Generic Object Recognition via Shock Patch Fragments

Özge C. Özcanlı and Benjamin B. Kimia
Division of Engineering
Brown University, Providence, RI, USA
{ozge, kimia}@lems.brown.edu

Abstract

We propose a new methodology to partition a natural image into regions based on the shock graph of its contour fragments. We show that these regions, or *shock patch fragments*, are often object fragments, thus effecting a partial segmentation of the image. We utilize shock patch fragments to recognize objects with dominant shape cues eliminating the need to segment out the entire object from the image first. Our preliminary results with minimal training are promising with respect to the state of the art recognition systems.

1 Introduction

There has been a paradigm shift in computer vision in the past decade moving away from relying on fully segmented images for recognition and other visual tasks, to using a collection of features that capture appearance and shape in a small area. The key point underlying this paradigm shift is the availability of a new generation of feature detectors such as Harris-Affine, Harris-Laplace and others [15] and a new generation of feature descriptors such as SIFT [13] and others [14] that are more stable to viewpoint variation, lighting change, *etc*. The main idea is that while these features may not remain present under all variations, given the sheer number of features, the common presence of a few discriminative features can discriminate between the presence or absence of a particular, previously observed object in an image.

In contrast to object instance recognition, generic object recognition, where the intra-category type variations must in addition be accounted for, has proven to be more challenging. Approaches to generic object recognition which are based on key feature detectors and descriptors span a continuum between two extremes. On one extreme, the spatial relationship between parts is represented such as in the constellation model [6, 4] and the k-fan model [3], and these are referred to as "part-based" models. On the other extreme which completely discards the spatial relationship, the approach relies on an unorganized collection of features which are coded in a lower dimensional vocabulary of visual words, or a codebook of appearance parts common to a collection of images, and are known as "bag of words" models. The former "part based" approach is faced with a combinatorial search arising from an exponential number of correspondences. The "bag of words" approach avoids the combinatorial difficulties [29, 5, 28, 35], but is more brittle to situations when some of the features have been removed, *e.g.*, due to partial occlusion. In approaches between these extremes, the geometric relationships between neighboring

BMVC 2007 doi:10.5244/C.21.104

features is modeled using specified geometric transformations [21, 22, 33], or using lose pairwise relationships [1, 34].

The success of the above approaches indicates the significance of the role of *appearance* in discriminating the presence or absence of objects in typical scenes where bottom-up segmentation could not conceivably segment figure from ground. However, these approaches are also limited in several significant ways. The chief drawback is that stable key features are not always available to the degree of abundance that the approach relies on. For example, the intensity variation in the interior of an object viewed in low lighting condition, *e.g.*, a cow at dusk/dawn, or equivalently against bright backgrounds, *e.g.*, a bird against a bright sky, is too limited to produce reliable key features. In this case the silhouette is a more reliable cue for recognition. As another example, objects especially man-made objects, may feature large homogeneous and therefore featureless areas. Similarly, cartoons, sketches, and line drawings which are readily recognizable would have no appearance-based key features. In these cases, the edge content is the sole information source for recognition. Finally, objects in low resolution imagery, *e.g.*, aerial video images of vehicles, where the total extent of the object is of the same order as that required for feature descriptors (25x25) [19], cannot be recognized using this approach.

A second short coming of the use of appearance-based key features for recognition is that the role of appearance may become severely diminished as the size of the database grows. This happens when the type variation increases the range of appearances on object category captures. For example, in recognizing bottles and cups, the surface markings are simply too varied to be useful [17]. In this case, the edge content of the silhouette and of the internal markings consistent across the category become the primary source of information for recognition.

A number of recent approaches to object recognition rely on the edge content of category, as represented in an unorganized edge map, or in a collection of curve fragments. As an example of use of edge maps in recognition, Belongie *et al.*'s [2] shape context approach assigns a signature to each edge representing the radial-polar histogram of other edges. This signature is sufficiently discriminative to enable correspondence and a similarity score after an image transformation. As an example of an approach relying on contour fragments, Nelson and Selinger [16, 24], motivated by the cubist approach to evoking the visual percept of form from a few fragmentary cues, modeled contour fragment maps by a collection of local context patches (21x21) which are normalized for size and orientation with respect to a centrally placed key curve. Fergus *et al.* [7] use segments of extended edge chains lying between bitangent points in their constellation model. Kumar *et al.* [12] used contours as a component in a graph-based pictorial structures. In the Boundary Fragment Model (BFM), a boundary fragment codebook is constructed by clustering those which are highly class-distinctive and predictive of the object centroid over a set of training data [17, 25].

A significant disadvantage of the above approaches is that either the relative spatial distribution of various contour fragments in an object is ignored altogether, or it is captured through the mediation of an object model, *e.g.*, requiring an object centroid. The lack of the relative spatial relationship among contour fragments restricts the discriminability of each fragment. The requirement of an object model to mediate the spatial relationship between fragments renders the approach sensitive to partial occlusion. For example, if only the head of a horse or a cow is visible, individual contour fragments for the head can match a large number of fragments from other objects, an effect which

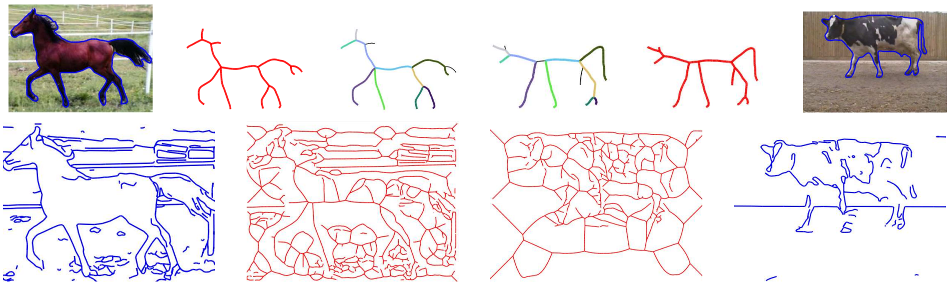

Figure 1: Top row illustrates how the shock graph of a horse is optimally transformed to the shock graph of a cow (colored edges are matching and the thinner black edges are editted out). While for segmented images this can be done with a polynomial time algorithm, the algorithm for matching the shock graphs in the bottom row is NP-complete (blue: boundaries, red: shock graphs).

becomes more significant as the size of the database increases. The spatial relationship among pairs of contour fragments on the head, on the other hand, make the joint-pair of fragments highly selective.

The main goal of this paper is to use the joint representation of a pair of contour fragments for recognition. The medial axis is a structure for the joint representation of pairs of contour fragments and our paper is focused on the use of this structure in the form of a shock graph as described below. The only previous work which takes advantage of pairs of contour fragments in such a "localized" sense is that of Jurie and Schmid [9] where edges are detected at multiple scales and annular regions are rated for the extent of significant non-accidental edge support on a wide range of angles around the region, see also [10]. The annular regions are localized over position and scale and used as distinctive and discriminative shape features. However, these shape features do not make use of the geometry of the curve fragments beyond the presence or absence in the small portions falling in the thin annular regions.

Our work builds on the success of shock graphs as a representation for generic object recognition from segmented images [23, 27]. Shock graph is a variant of the medial axis of the contour map of an image and it is obtained by viewing the medial axis as the locus of singularities (shocks) formed in the course of wave propagation (grass-fire) from boundaries [11, 26, 32]. The resulting shock graph is a richer descriptor of the contour map than the medial axis graph and it is a good intermediate representation since its nodes and edges signify presence of contour pairs and triplets, gaps and T-junctions. Loops in the graph signify groups of edges. See Figure 2c and 3c for example shock graphs of two contour sets. The use of the shock graph captures much of the intra-class object variability since articulation and metric variations in the part shape often leave the structure intact, while partial occlusion only affects parts of it. Those changes that lead to structural changes in the shock graph are captured in the context of considering deformation paths encoded by shock transitions. The precision-recall rates for large number of categories is excellent [23]: for a database of 1032 shapes roughly organized in 40 categories, the leave-one-out recognition rate is at 97% and drops to 82% for the last member of the category.

Generalizing the above approach from recognition of objects in segmented images to those in real images requires first that edge maps be represented by a shock graph, and furthermore that perceptual grouping operations, like removing an edge from the map or closing a gap, be represented as a sequence of shock transitions, *e.g.* as described in [30]. The key difficulty is that the shock graph which for segmented images is a tree and which therefore leads itself to a polynomial-time edit distance algorithm, Figure 1, is no longer a tree due to the presence of spurious edges and gaps. Thus, matching two shock graphs faces combinatorial explosion in the search space and becomes NP-complete. This is very much similar to the combinatorial search space of constellation models, which limits its use to simpler models. In a similar vein, the complexity of the shock graph edit-distance algorithm for edge maps in real images motivates the recognition of smaller portions of objects, or object fragments. Our approach, therefore, probes the presence of an object in a limited portion of the shock graph, as determined by a collection of subgraphs. Each shock subgraph models a patch of the image which we refer to as a *shock patch fragment*. The recognition of these object fragments is the basis of our approach to object category recognition. While the use of shock object fragments based on "shock patch fragments" enables the use of both shape and appearance cues, we focus on shape features in this paper. However, it is not difficult to construe how a region descriptor of the sort used for key feature description can be used for shock patch regions to augment the shape aspects with appearance.

The paper is organized as follows. Section 2 describes how an edge map is processed to produce a collection of contour fragments from which a shock graph is obtained, Figures 2 and 3. In Section 3, we explain extraction of shock patches and show examples. In Section 4 we describe the procedure to match shock patch sets for object recognition and we report our results on object detection task in Section 5.

2 Contour Fragments

The success of any method based on shock graphs of curve fragments heavily depends on the reliability and stability of image contours. We now explain the processing stages of our approach: edge detection, edge linking, and perceptual grouping.

Edge Detection and Edge Linking: We use a pair of recently proposed edge detection and edge linking algorithms [31] which robustly extract well-localized sub-pixel edges and stably links these into curve fragments. For edge detection, it advocates the use of a third-order edge detector with an extremely low threshold, to get as many edges as possible so that the linking stage has enough options to choose from. Since the low threshold creates many spurious curve fragments, we prune these after linking by thresholding a measure consisting of both length and color contrast in the LAB space as used in [20],

$$C_a = 1 - e^{\frac{-||\mu_{R^+} - \mu_{R^-}||}{\gamma_{app}}} \tag{1}$$

where μ_{R^+} and μ_{R^-} are the mean colors of regions on either side of the curve fragment, and $||.||$ is the L_2 distance in R^3. We set $\gamma_a = 14$. If a color image is not available we find the L_1 distance of the appearance means. Figure 2b and 3b shows the curve fragments resulting from this process using a length threshold of 2 pixels and a color contrast threshold of 0.5 with a support region width of 5 pixels.

Gap Closure: The shock graph of the resulting curve fragments is computed using the method in [30], see Figures 2c and 3c. There are numerous gaps and spurious curve fragments which interfere with the process of forming shock segments correspond to object fragments. The shock graph provides a clue to the existence of these elements and

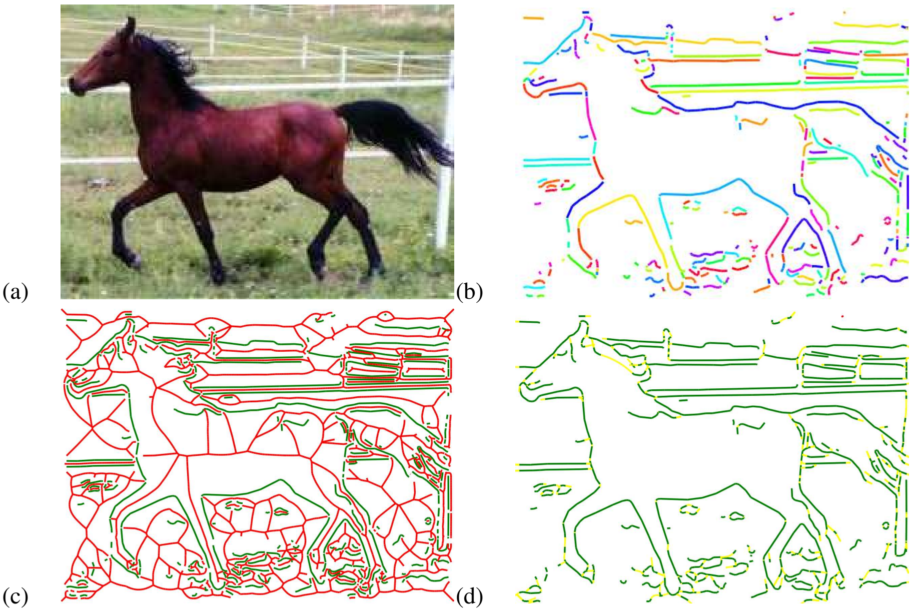

Figure 2: (a) An example image; (b) Curve fragments after pruning based on length and color contrast; (c) Shock graph of the curve fragment set (d) Curve fragments after the gap transform where the fragment set is shown in green and the gap completions in yellow.

transformations of it can be used to effect gap closure (*gap transform*) and the removal of spurious elements (*loop transform*). Specifically, observe that waves propagating radially from curve-ends meet, they form *degenerate* shocks in the shock graph, when they meet with normal waves propagating parallel to the contours, they form *semi-degenerate* shocks [8]. These edges signify gaps and possible T-junctions, respectively, in the curve fragment map of the image. See Figure 4a for an example of each kind. The gap transform is based on closing gaps and forming T-junctions by considering each case as rank-ordered by a measure reflecting both *(i)* good contour continuity and *(ii)* appearance discontinuity. The results are shown in Figures 4c, 2d and 3d.

3 Shock Patch Extraction

We now explore the notion of forming recognizable and stable image fragments which in effect are hypotheses for partial segmentations of the image. Assuming that the fragments have detectable boundaries, they must be anchored on curve fragments. Since a single curve fragment is not sufficiently distinctive, multiple contour fragments should be used to define image fragments. Since each pair of adjacent contour fragments give rise to a shock segment, selecting shock subgraphs provides a mechanism for selecting a group of curve fragments. Specifically, given a particular node in the shock graph, we traverse neighboring nodes in a depth-first manner to extract subgraphs at various depths. Since each shock segment typically describes a pair of curve fragments and the portion of the image in-between, we refer to this as a *visual fragment* Figure 5a, the shock subgraph describes an image fragment, which we refer to as the *shock patch fragment*, Figure 5e. The

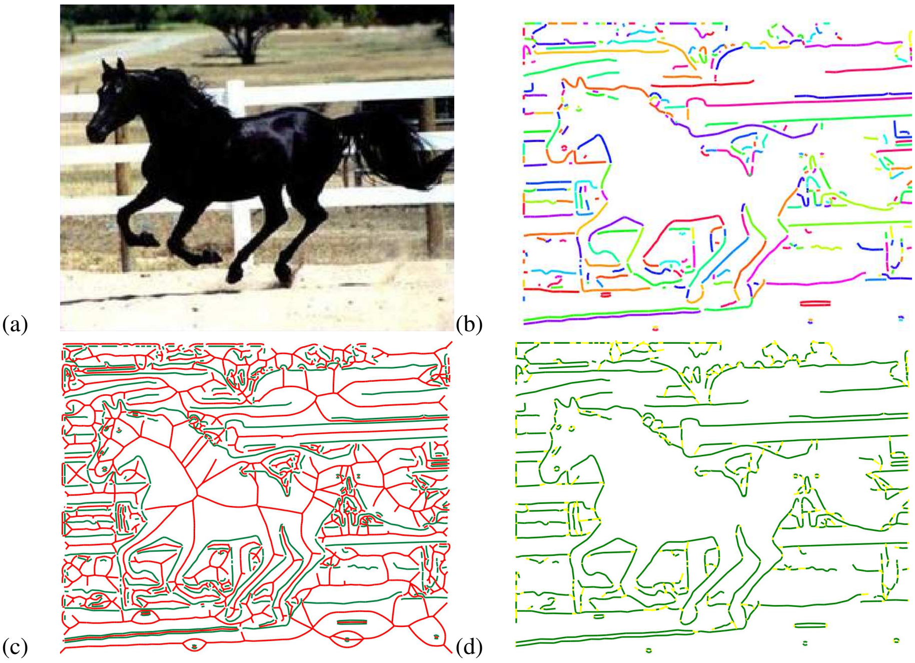

Figure 3: The same steps in Figure 2 are shown for another horse image.

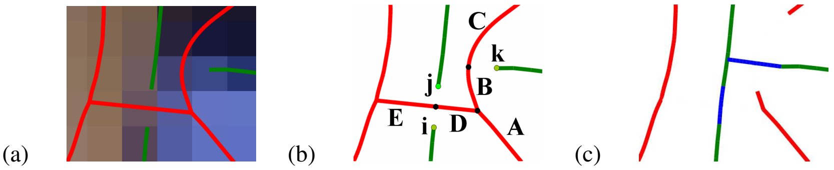

Figure 4: (a) Image curves shown in green and shock graph in red (b) D, E and A are degenerate edges, suggesting the closure of (i-j), (j-i) and (i-k) respectively. B and C are semi-degenerate edges suggesting to form a T-junction from k to the contour. (b) Completion curves in blue after the gap (i-j) is closed and a T-junction is formed based on the closure criteria.

boundary of this region is partially detected curve fragments shown in blue in Figure 5d, and partially by virtual contours imposed by end-nodes, shown in yellow. Figure 6a shows four subgraphs of increasing depths for a selected node on a real image example. Observe that when the subgraph contains a loop, *e.g.* due to a spurious edge, the fragment boundary does not contain this inner boundary, effectively removing it from consideration.

The shock patch fragments then consist of an outer contour as well as an appearance of the inner region. Each shock graph node produces shock patch fragments at all depths 1,2, *etc*. This collection of shock patch fragments is highly redundant, since adjacent nodes produce very similar fragments and since fragments from the same node but at different depths are similar. Furthermore, low-depth patches are often not very informative. Therefore, we subsample depths $(d_1, d_2, \ldots, d_n)$ = (6, 9, 12, 15, 18), and use the extent of overlap to remove similar patches generated by nearby nodes. All patches with 80% or more overlap are considered equivalent, and represented by the patch with the highest appearance contrast. This reduces the number of fragments from thousands to about 30-100

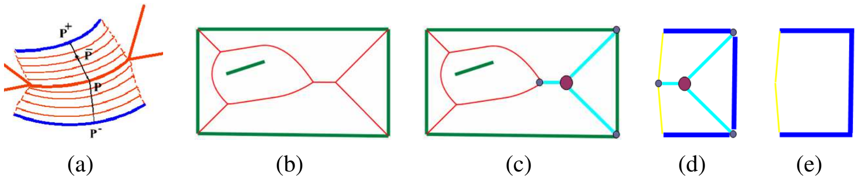

Figure 5: (a) Visual fragment (b) A simple shape with boundaries in green and shock graph in red (c) A subgraph at depth 1 (d) Induced boundaries in blue, virtual boundaries in yellow (e) Shock patch fragment.

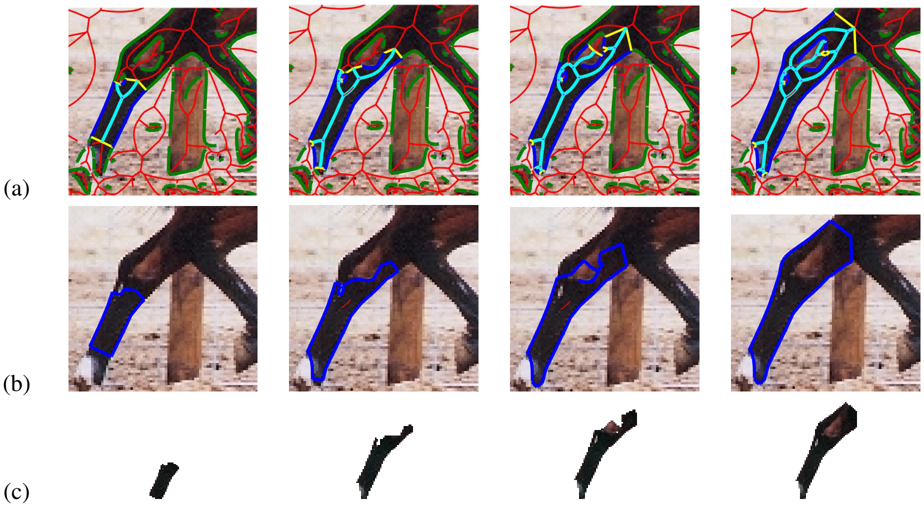

Figure 6: (a) Shock subgraphs at depths 1, 2, 3 and 4, respectively. The shock graph is shown in red and the subgraph in light green, image boundaries are shown in green, shock patch boundaries in blue. (b) shows the simple closed boundary in blue traced from the outer face of the subgraph. (c) Four shock patch fragments.

per image, as shown for the horse examples of Figures 2 and 3 in Figure 7.

4 Object Matching and Detection using Shock Patches

Shock patch fragments can depict either object fragments, effectively implementing a partial figure-ground segmentation, they can be pieces of the background, or object combined with the background, for example in Figure 7 some fragments depict meaningful object parts, *e.g.*, the horse head, limb, torso, *etc.*, while others do not clearly map to a distinguishable part. When we compare the two horse images, we do not expect any similarities between the second type of fragments, while we do expect some similarity between the head, limb, torso *etc.* between the two sets, and this can be confirmed in Figure 8.

Our approach therefore relies on finding similar fragments between the two sets. Fragment similarity can be measured by comparing the shape and appearance of the two fragments. As tempting as it is, we have excluded appearance from our current fragment similarity measure, both to explore the limits of a shape-based measure, and also because fragment appearance similarity is very well explored elsewhere in the patch-based object

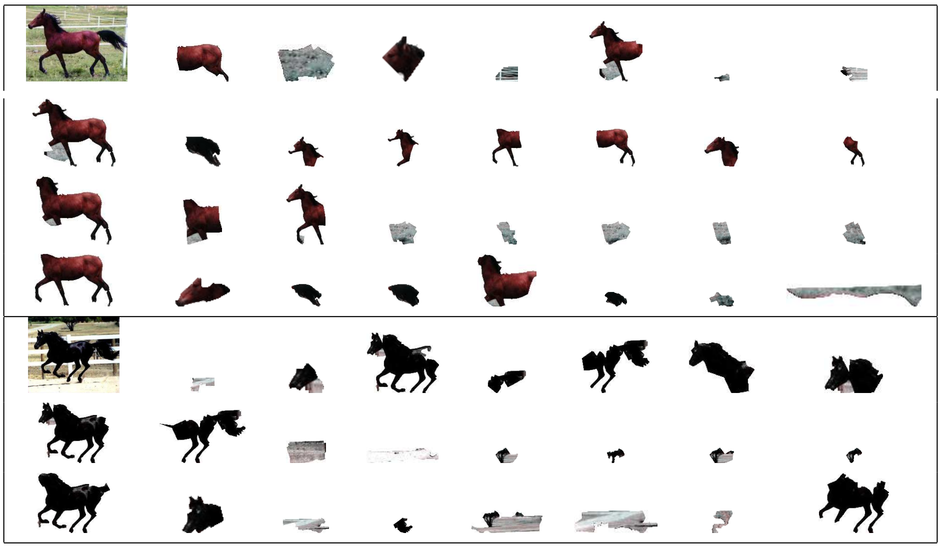

Figure 7: Example patches with depths (6, 9, 12, 15, 18) from the example horses. Observe that all major body parts are covered.

recognition work in the form of local descriptors [14]. We expect that the addition of appearance would improve our recognition results.

Fragment shape similarity should be measured using an algorithm that is robust against occlusions. This is because one can view the fragments as partial occlusions of the figure, *i.e.*, when a horse's torso is compared to a horse. In addition, it must capture intra-class shape variations very well. We therefore use the approach proposed by Sebastian *et al.* [23] which uses an edit-distance algorithm based on shock transitions [8] which handles both well. Figure 8 shows some example matches and non-matches.

5 Results

We propose an object detection and classification algorithm using only a few segmented object images as the training set (one in the case of this paper). We tested our system on the Horse-side class used in [18] consisting of 88 horse images and 88 background images. First, to explore the strength of the matching algorithm we used a single shock patch obtained from the silhouette mask of one of the images as the training set, shown in the second row of Figure 8. We matched all the test image patches to our model patch and declare a detection if top 3 matches are below a given similarity threshold and return the detection box to be the union of top 3 image patches. See Figure 8 for the top 5 matches of some test images. An object is deemed correctly detected if the overlap of the bounding boxes (detection vs ground truth) is greater than 50%. Our recall with best threshold settings is 75% with a precision of 85%. There are two reasons for the low recall rate, one is the use of a single model leading to large deviation, *e.g.* as the pose varies, and the other is that the partial matching of a single fragment to a full model degrades as the ratio of fragment area to the full model decreases. Observe from Figure 8 bottom row that the head of the example horse is correctly matching to the head of the training image

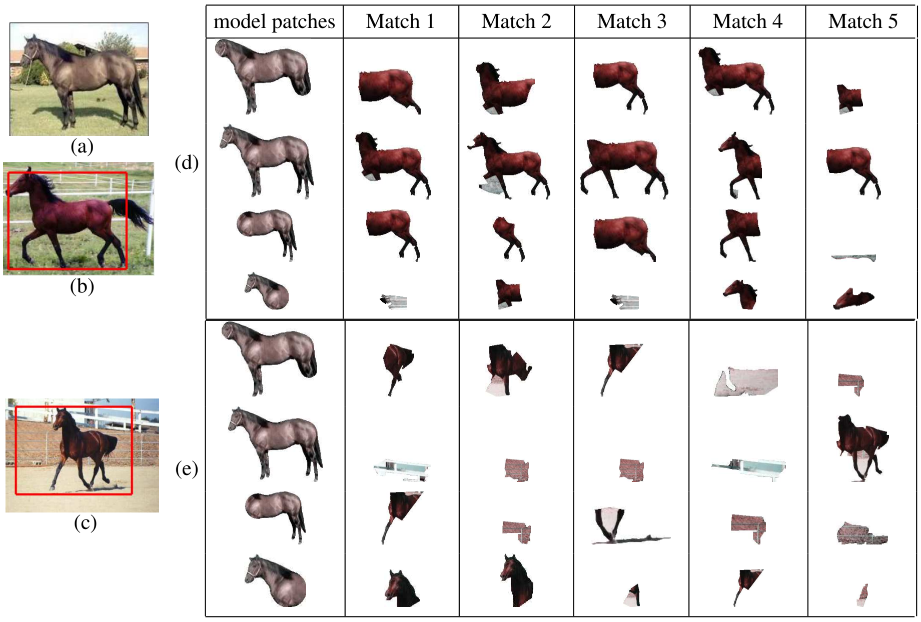

Figure 8: This figure illustrates the similarity between the model horse (a) fragments to fragments from two other horse images (b, c), as measured by the shock graph edit distance [23]. The detection boxes outputted by our system are superimposed on the test images shown in (b, c).

despite the pose difference, but there is not sufficient shape content to match against a full model. This motivates replacing the model by the model shock patch fragments. With this modification and the constraint that at least two model patches' top 3 matches should be within the similarity threshold, recall rate at the best threshold settings increases to 92%, with 85% precision. This second test image is classified correctly in this setting.

In conclusion, we have presented a shape based object detection and classification system which does not require involved training/learning stages and which has promising results on a difficult test set. These results can be improved by making use of spatial constraints which are naturally imposed by the shock topology of the training image, *e.g.*, head patch should be detected in correct relative position and orientation with respect to the torso patch *e.t.c*. Technical enhancements such as the implementation of the loop transform to further clean the curve fragment set, the inclusion of a few more training examples, and the use of appearance in the fragment similarity computation, all should lead to improvements in the recognition rate. Our main contribution in this paper is to present a novel method to generate fragments of images and illustrate their use in a generic object recognition and detection task.

Acknowledgements: This material is based on work supported by the National Science Foundation under Grant No. 0413215.

References

[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Trans. on Pattern Anal. Mach. Intell.*, 20(11):1475–1490, 2004.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, 2002.

[3] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *CVPR*, San Diego, CA, 2005.

[4] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, 28(4):594–611, 2006.

[5] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *CVPR*, pages 524–531, 2005.

[6] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, Madison, WI, 2003.

[7] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *ECCV*, 2004.

[8] P. J. Giblin and B. B. Kimia. On the local form and transitions of symmetry sets, medial axes, and shocks. *IJCV*, 54(Issue 1-3):143–157, August 2003.

[9] F. Jurie and C. Schmid. Scale-invariant shape features for recognition of object categories. In *CVPR*, 2004.

[10] M. F. Kelly and M. D. Levine. Annular symmetry operators: A method for locating and describing objects. In *ICCV*, 1995.

[11] B. B. Kimia, A. R. Tannenbaum, and S. W. Zucker. Shapes, shocks, and deformations, I: The components of shape and the reaction-diffusion space. *IJCV*, 15(3):189–224, 1995.

[12] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Extending pictorial structures for object recognition. In *BMVC*, pages 789–798, 2004.

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

[14] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630, 2005.

[15] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. J. V. Gool. A comparison of affine region detectors. *IJCV*, 65(1-2):43–72, 2005.

[16] R. C. Nelson and A. Selinger. A cubist approach to object recognition. In *ICCV*, pages 614–621, 1998.

[17] A. Opelt, A. Pinz, and A. Zisserman. A boundary-fragment-model for object detection. In *ECCV*, 2006.

[18] A. Opelt, A. Pinz, and A. Zisserman. Incremental learning of object detectors using a visual shape alphabet. In *CVPR*, 2006.

[19] O. C. Ozcanli, A. Tamrakar, B. B. Kimia, and J. L. Mundy. Augmenting shape with appearance in vehicle category recognition. In *CVPR*, 2006.

[20] M. A. Ruzon and C. Tomasi. Edge, junction, and corner detection using color distributions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1281–1295, 2001.

[21] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or "how do i organize my holiday snaps?". In *ECCV*, volume 1, pages 414–431, 2002.

[22] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE PAMI*, 19(5):530–535, 1997.

[23] T. Sebastian, P. Klein, and B. Kimia. Recognition of shapes by editing their shock graphs. *PAMI*, 26:551–571, May 2004.

[24] A. Selinger and R. C. Nelson. A perceptual grouping hierarchy for appearance-based 3d object recognition. *Computer Vision and Image Understanding*, 76(1):83–92, 1999.

[25] J. Shotton, A. Blake, and R. Cipolla. Contour-based learning for object detection. In *ICCV*, 2005.

[26] K. Siddiqi and B. B. Kimia. A shock grammar for recognition. In *Proc. CVPR*, pages 507–513, 1996.

[27] K. Siddiqi, A. Shokoufandeh, S. J. Shokoufandeh, and S. W. Zucker. Shock graphs and shape matching. In *ICCV*, pages 222–229, 1998.

[28] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their localization in images. In *ICCV*, pages 370–377, 2005.

[29] E. B. Sudderth, A. B. Torralba, W. T. Freeman, and A. S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, pages 1331–1338, 2005.

[30] A. Tamrakar and B. B. Kimia. Medial visual fragments as an intermediate image representation for segmentation and perceptual grouping. In *Proc. of POCV*, page 47, 2004.

[31] A. Tamrakar and B. B. Kimia. No grouping left behind: From edges to curve fragments. In *sub. to ICCV*, 2007.

[32] H. Tek and B. B. Kimia. Symmetry maps of free-form curve segments via wave propagation. In *ICCV*, 1999.

[33] D. Tell and S. Carlsson. Combining appearance and topology for wide baseline matching. In *ECCV*, 2002.

[34] M. Vidal-Naquet and S. Ullman. Object recognition with informative features and linear classification. In *ICCV*, pages 281–288, Nice, France, 2003.

[35] G. Wang, Y. Zhang, and L. Fei-Fei. Using dependent regions for object categorization in a generative framework. In *CVPR*, volume 2, pages 1597–1604, 2006.