Semi-local Affine Parts for Object Recognition

Svetlana Lazebnik, Cordelia Schmid, Jean Ponce

Semi-Local Affine Parts for Object Recognition

Svetlana Lazebnik[1] Cordelia Schmid[2] Jean Ponce[1]
slazebni@uiuc.edu cordelia.schmid@inrialpes.fr ponce@cs.uiuc.edu
[1] Beckman Institute, University of Illinois, 405 N. Mathews, Urbana, IL 61801, USA
[2] INRIA Rhône-Alpes, 665 Avenue de l'Europe, 38330 Montbonnot, France

Abstract

This paper proposes a new approach for finding expressive and geometrically invariant parts for modeling 3D objects. The approach relies on identifying groups of *local affine regions* (image features having a characteristic appearance and elliptical shape) that remain approximately affinely rigid across a range of views of an object, and across multiple instances of the same object class. These groups, termed *semi-local affine parts*, are learned using correspondence search between pairs of unsegmented and cluttered input images, followed by validation against additional training images. The proposed approach is applied to the recognition of butterflies in natural imagery.

1. Introduction

Achieving true 3D object recognition is one of the most important challenges of computer vision. As a first step towards this goal, it is necessary to develop geometrically invariant object models that can support the identification of object instances in novel images in the presence of viewpoint changes, clutter, and occlusion. To date, object representations based on distinctive local image regions (interest points) have shown great promise for recognizing different views of the same object [3, 14, 15, 17] as well as different instances of the same object class [1, 2, 12, 20]. In the latter case, local regions play the role of generic *object parts* (e.g., eyes of a person or wheels of a car). In this paper, we propose a novel object recognition framework based on composite *semi-local affine parts*, or geometrically stable configurations of multiple elliptical *local affine regions*. We introduce methods for learning collections of such parts to represent 3D object classes, and for detecting part instances in test images. An important advantage of our learning method is that it is *weakly supervised*, i.e., it works with unsegmented, cluttered training images.

The parts proposed in this paper are *affinely rigid* by construction, i.e., the mapping between two different instances of the same part can be well approximated by a 2D affine transformation. Note that we do not make the overly restrictive assumption that the entire object is planar and/or rigid — it is sufficient for the object to possess some (approximately) planar and rigid components. Because of this non-global notion of affine invariance, our method is suitable for modeling a wide range of 3D transformations, including viewpoint changes and non-rigid deformations. This exceeds the capabilities of most existing part-based category-level recognition schemes [1, 2, 12, 20], which are suited primarily for recognizing fronto-parallel views of objects.

The mechanism for learning semi-local affine parts, which is described in Section 2, is based on the idea that a direct search for visual correspondence is key to successful recognition. Thus, at training time we seek to identify groups of neighboring local affine regions whose appearance and spatial configuration remains stable across multiple instances. To avoid the prohibitive complexity of establishing simultaneous correspondence across the

whole training set, we separate the problem into two stages: Parts are initialized by matching pairs of images and then matched against a larger *validation set*. Even though finding optimal correspondence between features in two images is still intractable [6], effective sub-optimal solutions can be found using non-exhaustive constrained search. The promise of the proposed framework is demonstrated in Section 3 with an application to the automated acquisition and recognition of butterfly models in heavily cluttered natural images. Finally, Section 4 closes with a discussion of major conceptual issues raised in the paper.

2. Learning Semi-Local Affine Parts

This section presents the method for automatically identifying collections of semi-local affine parts to represent 3D object classes. The first step is feature extraction (Section 2.1), which consists of detection of local affine regions followed by computation of appearance descriptors. Next, *candidate parts* are formed by matching several pairs of training images (Section 2.2), and a validation step is used to discard spurious matches (Section 2.3).

2.1. Feature Extraction

Detecting local affine regions. We use an affine-adapted Laplacian blob detector based on [5]. While a few other affine- and scale-invariant detectors are available in the literature [9, 15, 16, 19], we chose the Laplacian because it finds perceptually salient blob-like regions that tend to be centered away from object boundaries. This detector finds the locations in scale space where a normalized Laplacian measure attains a local maximum and then applies an *affine adaptation* process (see [5, 16] for details). The elliptical regions found as a result of this process can be *normalized* by mapping them onto a unit circle. However, the normalizing transformation has an inherent *orthogonal ambiguity*, since the unit circle is invariant under rotation and flipping. We resolve this ambiguity by representing the appearance of each normalized patch by rotation-invariant descriptors.

Descriptors. In this work, we use two descriptors which complement each other by relying on different types of image information: *spin images* [8], which are based on normalized intensity values; and *RIFT* descriptors, which are based on gradient orientations. For details about descriptor computation, see [10].

An *intensity-domain spin image* is a two-dimensional histogram with bins indexed by two parameters: The first is d, the distance from the center of the patch, and the second is i, the intensity. Thus, the "slice" of the spin image corresponding to a fixed d is simply the histogram of the intensity values of pixels located at a distance d from the center. In our implementation, we use ten bins each for d and i, resulting in 100-dimensional descriptors. To achieve invariance to affine transformations of the intensity, we normalize the range of the intensity function within the support region of the spin image.

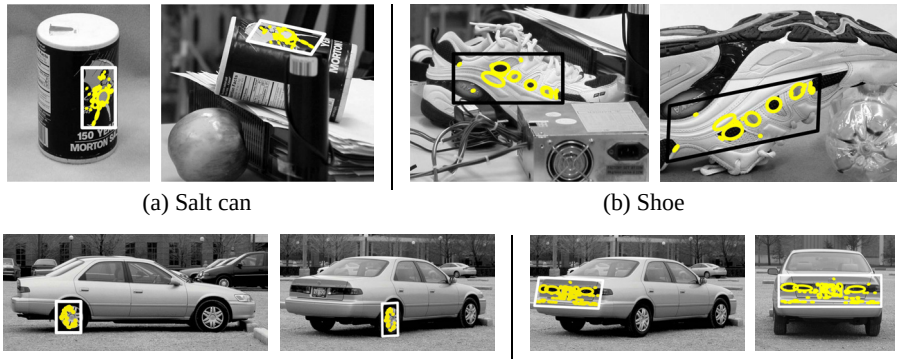
The representation of local appearance of a normalized patch is augmented with an additional *RIFT* descriptor, which is a rotation-invariant generalization of *SIFT* [14]. The *RIFT* descriptor is constructed as follows. The circular normalized patch is divided into concentric rings of equal width, and within each ring, a gradient orientation histogram is computed. To maintain rotation invariance, this orientation is measured at each point relative to the direction pointing outward from the center. We use four rings and eight histogram orientations, yielding 32-dimensional descriptors (the original SIFT has 128 dimensions). Note that the RIFT descriptor as described above is not invariant to reflection of the normalized patch, which reverses the order of directions in the orientation

histogram. Thus, when finding the distance between two RIFT descriptors, we must take the minimum over both orders.

The final issue is how to combine the two kinds of descriptors in determining the appearance-based dissimilarity or *matching score* between two patches. We set the matching score to be the *minimum* of the Euclidean distances between the two spin images or RIFT descriptors (note that both descriptors are normalized to have zero mean and unit norm, thus Euclidean distances between them lie in the same range and are comparable). Empirically, this approach performs better than other ways of combining descriptors, since it provides robustness against instabilities in region extraction and intensity normalization. The descriptors (particularly spin images) can be sensitive to transformations of the intensity values (i.e., noise, JPEG compression, sharpening) and to shifts in the position of the center of the normalized patch, so a large distance between two spin images or two RIFT descriptors is not always a reliable indication of perceptual difference. Thus, in determining the matching score between two patches, it makes sense to trust the descriptor that produces the lower distance.

2.2. Finding Candidate Parts

In this section, we describe the procedure for initializing local affine parts, which is based on determining correspondence between sets of regions in the two images. The space of all hypotheses is exponentially large, necessitating the use of constrained search [6] with strong geometric and appearance-based consistency constraints to prune this space.

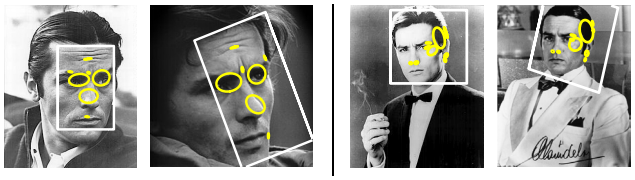


(a) Salt can (b) Shoe

(c) Car

Figure 1: Candidate parts for three 3D objects: (a) a salt can, (b) a shoe, and (c) a car. The two input images to the matching procedure are shown side by side, with the matched ellipses superimposed. For visualization purposes, we also show bounding boxes around the matched ellipses: the axis-aligned box in the left image is mapped onto the parallelogram in the right image by the affine transformation that aligns the matched ellipses.

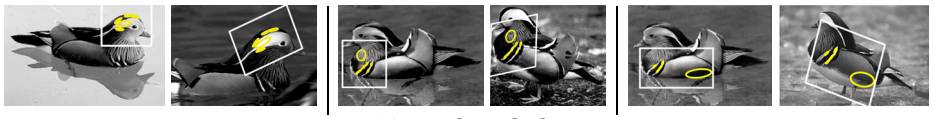
Correspondence search begins by identifying a set of neighboring (or *seed*) triples of regions (ellipses) in the first image. The neighborhood of an ellipse is defined by "growing" the axes of that ellipse by a constant factor (typically 2 to 4). Given an ellipse with index i and an ordered set of neighbor indices, we generate a fixed number of seed triples of the form (i, j, k), where (j, k) are all combinations of the first m (2 to 5) neighbors. For each seed triple (i, j, k) in the first image, we find all triples (i', j', k') in the second image such that i' (resp. j', k') is a potential appearance-based match of i (resp. j, k), and j' and k' are neighbors of i'. Potential matches are determined using a threshold on



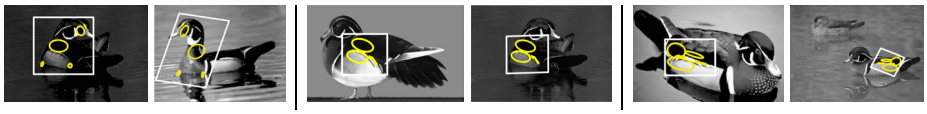
(a) Alain Delon

(b) Grace Kelly (c) Grace Kelly and Cary Grant

Figure 2: Candidate parts for face images. (c) Curiously, the face of Grace Kelly in the left image fails to match to her face in the right image, but the matched regions do reveal a structural similarity between her face and Cary Grant's.



(a) Mandarin duck



(b) Wood duck

Figure 3: Candidate parts for birds. Note that the hypotheses are confined to relatively rigid sections of the body (e.g., head, neck, breast, wing).

the matching score between the respective descriptors. Note that the total number of seed triples of matches is $O\left(n\binom{m}{2}r^3\right)$, where r is the maximum number of potential matches of a single ellipse and n is the total number of ellipses in the first image. This quantity is actually linear in n since m and r have fixed upper bounds in the implementation.

The next step is to judge the *geometric consistency* of each triple of matches. To do this, we determine the affine transformation mapping the three ellipse centers in the first image onto their putative correspondences in the second image. Once the two sets of ellipses have been aligned in the same coordinate system, their shape can be compared directly and consistency can be measured using thresholds on the difference of major and minor axis sizes, and the angle between major axes. Note that we cannot set an absolute threshold on the first quantity, which is measured in pixels and is therefore non-invariant. Instead, the threshold is defined as a fraction of a quasi-affine *local scale*, in our case, the average of the major axis lengths of all ellipses in the current hypothesis.

After finding a geometrically consistent triple of matches, the search algorithm attempts to extend it into a larger consistent hypothesis by searching for additional matches lying in the neighborhood of any of the original ellipses. We use a greedy strategy, where at each step, the most geometrically consistent pair of potential matches is selected, and the affine transform between corresponding ellipse centers is re-estimated using linear least squares. With four or more matches, the residual from this estimation is used as the

(a) Palace window: symmetry

(b) Badgers: symmetry (c) Badgers: repetition

Figure 4: Examples of symmetry and repetition detection in single images. (a) Note that because the geometry of this scene is not planar, multiple *local* symmetries are detected. (b) Reflective symmetry of the left badger's head. (c) A match between the heads of the two individuals.

measure of geometric consistency of the entire spatial configuration. The process continues until no match can be added without violating the consistency constraints. Note that the size of the final hypothesis is an indicator of its saliency. We have found that at least 6 to 8 matches are needed to represent a non-spurious correspondence. Therefore, we discard all hypotheses smaller than a certain minimum size. As a final step, we also merge hypotheses that overlap by a significant number of regions.

Because it does not assume a *global* affine correspondence between two images, our matching procedure is applicable to general 3D objects, as demonstrated by the examples of Figure 1. In the case of cars, consistent hypotheses arise from structurally important areas such as wheels and the license plate. Figures 2 and 3 show several matching hypotheses for faces and birds, which are non-rigid objects.

Interestingly, the above search procedure can be used to match an image to itself, and is thus applicable to the problem of detecting repeated structures and symmetries within an image [13, 18] (Figure 4). The only change necessary in the implementation is the addition of checks to prevent trivial hypotheses, i.e., hypotheses that match every region to itself. Note that our approach can discover patterns despite substantial clutter in the image, which is not possible with other, more specialized approaches, e.g., [13].

2.3. From Hypotheses to Parts

The next step is to convert a matching hypothesis into a unique representation for a *candidate semi-local affine part*. Informally, we form a part by "averaging" the two sets of regions brought into correspondence by the hypothesis. Given two sets of ellipse centers $\{x_i \leftrightarrow x_i'\}$, we want to find two point sets $\{\hat{x}_i\}$ and $\{\hat{x}_i'\}$ that are exactly related by an affine transformation, such that the "Procrustean" distance $\sum_i (x_i - \hat{x}_i)^2 + (x_i' - \hat{x}_i')^2$ is minimized. The solution for $\{\hat{x}_i\}$ and $\{\hat{x}_i'\}$ is easily obtained using 2D affine factorization [4]. Note that our representation derives its invariance from alignment, so it is not necessary to have an invariant representation of coordinates. Thus, either of the two affinely equivalent coordinate sets $\{\hat{x}_i\}$ or $\{\hat{x}_i'\}$ can be used to represent the geometric configuration of the part. Next, pairs of corresponding ellipses are registered in the chosen coordinate system, and their attributes (descriptors, axis lengths, orientations) are averaged.

After initializing a candidate part as described above, we *validate* it by matching it to additional sample images of the object. The purpose of validation is to reject spurious

hypotheses, or, less drastically, to remove individual correspondences arising from clutter. The process of detecting a part in a new image is much the same as two-image matching, made simpler and more efficient by the part's relatively small size and (presumed) lack of clutter. Note that the detections of the part in a validation image are allowed to have missing regions to account for occlusion or failure of the region detector. We define the *repeatability* of an individual region as the proportion of hypotheses in which the region was detected. Regions with repeatability below a certain threshold are rejected. Similarly, a repeatability score is defined for parts as the average number of detected regions per hypothesis. This score provides us a way to rank parts according to their "quality," so that we can take a fixed number of top parts to serve as a "vocabulary" for representing the object class, or simply to discard parts whose repeatability falls below a certain threshold.

Note that in principle, we can use the correspondences between the parts and the validation images to improve the parts by modifying their appearance and shape. At present, we have not implemented this extension, which would involve factorization in the presence of missing data. Moreover, in our experiments we have generally been satisfied with the quality of parts obtained following verification.

3. Recognition Experiments

In this section, we exercise the proposed part extraction method for the challenging application of identifying butterflies in natural imagery. We use an extremely simple recognition framework, so that the burden for achieving good performance is placed entirely on the expressiveness and invariance of semi-local affine parts. Briefly, the matching and validation procedures described in the previous section are used to identify a fixed-size collection of parts for representing the classes. We define a cumulative repeatability score that combines all part detections for a given class in a test image, which enables us to evaluate performance either using multi-class classification or binary detection.

Figure 5 shows a dataset composed of 619 images of seven classes of butterflies. The pictures, which are collected from the Internet, are extremely diverse in terms of size and quality. Motion blur, lack of focus, resampling and compression artifacts are common. This dataset is appropriate for exercising the descriptive power of local affine parts, since the geometry of a butterfly is locally planar for each wing (though not globally planar). In addition, the species identity of a butterfly is determined by a basically stable geometric wing pattern, though appearance can be significantly affected by variations between individuals, lighting, and imaging conditions. It is crucial to point out that butterfly recognition is beyond the capabilities of many current state-of-the-art recognition systems [1, 2, 12, 20]. For example, the system developed by Fergus et al. [2] is, according to its authors, limited to models consisting of up to 6 or 7 features learned from images containing 20 to 30 features. By contrast, a typical butterfly pattern is sufficiently complex to require at least a dozen regions to be adequately represented, while the clutter is measured by hundreds or even thousands of regions. Moreover, the levels of invariance (translation and scale) possessed by existing algorithms are clearly insufficient for recognizing butterflies, which can and do appear at a wide range of scales and orientations, and are rarely fronto-parallel with respect to the camera.

Candidate parts are formed by matching between eight randomly chosen pairs of training images. Ten verification images per class are used to rank candidate parts according to their repeatability score, and top ten parts per class are retained for recognition. Figure 7 (a) shows the part having the highest repeatability for each of the classes. At testing time,

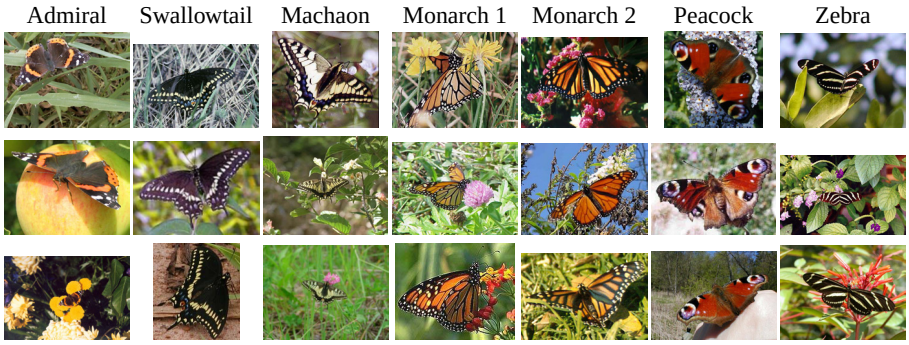


Figure 5: The butterfly dataset. Three samples of each class are shown in each column.

(a)

Class	Part size	Test images	Correct (rate)
Admiral	179 (12/28)	85	74 (0.871)
Swallowtail	252 (18/29)	16	12 (0.750)
Machaon	148 (12/21)	57	55 (0.965)
Monarch 1	289 (14/67)	48	35 (0.729)
Monarch 2	275 (19/36)	58	53 (0.914)
Peacock	102 (8/14)	108	108 (1.000)
Zebra	209 (16/31)	65	58 (0.892)
Total		437	395 (0.904)

(b)

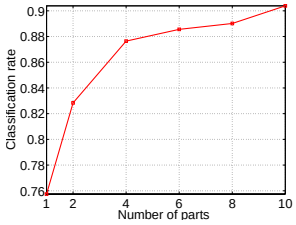


Figure 6: Classification results for the butterflies. (a) The second column shows the total model size for each class (the sum of sizes of individual models), and the size of the smallest and the largest models are listed in parentheses. (b) Classification rate vs. number of parts.

the parts for all classes are detected in each training image. Though multiple instances of the same part may be found, we retain only the single instance with highest number of detected regions. Figure 7 (b) shows examples of part detections in individual test images. The cumulative score for a given class is given by the *relative repeatability* of all its parts, or the total number of regions detected in all parts divided by the sum of part sizes. For multi-class classification, each image is assigned to the class having the maximum relative repeatability score. Figure 6 (a) shows classification results obtained using the above approach (the average rate is 90.4%), and Figure 6 (b) shows how performance is improved by using multiple parts.

We can get an alternative assessment of performance by considering the binary detection task, where for each image and each class, we ask whether an instance of the from this class is present. This decision can be made by setting a threshold on the relative repeatability. By considering all possible thresholds, we get an *ROC curve*, a plot of the true positive rate vs. the false positive rate. These curves, given in Figure 7 (c), also show ROC *equal error rates* (false positives $= 1-$true positives). The true positive rates range from 87% to 94.8%, showing that detection can indeed be performed successfully.

4. Discussion

In this paper, we have presented a weakly supervised framework for modeling 3D objects in terms of geometrically invariant *semi-local affine parts*. The two-image matching procedure that forms the core of our method is also applicable to identifying repeated structures and symmetries within a single image — an interesting application which is rarely treated in the same context as recognition. For our primary goal of 3D object recognition, the proposed approach has the advantages of robustness and flexibility. Namely, it is capable of learning multiple variable-sized parts from images containing a significant

amount of noise and clutter. We conclude this presentation with a discussion of several conceptual issues relevant to our work.

Probabilistic vs. geometric approaches. Recently, Bayesian approaches have shown considerable promise for weakly supervised learning of part-based models [2, 12, 20]. However, while the generative framework provides a principled way of modeling intra-class variability, it is not ideally suited for designing geometrically invariant representations. In particular, a rigorous probabilistic treatment of affine invariance is quite daunting [11]. In our own research, we have taken a direct geometric approach to invariance, thus greatly gaining in simplicity and flexibility. For example, our search method can automatically determine the number of regions in a semi-local affine part, something that is not straightforward to achieve in a probabilistic framework.

EM vs. unique correspondence. Another practical limitation of Bayesian methods comes from the use of the *Expectation Maximization* (EM) algorithm to estimate model parameters. EM treats correspondence as missing data to be integrated out, which in principle involves computing expectations over the exponentially large space of all possible correspondences. Despite the use of various approximations, the combinatorics of EM severely limits the size of the model and the amount of clutter that can be tolerated during learning. By contrast, our approach is built on the idea that establishing *unique* correspondence between model and image features is central for successful recognition.

Alignment. The search algorithm described in Section 2.2 is reminiscent of the alignment techniques used in model-based vision [6, 7]. While the process of detecting an existing part in a test image may indeed be thought of as affine alignment, our overall modeling approach follows a different strategy. Whereas in classical alignment globally rigid models are built manually and/or from segmented/uncluttered images, our method is capable of handling heavily cluttered input since it does not seek a *global* transformation between two images, nor does it assume that the entire object is either planar and/or (affinely) rigid. Instead, it exploits the fact that smooth surfaces are planar in the small, and that semi-local affine parts are sufficient to handle large viewpoint variations for approximately coplanar, close-by patches, as well as small non-rigid transformations.

Training set size. Our procedure for initializing semi-local affine parts uses only two images. Recently, it has been observed that very few training images are actually necessary for learning of object models provided the learner is equipped with a good prior on the parameters of the model [12]. In our case, the "prior" is the strong notion of visual similarity, making it possible to learn candidate parts from pairs of input images.

Modeling the background. Most object detection schemes, whether classifier-based (discriminative) or probabilistic (generative), require an explicit model of the background and a "negative" training set. By contrast, our approach avoids these requirements because of its reliance on strong geometric consistency constraints and the implicit assumption that the background is non-repeatable.

Relations between parts. As a major extension of our method, which in its current form stops short of *category-level* recognition, we plan to develop a method for reasoning about the *spatial relations* between parts. Although we currently do learn multiple parts for the butterfly dataset, we do not define any relations between them. In this case, relations are not necessary because multiple butterfly parts are redundant, i.e., they all represent wings. In this case, one part is (in principle) sufficient for successful recognition, and the cumulative scoring scheme based on relative repeatability of *all* parts is appropriate. However,

for more complicated 3D classes, parts may correspond to different structures (e.g., body parts), and certain parts may only be detected for some poses of the object, or only for some instances of the class. Thus, developing inter-part relations will be necessary for recognizing non-rigid and/or articulated objects such as animals and humans.

Acknowledgments. This research was partially supported by the National Science Foundation under grant IIS-0308087, the European project LAVA (IST-2001-34405), the UIUC-CNRS collaboration agreement, and the Beckman Institute for Advanced Science and Technology.

References

[1] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *Proc. ECCV*, volume 4, pages 113–130, 2002.

[2] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. CVPR*, volume 2, pages 264–271, 2003.

[3] V. Ferrari, T. Tuytelaars, and L. Van Gool. Simultaneous object recognition and segmentation by image exploration. In *Proc. ECCV*, 2004.

[4] A. Fitzgibbon and A. Zisserman. On affine invariant clustering and automatic cast listing in movies. In *Proc. ECCV*, volume 3, pages 304–320, 2002.

[5] J. Gårding and T. Lindeberg. Direct computation of shape cues using scale-adapted spatial derivative operators. *IJCV*, 17(2):163–191, 1996.

[6] W. E. L. Grimson. The combinatorics of object recognition in cluttered environments using constrained search. *AIJ*, 44(1-2):121–166, 1990.

[7] D. P. Huttenlocher and S. Ullman. Object recognition using alignment. In *Proc. ICCV*, pages 102–111, 1987.

[8] A. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Trans. PAMI*, 21(5):433–449, 1999.

[9] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *Proc. ECCV*, 2004.

[10] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. Technical Report CVR-TR-2004-01, Beckman Institute, University of Illinois, 2004. Available at `http://www-cvr.ai.uiuc.edu/ponce_grp`.

[11] T. Leung, M. Burl, and P. Perona. Probabilistic affine invariants for recognition. In *Proc. CVPR*, pages 678–684, 1998.

[12] F.-F. Li, R. Fergus, and P. Perona. A Bayesian approach to unsupervised one-shot learning of object categories. In *Proc. ICCV*, 2003.

[13] Y. Liu, R. Collins, and Y. Tsin. A computational model for periodic pattern perception based on frieze and wallpaper groups. *IEEE Trans. PAMI*, 26(3):354 – 371, March 2004.

[14] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[15] J. Matas, O. Chum, U. Martin, and T Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proc. BMVC*, volume 1, pages 384–393, 2002.

[16] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. ECCV*, volume 1, pages 128–142, 2002.

[17] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3d object modeling and recognition using affine-invariant patches and multi-view spatial constraints. In *Proc. CVPR*, volume 2, pages 272–277, 2003.

[18] A. Turina, T. Tuytelaars, T. Moons, and L. Van Gool. Grouping via the matching of repeated patterns. In *Proc. CAPR*, pages 250–259, 2001.

[19] T. Tuytelaars and L. Van Gool. Matching widely separated views based on affinely invariant neighbourhoods. *IJCV*, 59(1):61–85, 2004.

[20] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proc. ECCV*, volume 1, pages 18–32, 2000.

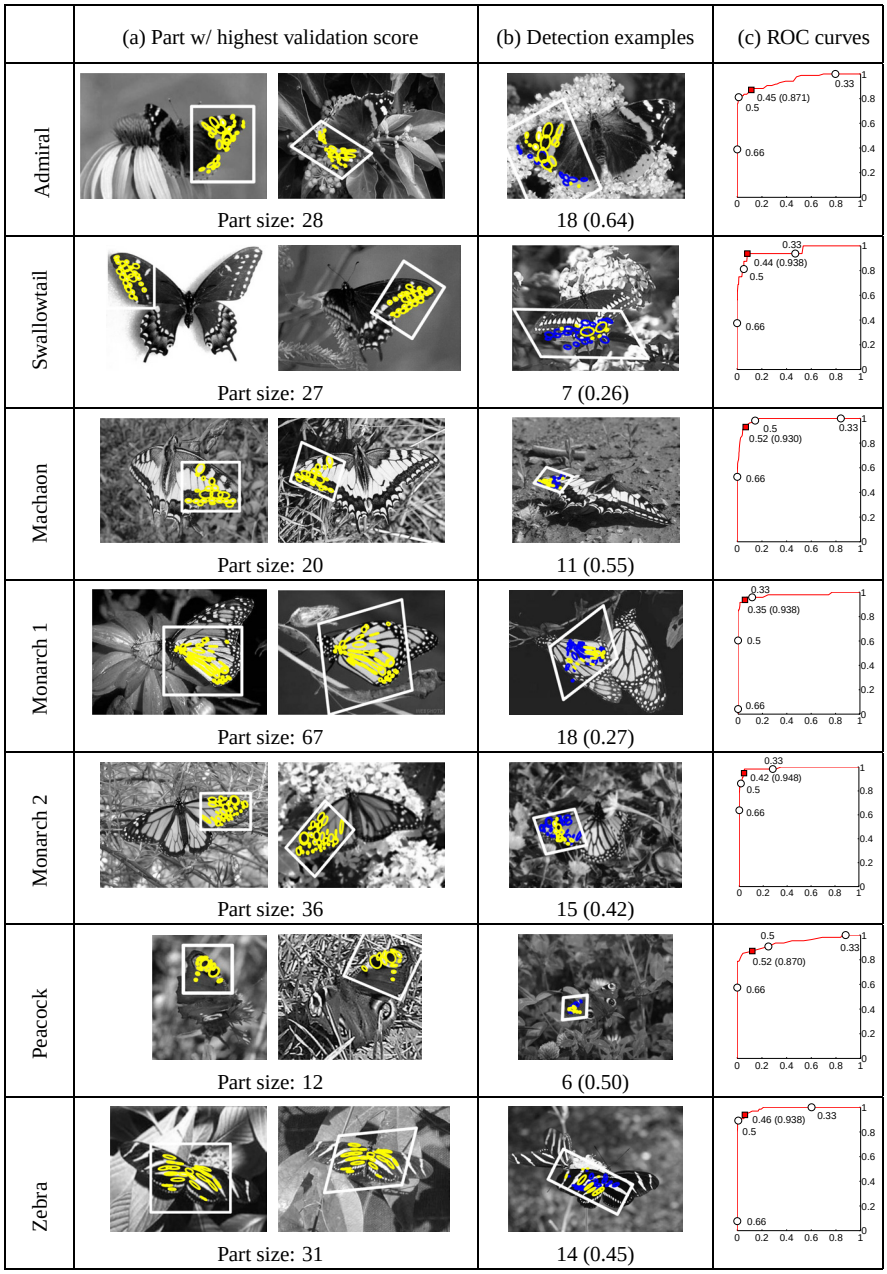


Figure 7: Butterfly modeling and detection examples. (a) The part with the highest validation score for each class. The part size is listed below each modeling pair. (b) Example of detecting the part from (a) in a single test image. Detected regions are shown in yellow and occluded ones are reprojected from the model in blue. The total number of detected regions (absolute repeatability) and the corresponding repeatability ratio are shown below each image. Note that for the swallowtail and zebra detections the correspondences between part and image regions are incorrect. (c) ROC curves for detection. Three different thresholds for relative repeatability (0.33, 0.5, 0.66) are marked on the curve. The dark square marks the ROC equal error rate, which is listed in parentheses next to the threshold value at at which it is attained.