PICTORIAL STRUCTURES FOR OBJECT RECOGNITION AND PART LABELING IN DRAWINGS

Amir Sadovnik and Tsuhan Chen

Department of Electrical and Computer Engineering, Cornell University

ABSTRACT

Although the sketch recognition and computer vision communities attempt to solve similar problems in different domains, the sketch recognition community has not utilized many of the advancements made in computer vision algorithms. In this paper we propose using a pictorial structure model for object detection, and modify it to better perform in a drawing setting as opposed to photographs. By using this model we are able to detect a learned object in a general drawing, and correctly label its parts. We show our results on 4 categories.

***Index Terms*—** sketch recognition, pictorial structures, object detection

1. INTRODUCTION

Sketch recognition and computer vision algorithms attempt to solve similar problems in different domains. For example, the tasks of object detection and localization in computer vision are closely related to the task of interpreting strokes as certain objects in drawings. In both, the goal is to take a 2D image and identify the existence and position of a previously learned object. However, many sketch recognition algorithms do not try to utilize the advancements made in the computer vision field on drawings. This is mainly because drawings are analyzed using ink data, and therefore there has been a lot of focus on developing algorithms which take advantage of this representation of the image.

In this paper we approach the task of object recognition in sketches using a similar approach to Sharon et al. [1]. Although we use the ink data from the images, we are able to integrate ideas from Felzenszwalb et al. which detects objects in photographs, and modify them for drawings [2]. Because of the basic differences between object appearance in photographs vs. drawings, these modifications include the transition from pixel space to scribbles and the selection of more appropriate features. We also account for multiple scales of objects in a single image. This combination of ink data together with part-based-detection allows us to solve more complicated problems than previous works on a wider variety of drawings.

More specifically, our contribution is the use of pictorial structures for object detection in drawings. The input to the algorithm is a general drawing made up of separate strokes (Fig. 1(a)). In each of the drawings, one or more objects of different catagories are drawn, in addition to strokes which do not belong to any known object. An object detector for each category is then run on the drawing. The output of each detector is a new image, in which each stroke is labeled as part of the object, or as background (Fig. 1(b)). Each detected object also receives a confidence score. Objects with low confidence are removed from the output.

There are many uses for this type of algorithm. First, it can be used in sketch retrieval applications. For example, using the algorithm we would be able to search for sketches which include a drawing of a house. In addition, we can use it to perform semantic segmentation, in which we retrieve only the strokes which are relevant to a certain object.

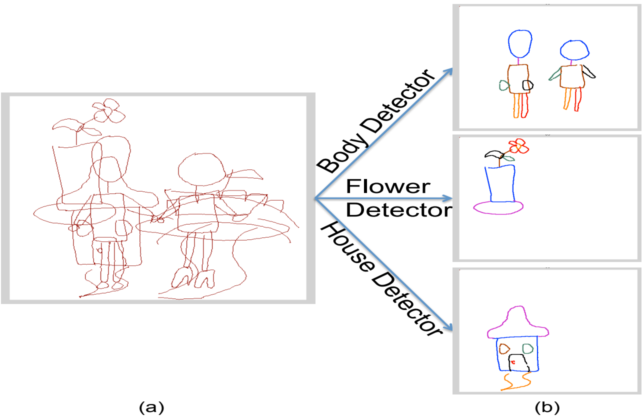


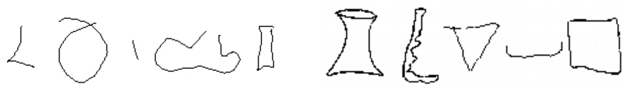
Fig. 1. Example of the input (a) and output (b) of the algorithm. The color legend for each catagory can be found in table 1. Note that background strokes are successfully removed by our algorithm. (best viewed in color)

The rest of the paper is organized as follows. In section 2 we provide an overview of related works. Section 3 describes the modified pictorial structure for drawings. Section 4 describes our implementation of the model. Section 5 describes the experiment and results. We conclude in section 6.

2. RELATED WORK

There are many different algorithms for sketch recognition which use ink data (for a short survey on the different approaches see [3]). Most of these algorithms try to decompose the sketch into known symbols from a certain vocabulary based on some similarity measures. This approach makes sense for the task of interpreting graphical diagrams in which each symbol has a specific shape. For example, when interpreting an electrical circuit, the shape of each component leads to its semantic label [4]. Another approach is that of object recognition using shape contexts [5]. Although this method attempts to recognize objects which have a more flexible appearance, it still relies heavily on the shape of the object.

However, in free hand drawings these approaches would not be sufficient, since strokes with similar semantic labels may have very different shapes (see Fig. 2). Therefore in order to interpret these symbols, a more global approach is necessary in which inter-stroke relationships are taken into account. Sharon et. al. use this information to label parts of a known object using a similar approach to ours [1]. However, we differ from their approach in that we use the model to perform multiple object detection as opposed to only part



(a) nose examples (b) flower pot examples

Fig. 2. Example of intraclass variance in collected drawings

labeling. In addition, we reduce the complexity of the algorithm by using a tree structure in our graphical model.

Other works have attempted different ways in which to deal with a more global approach. For example, Hall et al. use a combinatorial approach [6]. Sezgin et al. use temporal data [7]. Although they both use the entire sketch to label each symbol, they do not use any learned spatial relationships between the strokes. Qi et al. do use spatial information, but they perform a binary labeling problem using a CRF [8]. Mackenzie et al. use an agent based approach as opposed to a probabilistic approach, and rely on a predefined grammar [9].

Our work is an extension of Felzanszwalb et al. [2]. Although there are many other approaches for object detection in images this part based approach seemed the most suitable to use in drawings. First, drawings are usually constructed by parts. For example, when drawing a face one would usually draw the outline of each part separately. In addition, the algorithm allows to put a lot of weight on the relationships between parts vs. their actual appearances. This is crucial for our task since appearance cannot be counted upon.

3. THE PICTORIAL STRUCTURE

We approach the problem of detection and localization of an object in a similar manner to Felzanszwalb et al. where the object we are trying to detect can be expressed in terms of an undirected graph $G = (V, E)$ [2]. $V = \{v_1, ..., v_n\}$ are the n parts that need to be identified, and an edge $(v_i, v_j) \in E$ describes the relationship between the two parts. Given a sketch $S = \{s_1, ..., s_m\}$ where each s_j is a stroke, our goal is to find $L = \{l_1, ...l_n\}$, where each l_i is an assignment of a stroke s_j to a part v_i.

We use a statistical setting, where the $p(L|S, \theta)$ is the probability of assignment L given the sketch S and a set of parameters θ. Assuming all drawing are equiprobable we can use Bayes' rule and write this posterior probability as,

$$p(L|S, \theta) \propto p(S|L, \theta)p(L|\theta) \qquad (1)$$

Our set of parameters θ will include three subsets $\theta = (U, E, C)$. $U = \{u_1, ...u_n\}$ describe the visual aspect of each part regardless of other strokes. This can include different shape descriptors such as rectangularity, length to height ratio, etc. E is the set of edges which specifies which parts are connected, and $C = \{c_{ij}|(v_i, v_j) \in E\}$ are the parameters for each relationship. These relationships can include position differences, size differences, etc. Using the same method as in [2] we can write the likelihood and prior as

$$p(S|L, U) \propto \prod_{i=1}^{n} p(S|l_i, u_i) \qquad (2)$$

$$p(L|E, C) = \prod_{(v_i, v_j) \in E} p(l_i, l_j|c_{ij}) \qquad (3)$$

Finally we can plug equations 2 and 3 into 1 and arrive at our final probability

$$p(L|S, \theta) \propto \prod_{i=1}^{n} p(S|l_i, u_i) \prod_{(v_i, v_j) \in E} p(l_i, l_j|c_{ij}) \qquad (4)$$

We are interested in L^* which maximizes this equation . Taking the negative logarithm we find:

$$L^* = \underset{L}{\operatorname{argmin}} \left(\sum_{i=1}^{n} m_i(l_i) + \sum_{(v_i, v_j) \in E} d_{ij}(l_i, l_j) \right) \qquad (5)$$

where $m_i(l_i)$ and $d_{ij}(l_i, l_j)$ are the negative logarithm of the likelihood and the prior respectively.

We learn the structure of G in a similar fashion to [2]. First, a fully connected graph is constructed by finding the maximum likelihood estimates for U and C. In this graph each node represents a part v_i and has an associated probability $p(S|l_i, u_i)$ and the connections represent the relationships between the parts (v_i, v_j) and have an associated probability $p(l_i, l_j|c_{ij})$. A tree structure is then derived by finding the minimum spanning tree of the graph, where the edge weights are set to $-log(\prod_{k=1}^{r} p(l_i^k, l_j^k|c_{ij}))$ for r training drawings.

The edges which are left in the tree describe for each part its strongest relationship. For example, the left eye and right eye in a face will most likely be connected since their spatial relationship is very well defined (usually very little vertical distance), and their sizes are very similar.

Finally, we can find L^* by using the min-sum algorithm (also known as the Viterbi algorithm), which guarantees an optimal minimum. Although the tree limits the amount of relationships we have between parts, and thus uses less information, it reduces the computational complexity of the inference from $O(n^m m^2)$ for a full graph to $O(nm^2)$ (where n is the number parts, and m is the total number of strokes). Thus, the algorithm can scale much better for images which contain many strokes, which could be true in a general case.

4. IMPLEMENTATION

The features that were selected to model both $p(S|L, \theta)$ and $p(L|\theta)$ emphasizes the differences that exist between photographs and drawings. For example, in drawings there is a large intraclass variance between the appearance of the same parts (Fig. 2) and so no good features could be found to evaluate $p(S|L, \theta)$. Many features were examined, among them rectangularity, height to width ratio, bounding box angle and compactness. However, all led to a degradation in the results since none was able to successfully describe a specific part. Therefore, a uniform probability was given to all strokes.

Since the appearance of the part does not signify its semantic meaning, the inference was based on $p(L|\theta)$ which contains the prior information about the relationships between the parts. We model each of these relationships as a gaussian random variable. Therefore each c_{ij} is the mean and variance of the distribution, computed by calculating the mean and variance of each feature from the training data. The following relationships were found to exhibit the best results:

1. The horizontal distance between the bounding boxes' centroids (Δx).
2. The vertical distance between the bounding boxes' centroids (Δy).
3. The difference in size of the bounding boxes.
4. The difference in the rectangularity of the parts.
5. The difference in angles of the bounding boxes

In order to detect objects of different scales, the first three connection parameters c_{ij} are learned with respect to the size of part v_i. When each of the c_{ij} is learned, the measured quantity is first normalized with respect to the size of the bounding box of s_k, the

Output Color	Body	Face	House	Flower
blue	head	face	building	pot
purple	neck	left eye	roof	saucer
brown	torso	right eye	left window	stem
green	left arm	nose	right window	right leaf
black	right arm	mouth	door	left leaf
orange	left leg	left ear	path	stigma
red	right leg	right ear	door knob	petals

Table 1. Names and colors of parts for each category.

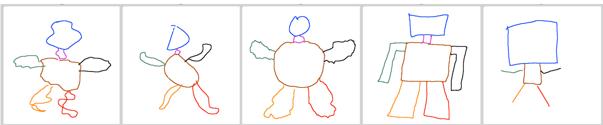


Fig. 3. A sample of body types detected by our algorithm.

strokes which represents part v_i in the training phase. Therefore the final distribution of each connection parameter, is over the measured parameter normalized by the size of the part. During testing we divide each measured quantity by the size of the bounding box of each s_i and thus are able to detect objects of different scales.

We did experiment with other relationships (i.e relative circularity, absolute distance, bounding box overlap, etc.). However, adding these features yielded poorer classification results, and so finally the 5 features mentioned above were selected. The first two features are the most important ones for many parts. In many drawing the parts are mainly defined by their relative location. For example, in a flower we would always expect to find the saucer beneath the pot. The third relationship is only important for some parts. For example, we would always expect the face to be bigger than the nose, but the relative sizes of the nose and mouth can vary widely. The final two features mainly allow to describe parts that are similar. For example, The two eyes would usually have a similar shape, and their angles would be symmetrical to each other.

Once the object location L^* which minimizes eq. 5 is detected, the strokes which comprise it are hidden, and a new search begins. This is repeated until a search finds an object whose score is above a threshold t. This last detection is discarded, and all previous ones are labeled as correct detections.

5. EXPERIMENT

5.1. Experimental setup

The experiments were done on 4 object categories: body, face, house and potted flower. Each one of the objects was constructed out of seven parts as seen in table 1.

We collected a total of at least 60 training drawings and 60 test drawings for each category from 12 different people on a tablet pc (a total of 500 images). Using a java interface, each person would provide the ground truth while drawing, by selecting the correct color to match the part he/she is drawing. The only constraint we put on the drawers was that each part of the object would be drawn as one stroke, since the model relies on this fact. During the training phase people were requested to draw only a specific object using the required parts. During the testing phase, subjects were allowed to draw multiple instantiations of an object in one drawing, and add distracting strokes, which are not part of any specific object. This dataset is available for further research [10].

To test the algorithm's robustness , 1-3 test images from different categories where randomly overlaid on top of each other to create a new set of 103 test images from our original 250. This increased the amount of noise in the test images, while allowing different object categories to appear in one drawing. After learning the parameters for each detector using the relevant training drawings, we ran each of the 4 detectors on the overlaid images.

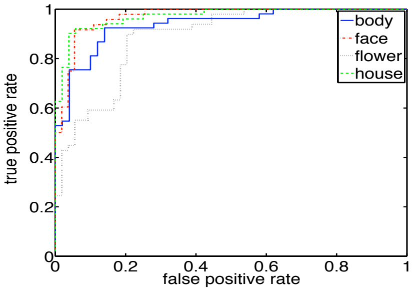


Fig. 4. ROC curves for each of the four category detectors (best viewed in color).

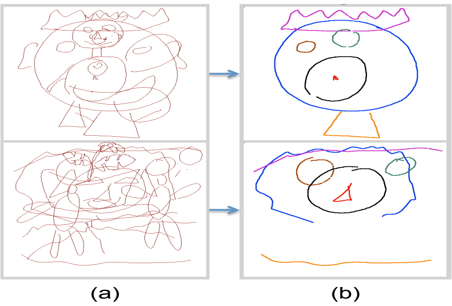


Fig. 5. A sample of false positive detections for house (best viewed in color).

5.2. Results & Discussion

Sample results are shown in Fig. 6. The ROC curves in Fig. 4 show promising results given that many of the drawings are cluttered, and consist of many strokes (20-100 strokes per image). They also show the relative performance of the different detectors. It can be observed that face and house detectors perform the best, while the flower detector has the poorest result. This makes sense, since the structure of both the face and house categories is much more rigid than that of the flower category.

Fig. 3 displays a sample of bodies that are detected by our algorithm. These different body shapes again emphasize the importance of the relationships between the parts vs. the actual appearance of the stroke. For example, notice that the head can be drawn as different shapes (square, rectangle, circle, triangle, etc,). Fig. 5 displays a sample of false positive results from the house detector. Although there was no house in any of the original drawings which were overlaid to create the input (Fig. 5(a)), it is clear to see why the algorithm detected a house. The overlaid drawings created a configuration which could definitely appear to look like a house

The algorithm is also extremely fast. Using 60 images, the learning phase for each detector takes approx. 0.6 sec, while the average testing time is just 0.07 sec/image for each detector including feature extraction. Even the most cluttered images with 100 strokes take only about 0.2 seconds. This is mainly due to the computation reduction we achieve by using a tree structure.

For each detector we also compute the equal error rate, and the confidence score at that rate. We then compute a confusion matrix for each detector. The confusion matrix shows the performance of the part labeling. Each stroke is labeled as either a certain part or

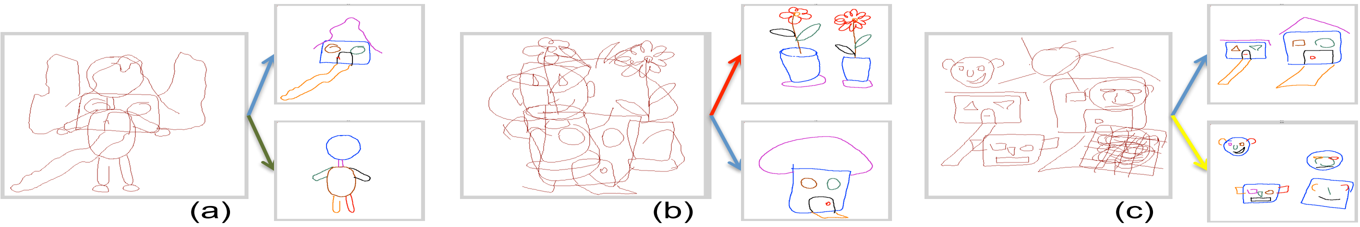
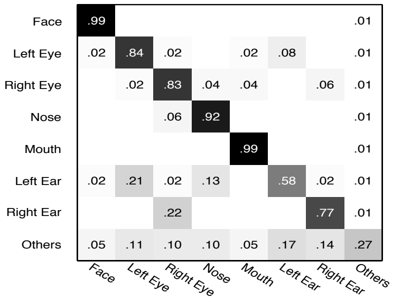
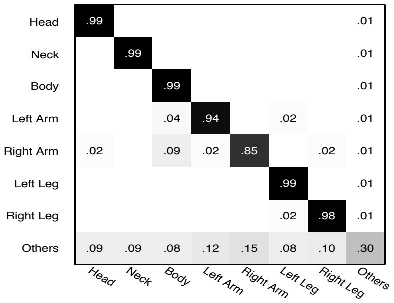


Fig. 6. Examples of inputs and results obtained by our different detectors. Arrow colors signify which detector has been used. (Green-Body, Yellow-Face, Red-Flower, Blue-House). Part labels and colors are described in Table 1. Notice error in face labeling in (c). See discussion section for explanation.



	Face	Left Eye	Right Eye	Nose	Mouth	Left Ear	Right Ear	Others
Face	.99							.01
Left Eye	.02	.84	.02		.02	.08		.01
Right Eye		.02	.83	.04	.04		.06	.01
Nose			.06	.92				.01
Mouth					.99			.01
Left Ear	.02	.21	.02	.13		.58	.02	.01
Right Ear			.22				.77	.01
Others	.05	.11	.10	.10	.05	.17	.14	.27

Fig. 7. Confusion matrix for face part labeling.



	Head	Neck	Body	Left Arm	Right Arm	Left Leg	Right Leg	Others
Head	.99							.01
Neck		.99						.01
Body			.99					.01
Left Arm			.04	.94		.02		.01
Right Arm	.02		.09	.02	.85		.02	.01
Left Leg						.99		.01
Right Leg						.02	.98	.01
Others	.09	.09	.08	.12	.15	.08	.10	.30

Fig. 8. Confusion matrix for body part labeling.

background and then compared with the ground truth. We present only the confusion matrices for the face and body detectors in Fig. 7 & Fig. 8 because of lack of space.

The face detector's confusion matrix highlights some of the problems which arise because of the reduction of the graphical model to a tree structure. For example, the tree structure for the face detector contains edges between the right eye and the left eye, and between the right ear and the left ear. This would be expected since they have relationships which are relatively consistent across drawings. However, in this model there is no edge that connects any of the eyes to the ears. This results in a common error where the left ear would be mistaken for the left eye, and similarly for the right (Fig. 6(c)). This can be seen in the confusion matrix in Fig. 7. As seen in Fig. 8 a similar error appears in the body detector between the body and right arm, since the right arm only has a connection to left arm. However, even with these common errors, the detector themselves still perform well. A full graph will be able to correct these errors, but would be more computationally expensive.

These results also highlight further work which can be done in this field. First, since a tree structure is not always sufficient, other structures such as k-fans could be used [11]. These structures present a good balance between computational complexity and representational power. Also, since no appearance features were found based on a gaussian assumption, perhaps a multimodal distribution can be learned which can better explain the shape (as suggested in [1]). Finally, in order to allow for multiple parts per stroke, and multiple strokes per part, a shape segmentation algorithm would need to be incorporated into the model.

6. CONCLUSION

We propose a pictorial structure model for object recognition and part labeling in drawings. We modify the original model to better perform on drawings by doing the inference over strokes, and choose useful features for the task of object detection. We present results on a cluttered data-set in which drawings are overlaid on top of each other, and mange to achieve promising results.

7. REFERENCES

[1] D. Sharon and M. van de Panne, "Constellation models for sketch recognition," *EUROGRAPHICS Workshop on Sketch Based Interfaces and Modeling*, 2006.

[2] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, pp. 55–79, 2005.

[3] B. Paulson T. Hammond and B. Eoff, "Eurographics tutorial on sketch recognition," *EUROGRAPHICS tutorial*, 2009.

[4] Christine Alvarado and Randall Davis, "Sketchread: a multi-domain sketch recognition engine," in *UIST*, 2004, pp. 23–32.

[5] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 509–522, April 2002.

[6] A. Hall, C. Pomm, and P. Widmayer, "A combinatorial approach to multi-domain sketch recognition," in *SBIM*, 2007, pp. 7–14.

[7] Tevfik Metin Sezgin and Randall Davis, "Hmm-based efficient sketch recognition," in *IUI*, 2005, pp. 281–283.

[8] Yuan Qi, Martin Szummer, and Thomas P. Minka, "Diagram structure recognition by bayesian conditional random fields," in *CVPR*, 2005, pp. 191–196.

[9] Graham Mackenzie and Natasha Alechina, "Classifying sketches of animals using an agent-based system," in *Computer Analysis of Images and Patterns*, 2003, pp. 521–529.

[10] Amir Sadovnik, "Pictorial structures for object recognitoin and part labeling in drawings," http://chenlab.ece.cornell.edu/projects/drawing_object_recognition/.

[11] David Crandall, Pedro Felzenszwalb, and Daniel Huttenlocher, "Spatial priors for part-based recognition using statistical models," in *CVPR*, 2005, pp. 10–17.