Shape Detection with Nearest Neighbour Contour Fragments

Kasim Terzić
w3.ualg.pt/~kterzic
Hussein Adnan Mohammed
a48025@ualg.pt
J.M.H. du Buf
w3.ualg.pt/~dubuf

Vision Lab (LARSyS)
University of the Algarve
Faculdade de Ciências e Tecnologia
Gambelas, Faro, Portugal

Shape is probably the single most important feature for object detection and much research has gone into developing deformable shape models. However, contours extracted by bottom-up edge detectors are notoriously unreliable, especially in natural images. In this paper, we present a very simple, yet powerful method for model creation, hypothesis generation, and hypothesis verification, which is competitive with much more complex methods.

Let $\mathbf{s}$ be a descriptor in some high-dimensional space of an edge fragment representing part of an object contour. In Bayesian terms, this fragment was likely to be generated by some class $c \in C$ if the conditional likelihood $P(\mathbf{s}|c)$ is greater for c than that for any other class, $P(\mathbf{s}|c' \neq c)$, including the background class. $P(\mathbf{s}|c)$ can be estimated non-parametrically, for example by using Gaussian kernels associated with some nearby samples in the feature space. In [2] it was shown that a good approximation can be obtained by using only the nearest sample. A fragment is *discriminative* if it is much more likely to belong to a class c than any other class $c' \neq c$. We thus define the discriminative power of a fragment by:

$$d(\mathbf{s}) := \log \frac{P(\mathbf{s}|c)}{P(\mathbf{s}|c' \neq c)} \,. \tag{1}$$

As proposed in [3], we can approximate the second likelihood using the distance to the nearest sample of any other class:

$$d(\mathbf{s}) := \|\mathbf{s} - NN_{c'}(\mathbf{s})\|^2 - \|\mathbf{s} - NN_c(\mathbf{s})\|^2; \qquad c' \neq c \,, \tag{2}$$

which is a simple, distance-based criterion. We further define the *relevance* $r(\mathbf{s},c)$ of a fragment for class c as the probability that a similar fragment $\mathbf{s}'$ appears in an annotated sub-image I_c containing c. We consider a fragment $\mathbf{s}'$ similar to $\mathbf{s}$ if the distance between the two is less than some threshold T:

$$r(\mathbf{s},c) = P(\|\mathbf{s} - \mathbf{s}'\| < T|c); \qquad \text{where } \mathbf{s}' = NN_c(\mathbf{s}). \tag{3}$$

We estimate T by using the nearest neighbour of $\mathbf{s}$ in all training images *not* containing c. The probability in Eqn. 3 can thus be approximated by counting the number of different annotated sub-images of c in the training set, in which there is at least one fragment closer to $\mathbf{s}$ than T.

We begin by extracting all maximally long contiguous edge fragments from the image, by linking edges produced by the Berkeley detector. We extend this set by adding subsets of all fragments with one half and one quarter length shifted to cover different parts of each fragment, which strikes a good balance between the number of fragments and sufficient statistics. Each fragment obtained is stored as a vector of 2D coordinates, and a shape context $\mathbf{s}$ is computed for each vector, using standard parameters. [1]

Next, we build a canonical model for each shape. Relevance $r(\mathbf{s},c)$ is calculated for each fragment belonging to each class c and they are sorted in descending order. We keep the most relevant 5% of the fragments and discard the rest. Thus an ensemble of fragments is used to represent each shape.

We then generate object hypotheses by determining the discriminative power $d(\mathbf{s})$ of each fragment $\mathbf{s}$ extracted from a test image. We keep the top 20% of most discriminative fragments and discard the rest. For each of the remaining fragments (called "trigger fragments"), we generate a hypothesis by projecting the bounding box of the nearest neighbour of $\mathbf{s}$ into a new image, scaling it to fit the position and scale of $\mathbf{s}$. Hypotheses are clustered around few promising spots, and after removing obvious overlaps, we are left with about 20 hypotheses per class per image. Each hypothesis is generated by one "trigger" fragment. We refine the detections by collecting fragments inside each hypothesis and fitting a bounding box which jointly minimises the bounding box error for all fragments using least square minimisation.

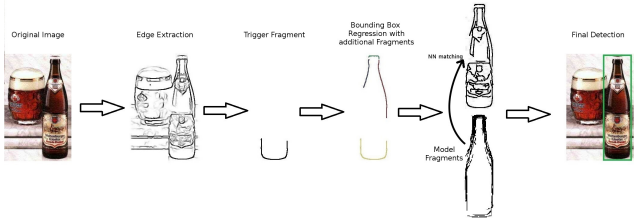

Figure 1: Overview of our method.

Apple	Bottles	Giraffes	Mugs	Swans	**Mean**
95/95	**96.4/100**	**81.3/85.4**	**87.1/87.1**	**88.2/94.1**	**89.6/92.3**

Table 1: Results on ETHZ (recall at 0.3/0.4 FPPI, respectively).

The final step in our algorithm is the verification of each hypothesis. A score for a hypothesis can be obtained from the average distance between fragments $\mathbf{s}_i$ in the hypothesis, and fragments $\mathbf{q}_m$ in the model of the corresponding class, but we have to be careful. Matching a hypothesis to the model will include the influence of clutter fragments and ignore missing parts of the shape. Both issues can be avoided by matching the fragments $\mathbf{q}_m$ from the model to the segments in the hypothesis in order to obtain the distance. Clutter fragments in the hypothesis will tend to be ignored because they do not resemble the models, and parts of the model contour that do not have a close match will increase the overall distance, thus penalising incomplete contours. We define the final scoring function of a hypothesis h in the following way:

$$\text{score}(h) = \frac{1}{M} \sum_m (\mathbf{q}_m - NN_h(\mathbf{q}_m))^2 \,, \tag{4}$$

where $\mathbf{q}_m$ is the position-enhanced shape context descriptor of the m^{th} fragment in the model and $NN_h(\mathbf{q}_m)$ is its nearest neighbour in the hypothesis.

We evaluated our algorithm on the popular ETHZ shape dataset, with results shown in Table 1. We note that there are several difficult aspects in shape detection in natural images: i) model creation, ii) hypothesis generation, and iii) hypothesis verification/classification. Our method tackles all these problems consistently based on Bayesian criteria and nearest neighbours.

We are currently investigating a combination of our approach with advanced contour extraction methods and more powerful classifiers. We believe that contour linking and a more powerful verification step can significantly boost the results. There is also potential for further optimisation, making the algorithm more suitable for real-time applications. Finally, we would like to explore using prior information from scene models to improve detection.

Acknowledgements This work was supported by the EU under the FP-7 grant ICT-2009.2.1-270247 *NeuralDynamics* and by the FCT under the grant UID/EEA/5009/2013.

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE T-PAMI*, 24(4):509–522, 2002.

[2] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, Anchorage, 2008.

[3] S. McCann and D.G. Lowe. Local naive bayes nearest neighbor for image classification. In *CVPR*, pages 3650–3656, Providence, 2012.