general method for selecting features by automatically determining various hyper-parameter, which are currently empirically calculated.

References

1. S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *ECCV*, pages 113–130, 2002.
2. A. Baumberg. Reliable feature matching across widely separated views. pages 774–781.

3. E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *ECCV*, pages 109–124, 2002.
4. Gy. Dorkó and C. Schmid. Selection of scale-invariant parts for object class recognition. In *ICCV*, pages 634–640, 2003.
5. Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.
6. R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR (2)*, pages 264–271, 2003.
7. R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *CVPR*, 2005.
8. M. Fischler and R. Elschlager. The representation and matching of pictorial structures, 1973. *IEEE Transaction on Computer c-22(1)*: 67-92.
9. T. Kadir and M. Brady. Scale, saliency and image description. *IJCV*, 2001.
10. Thomas K. Leung and Jitendra Malik. Recognizing surfaces using three-dimensional textures. In *ICCV (2)*, pages 1010–1017, 1999.
11. David G. Lowe. Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision ICCV, Corfu*, pages 1150–1157, 1999.
12. Andreas Opelt, Michael Fussenegger, Axel Pinz, and Peter Auer. Weak hypotheses and boosting for generic object detection and recognition. In *ECCV (2)*, pages 71–84, 2004.
13. Frederik Schaffalitzky and Andrew Zisserman. Multi-view matching for unordered image sets, or "how do i organize my holiday snaps?". In *ECCV (1)*, pages 414–431, 2002.
14. Cordelia Schmid and Roger Mohr. Local grayvalue invariants for image retrieval. *IEEE PAMI*, 19(5):530–535, 1997.
15. H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. pages 45–51, 2000.
16. Antonio B. Torralba, Kevin P. Murphy, and William T. Freeman. Sharing visual features for multiclass and multiview object detection. In *CVPR*, 2004.
17. Tinne Tuytelaars and Luc J. Van Gool. Wide baseline stereo matching based on local, affinity invariant regions. In *BMVC*, 2000.
18. Paul Viola and Michael Jones. Robust real-time object detection. *International Journal of Computer Vision* 2002.
19. Markus Weber, Max Welling, and Pietro Perona. Unsupervised learning of models for recognition. In *ECCV (1)*, pages 18–32, 2000.
20. Haim J. Wolfson and Isidore Rigoutsos. Geometric hashing: An overview. *IEEE Computational Science & Engineering*, 4(4):10–21, /1997.
21. Jian Yang, David Zhang, Alejandro F. Frangi, and Jing-Yu Yang. Two-dimensional pca: A new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(1):131–137, 2004.

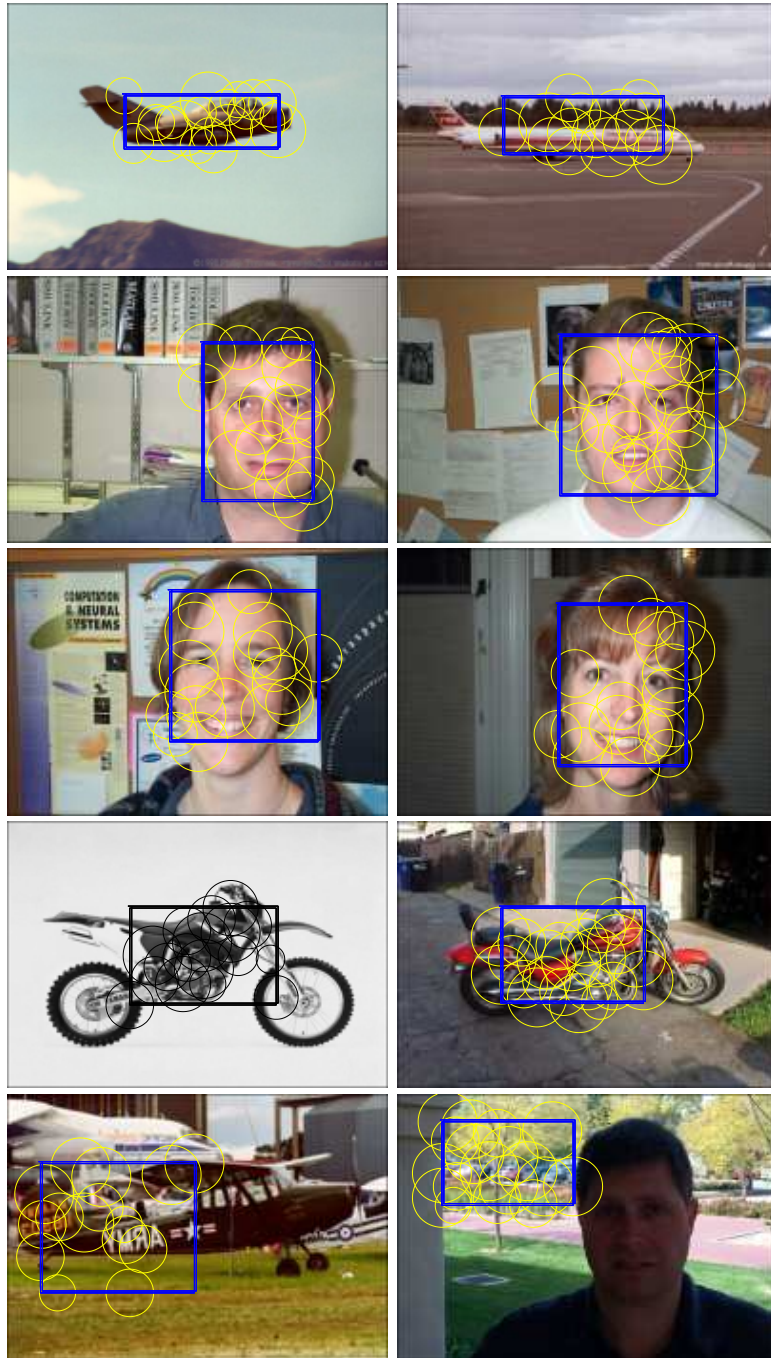


Fig. 2. This figure demonstrates the estimation of object frame in some typical testing image using statistical part selection. The estimated centroid is indicated by a rectangle. All the image patches contributed to this estimation are indicated by yellow circles. The bottom row of the images are some misclassification examples.