Sketch Recognition by Ensemble Matching of Structured Features

Yi Li
http://www.eecs.qmul.ac.uk/~yl303
Yi-Zhe Song
http://www.eecs.qmul.ac.uk/~yzs
Shaogang Gong
http://www.eecs.qmul.ac.uk/~sgg

School of Electronic Engineering and Computer Science,
Queen Mary, University of London,
London E1 4NS, UK

Sketch recognition aims to automatically classify human hand sketches of objects into known categories. This has become increasingly a desirable capability due to recent advances in human computer interaction on portable devices. The problem is nontrivial because of the sparse and abstract nature of hand drawings as compared to photographic images of objects, compounded by a highly variable degree of details in human sketches.

Current methods for sketch recognition and sketch-based image retrieval all employ a bag-of-features (BOF) representation of object sketches without considering their spatial structures. A characteristic of sketch is that its basic strokes do not necessarily exhibit strong discriminative cues in isolation. Instead, the structures of a sketch both locally and holistically contain informative visual cues for discriminating different sketches.

In this work, we exploit a star graph as a structured feature representation to encode both local features and the holistic structure of a sketch. We also exploit the ensemble matching for computing the distance metric when comparing star graph representations of different sketches [6]. While star graphs can represent holistic structure of sketches well, current BOF approaches [2, 4, 5] have the benefit of being able to better capture subtle structural details. To that end, we further propose a unified framework to address both holistic structural variations and local detail differences. More specifically, we introduce a separate category filtering process as a first step prior to ensemble matching to keep only a few categories most similar in local details, utilizing Support Vector Machines (SVM) classification on BOF.

For ensemble matching, we consider it as a graph matching problem, where each ensemble is encoded as a star graph. More precisely, we denote a star graph as $G = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, where $\mathcal{V}$, $\mathcal{E}$, $\mathcal{A}$ represent respectively a set of nodes, edges and attributes of the graph. In particular, $\mathcal{V} = \{v_i\}_{i=1}^{N_s} \cup c$ is the set of all N_s sample points $\{v_i\}_{i=1}^{N_s}$ and the center c, and $e_i \in \mathcal{E}$ is the implicit link between v_i and c. Moreover, $\mathbf{a}_{ic} \in \mathcal{A}$ represents the geometrical relationship between v_i and c, and $\mathbf{a}_i \in \mathcal{A}$ denotes the corresponding feature descriptor of v_i.

The computation of the similarity between ensemble q (query) and t (target) is formulated as follows:

$$P(G^q, G^t) = \sum_i P(\mathbf{a}_i^t | \mathbf{a}_i^q) P(\mathbf{a}_{ic}^t | \mathbf{a}_{ic}^q) \quad (1)$$

where $G^q = (\mathcal{V}^q, \mathcal{E}^q, \mathcal{A}^q)$ and $G^t = (\mathcal{V}^t, \mathcal{E}^t, \mathcal{A}^t)$ are their corresponding star graphs. $P(\cdot,\cdot)$ denotes the normalized distance metric value and is considered as probability. The feature similarity term $P(\mathbf{a}_i^t | \mathbf{a}_i^q)$ accounts for the similarity between features and the feature location correlation term $P(\mathbf{a}_{ic}^t | \mathbf{a}_{ic}^q)$ stands for the location correlation between two features.

We modify traditional ensemble matching [6] in several minor ways. First, similar to [1], we employ a two steps ensemble matching algorithm to accelerate the matching process. It first finds the most similar D target features $\{\mathbf{a}_j^t\}_{j=1}^{D}$ for each feature in the query (D is much smaller than the total feature amount in the target), then calculates location correlations only for these D features. The similarity between the query and the target is then:

$$P(G^q, G^t) = \sum_i \max_j P(\mathbf{a}_j^t | \mathbf{a}_i^q) P(\mathbf{a}_{jc}^t | \mathbf{a}_{ic}^q) \quad (2)$$

Second, the sum rule is employed to obtain the overall matching score (c.f. Equation (1)) instead of the product rule employed in [1], as it is proven to be the most resilient to estimation errors [3]. Third, we get rid of the center estimation and multi-scale matching, due to the single subject essence of the sketch images and pre-scale procedure for the sketches. And the center of the ensemble is set to the geometrical center of the sketch. Fourth, to ensure the matching score between two sketches is constant, two sides comparison is employed by swapping the query and the target. And a partial matching penalty factor is added to penalize the matching score according to how many points in the target is not matched. The final matching score is then:

$$P^f(G^q, G^t) = w_1 * P(G^q, G^t) + w_2 * P(G^t, G^q) \quad (3)$$

where, w_i is the proportion of points being matched in the current target.

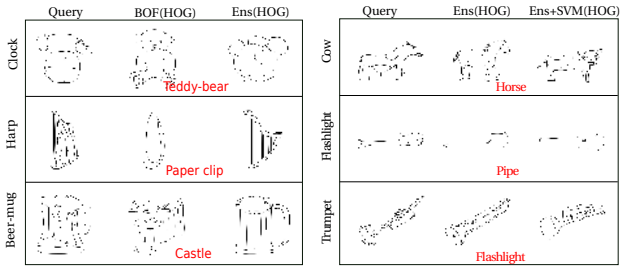

Figure 1: (a) qualitative comparisons of top retrieval results of three sketch probes on K nearest neighbors classification on bag-of-features (BOF(HOG)) and ensemble matching only method (Ens(HOG)). Ensemble matching preserves better holistic structure correspondence. (b) qualitative comparisons of top retrieval results of three sketch probes on ensemble matching only (Ens(HOG)) and ensemble matching with category filtering (Ens+SVM(HOG)). Category filtering helps to address subtle structural details.

For category filtering, we employ SVM classifiers to filter sketch categories prior to ensemble matching, therefore keep N categories closest to the query in term of local details other than holistic structure. More specifically, we represent a sketch by a n-dimensional BOF histogram $\mathbf{h}$. A set of SVM classifiers are trained with respect to the number of sketch categories in a training dataset. For a probe sketch image to be classified, the following voting function classifies a given probe sketch image into the ith category:

$$c^i(\mathbf{h}) = \sum_j w_j^i K(\mathbf{s}_j^i, \mathbf{h}) + b \quad (4)$$

where K is a kernel function, $\mathbf{s}_j^i$ are the support vectors, w_j^i are weights, and b is the bias. $c^i(\mathbf{h})$ is therefore the classification response measuring similarity between the probe and the ith category.

The conclusion is that by integrating ensemble matching into sketch recognition, the structure information of a sketch can be effectively represented using a star graph, and an unified ensemble matching with multi-SVM classification based category filtering benefits from both holistic structure and subtle local details of sketches.

[1] O. Boiman and M. Irani. Detecting irregularities in images and in video. *IJVC*, 74(1):17–31, 2007.

[2] R. Hu, M. Barnard, and J. Collomosse. Gradient field descriptor for sketch based retrieval and localization. In *ICIP*, pages 1025–1028, 2010.

[3] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *PAMI*, 20(3):226–239, 1998.

[4] E. Mathias, H. Kristian, B. Tamy, and A. Marc. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *TVCG*, 17(11): 1624–1636, 2011.

[5] E. Mathias, H. James, and A. Marc. How do humans sketch objects? *ACM TOG (Proceedings SIGGRAPH)*, 31(4):44:1–44:10, 2012.

[6] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *CVPR*, pages 1–8, 2007.