SPATIAL PYRAMID MINING FOR LOGO DETECTION IN NATURAL SCENES

Jim Kleban[*]

University of California Santa Barbara
ECE Department
kleban@ece.ucsb.edu

Xing Xie, Wei-Ying Ma

Microsoft Research Asia
No. 49 Zhichun Road, Haidian District
Beijing, 100080, P. R. China
xingx, wyma@microsoft.com

ABSTRACT

This work introduces a novel data mining scheme, spatial pyramid mining, to discover association rules at multiple resolutions in order to identify frequent spatial configurations of local features that correspond to classes of logos appearing in real world scenes. By indexing representative examples by the mined rules we can efficiently detect a variety of different lettering or design marks associated with a brand. Features in an image are marked by matching rules to representative examples selected via a weighted cosine similarity measure. Logos are localized in an image via density-based clustering of matched features. Precision vs. recall curves are presented for experiments on a dataset of web images of nearly 1,000 images containing seven popular logo types.

Index Terms— Data Mining, Object Detection, Mobile Search, Logo Recognition

1. INTRODUCTION

A new wave of services is emerging with pervasive wireless networking and cheap computing that will enable mobile users to retrieve information from their surroundings. Towards that end, this work addresses the problem of identifying logos on objects as they occur in images captured in natural settings. Object detection in real world environments is difficult due to clutter, occlusions, and variation in photometric conditions and perspective. As brand markings occur in wide varieties, template-matching and point correspondence approaches may suffer from poor recall. Instead we approach logo detection by treating each brand as a class of objects for which to discover frequent local feature configurations and then match in images from indexed templates. This work employs a novel multiresolution spatial pyramid mining technique to identifies useful rules at varying quantization levels in feature space and geometric configuration.

Logos are a tempting target for detection as they often appear in high contrast regions with distinctive shape and edge information. Applied logo detection systems may enable a variety of new web-based applications. For instance, mobile users may submit image search queries via camera-enabled devices [1] that can help identify products, offer discounts at store locations, or provide context-based advertisements.

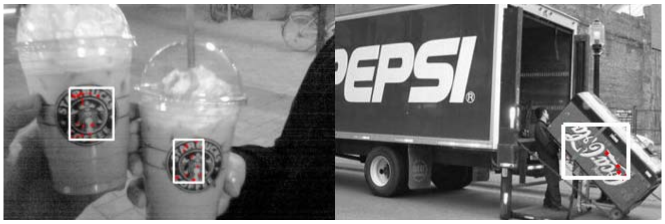


Fig. 1. Logo detection by clustering matching frequent spatial configurations of local features found by data mining.

Previous work in logo detection has primarily focused on the problem of locating design marks in documents [2] or similarity matching for trademark retrieval [3]. Shape and contour based features are employed for analysis of well-registered data sets. Other work attempts detection in real world scenes. For logos in sports videos, Den Hollander and Hanjalic [4] assume rigid planar backgrounds and utilize line-based intensity profiles. As local interest point descriptors like SIFT [5] have become popular, there have been approaches that use point correspondences for matching to templates. Bagdanov et.al. [6] employ normalized matching using a bag of local features and cloud localization in sports videos. The Logoseeker system [7] fits an affine transformation in order to mark the convex hull of detected regions.

This paper approaches logo detection by data mining [8] association rules that capture frequent spatial configurations of quantized local SIFT descriptors. Association rule mining was initially employed for market basket analysis and has been extended to image based domains [9]. Recently, Quack et.al. [10] introduced the idea of employing association rules to select features for object detection. The collection of rules as "discrete words" are indexed to retrieve representative training templates for matching. Our work extends [10] and applies an indexing method to the localization of logos in an image. Figure 1 shows an example of matched

[*] While at MSRA and supported by IGERT NSF Grant# DGE-0221713.

feature points using this method.

The paper proceeds as follows: section 2 outlines the data mining method used to extract association rules and describes the spatial pyramid mining scheme. Section 3 describes rule indexing, template selection, and localization in test images. Section 4 discusses the results of experiments using a dataset of logos collected from the web. Section 5 concludes.

2. MINING ASSOCIATION RULES FROM IMAGE TRANSACTIONS

The goal of association rule mining [8] here is to discover frequent item sets that infer the presence of the target object class with high confidence. The aim is to discover association rules of the form $A \Rightarrow B$. The discrete items in set A correspond to local interest descriptors discretized by both hierarchical k-means clustering and spatial configuration (see the next subsection.) Set B contains only the class label item. Transactions are sets of items $T = \{i_1, \ldots, i_n\}$. Set A is a K-*itemset* if $A \subseteq T$ with $|K|$ items, and the frequency at which A occurs over the set of all transactions, database D, is called the support:

$$supp\,(A) = \frac{|\{T \in D : A \subseteq T\}|}{|D|} \qquad (1)$$

We wish to discover frequently occurring K-itemsets existing in the transaction supersets that imply the positive training class label, itself included as an appended item in T. Association rules $A \Rightarrow B$ can be found efficiently using the Apriori algorithm [8] and must meet a minimal statistical significance or *support* of a rule defined as:

$$supp\,(A \Rightarrow B) = \frac{|\{T \in D : (A \cup B) \subseteq T\}|}{|D|} \qquad (2)$$

Additionally, rules must satisfy a minimum *confidence*, the frequency over which occurrences of antecedent A imply consequent B. Rule confidence is an estimate of the conditional probability $P(B \subset T | A \subset T)$ and is defined as:

$$conf\,(A \Rightarrow B) = \frac{|\{T \in D : (A \cup B) \subseteq T\}|}{|\{T \in D : A \subseteq T\}|} \qquad (3)$$

2.1. Transaction Items for Spatial Word Configurations

Transactions are constructed from items of discretized configurations of local interest points in an image. Each item corresponds to one feature point detected by the Harris-Affine method [11] and represented as the SIFT feature descriptor. Each transaction consists of items located in the semi-local neighborhood of a single central feature point. The motivation for sampling spatial neighborhoods as an extension of the bag-of-features process can be found in [10]. In essence, we extend their method and apply it to logo detection.

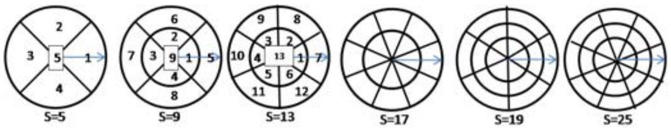


Fig. 2. Grids for coarser to finer levels of spatial quantization

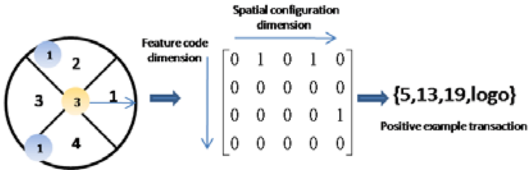


Fig. 3. Example transaction. 3 local features, in 1 of 4 clusters and lying in 1 of 5 regions, activate matrix elements.

We construct transaction databases at multiple resolutions as follows. The first step is to employ a tree-based hierarchical k-means clustering of the SIFT [5] descriptors from the Harris-Affine output. The tree has $Q(k) = 2^k$ clusters at each level k. These detected feature points are assigned to one of the $Q(k)$ clusters at a selected level k. To incorporate shape information we next iteratively overlay a circular grid centered at each feature point. The grid has a radius proportional to the base feature's scale, and rotation invariance of the configurations is achieved by rotating the circular grid by the base feature's orientation. Each transaction contains one item for each feature lying within this circular area. Feature points within the grid are additionally quantized according to their relative position in $S(z)$ circular spatial regions, where z is the given level of spatial quantization. Figure 2 shows sets of the dividing regions for increasingly finer levels.

Creating an example transaction is illustrated in figure 3. Each feature point in an image will create one such transaction. In the figure, the base feature, assigned to the 3rd feature space cluster (q=3), is found to have two other features in its spatial neighborhood. These are both assigned to the 1st feature space cluster (q=1) and lie in sectors (s=2) and (s=4). If the transaction database has at this resolution Q = 4 clusters and S = 5 sectors (4 plus 1 for the origin,) then the items in a transaction will be indices ranging from 1 to $Q \cdot S$ = 20 found as (s-1)*Q+q. Repeated items are discarded. The example transaction has 4 total items, three features and a class label.

2.2. Spatial Pyramid Mining

Two innovations in object detection motivate the idea:

1. Computation of a pyramid match kernel of coarse-to-fine histogram intersections in feature space to efficiently learn with sets of features [12]
2. Simultaneously matching in spatial subregions [13]

The idea of spatial pyramid mining is to compute association rules found in transaction databases at multiple resolutions in both feature space and semi-local spatial regions which when taken together outperform any single resolution. Han and Fu [14] introduce the idea of multiple-level association rule mining but focus on using top level results to efficiently find lower level ones. Here the idea is that mining at multiple resolutions finds different types of rules that can be appropriately weighted. The interplay between resolution in spatial and feature space leads to complementary rules ranging from a) distinct types of local features in coarse spatial regions to b) more variant local features with specific spatial arrangements.

The procedure for spatial pyramid mining is as follows:

1. Compute over the training set multiple transaction databases $D_{k,z}$ where k is a level on the hierarchical k-means tree with $Q(k) = F^k$ cluster nodes and z is the circular grid type with $S(z)$ sectors .
2. For each $D_{k,z}$, iteratively employ Apriori algorithm to find rules r with varying confidence and support minima that imply the positive class label item.
3. For a test image G, find rules matching features by calculating matrix $M_{k,z}$:

$$M_{k,z} = R_{k,z} T_G \tag{4}$$

   if rule antecedents are denoted as r_j of J total rules from $D_{k,z}$, and the transactions are t_p of P total in the test image, then $R_{k,z}$ is the sparse rules activation matrix with J rows, $Q(k) \cdot S(z)$ columns, and elements $R_{k,z}(j,i) = \frac{1}{|r_j|}$ for items $i \in r_j$. T_G is the sparse test image transactions matrix with $Q(k) \cdot S(z)$ rows, P columns and elements $T_G(i,p) = 1$ where $i \in t_p$. Rule j matches at feature p in G when $M_{k,z}(j,p) = 1$.
4. The weighted confidence f that feature p in test image G belongs to the target class is:

$$f(p) = \sum_{\forall k} \sum_{\forall z} \sum_{\forall j : M_{k,z}(j,p)=1} w_{k,z}(r_j) \tag{5}$$

   where rule weights $w_{k,z}$ are the product of scalars:

$$w_{k,z}(r_j) = c(r_j) \cdot s(r_j) \cdot |r_j| \cdot Q(k) \cdot S(z) \tag{6}$$

   with $c(r_j)$ as rule confidence, $s(r_j)$ rule support, and $|r_j|$ rule length. The term $Q(k) \cdot S(z)$ accounts for the lower support at finer levels.

3. INDEXING AND LOGO LOCALIZATION

To find logos in a scene we group selected features by mutual rule matches with representative training examples. An enclosing region is found by density-based clustering[15] of mutually matching points. This method can efficiently approximate pointwise matching over the templates in the training set by indexing the mined frequent itemset rules.

For a set of N total mined rules $R = \{r_1, r_2, \ldots, r_N\}$ with confidence weight vector $\overline{w} = [w_1 \; w_2 \; \ldots \; w_N]$ as found by (6), for each of M training templates create the N-dimensional indexing vector $\bar{t}_m$ with $t_m(n) = \sqrt{w(n)}$ when training image m contains a match to rule n, otherwise $t_m(n) = 0$. Then, for a test image with index vector $\bar{t}_{test}$ constructed in similar fashion, select a representative template, l, via weighted cosine similarity as:

$$l = \underset{m}{\operatorname{argmax}} \; \frac{\bar{t}_m \cdot \bar{t}_{test}^{\,T}}{\left\|\bar{t}_m\right\| \left\|\bar{t}_{test}^{\,T}\right\|} \tag{7}$$

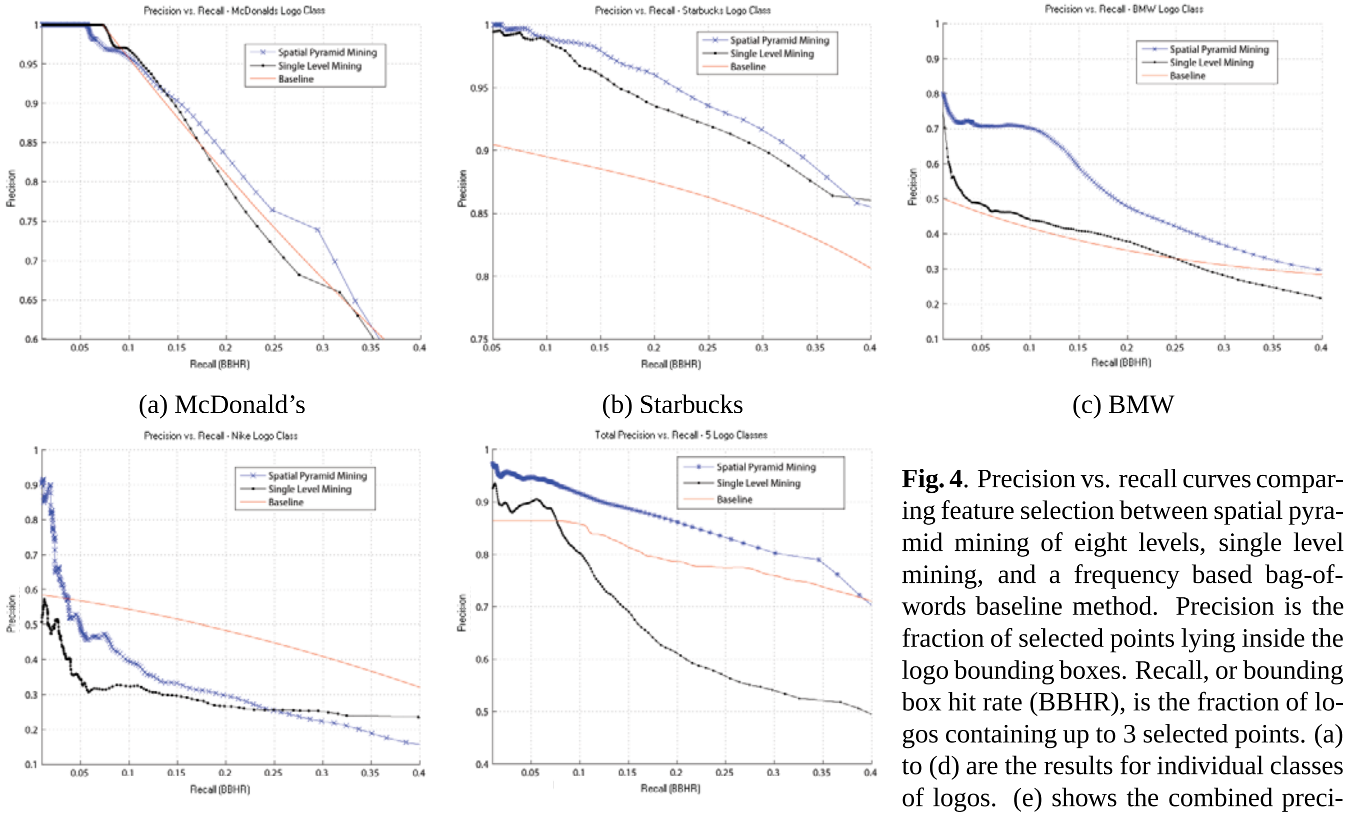
Points corresponding to mutual rule matches between the test image and template l are clustered by the DBSCAN algorithm. A bounding box encloses the convex hull of points in clusters with sufficient membership. A threshold on the similarity can be used to identify different logos co-occurring in the same image.

4. EXPERIMENTAL RESULTS

To test our method we collected a dataset of 974 images from the web containing logos on products, buildings and signs for the McDonald's (130 images), Starbucks (148), Nike (124), Yankees (119), BMW (139), Coca-Cola (179) and Apple (135) brands. A bounding box is annotated for each logo appearing in an image. For rule discovery, logo-containing transactions are combined at a 1:3 ratio with background class transactions taken from regions outside the bounding boxes and from 200 non-related images. The databases contain 50,000 total transactions, and at least 500 rules are extracted from each resolution with varying minimum support and at least 90% minimum confidence. Results are generated by 6-fold cross validation over random 80/20 training-test splits.

Figure 4 shows precision and recall curves for selected points over four classes separately (a-d) and a total curve for 5 selected classes (e). Precision is defined as the fraction of selected feature points in the image located inside the annotated boxes. Recall is the fraction of bounding boxes hit with at least K selected feature points, in this case K=3. The total number of mined rules limits the maximum recall compared with a baseline bag-of-words method that selects feature points by their hit frequency at 2^{13} clusters. Spatial pyramid mining rules are combined over 8 database resolutions with $Q = \left\{2^{13}, 2^{12}, 2^{11}, 2^{11}, 2^{10}, 2^{10}, 2^{9}, 2^{9}\right\}$ clusters and $S = \{5, 9, 9, 13, 13, 17, 19, 25\}$ sectors respectively. We also plot against one of the better performing single mining resolutions ($Q = 2^{13}, S = 5$) as in [10].

Results for the seven types of logos vary. The best overall performance is for the Starbucks and Coca-cola classes where the data contains large examples that have many aligning spatial configurations. The Nike and Yankees logos are difficult as the symbols tend to appear on small areas in the image. Mining spatially configured rules offers no advantage over baseline as seen in figure 4(d) and for this reason the two poor performing classes are excluded in 4(e).



(a) McDonald's

(b) Starbucks

(c) BMW

(d) Nike

(e) Total Precision vs. Recall

Fig. 4. Precision vs. recall curves comparing feature selection between spatial pyramid mining of eight levels, single level mining, and a frequency based bag-of-words baseline method. Precision is the fraction of selected points lying inside the logo bounding boxes. Recall, or bounding box hit rate (BBHR), is the fraction of logos containing up to 3 selected points. (a) to (d) are the results for individual classes of logos. (e) shows the combined precision vs. recall curves for 5 of the classes.

5. CONCLUSION

This work presents a method for logo detection based on mining frequent spatial configurations of discretized local features at multiple resolutions. The technique can discover a variety of logo types in real world settings. By indexing based on these rules the system can scale to a large database of templates. Experiments on a challenging real world dataset show that the rules can successfully select features that appear densely on logos. An important limitation of this method is image resolution, as multiple local features are required to mine robust spatial configurations.

References

[1] Jia M., Fan X., Xie X., Li M., and Ma W. Y., "Photo-to-search: Using camera phones to inquire of the surrounding world," in *MDM*, 2006.

[2] F. Cesarini, E. Francesconi, M. Gori, S. Marinai, J. Q. Sheng, and G. Soda, "A neural-based architecture for spot-noisy logo recognition," in *ICDAR*, 1997.

[3] A. Jain and A. Vailaya, "Shape-based retrieval: A case study with trademark image databases," in *Pattern Recognition, 31(9):1369–13990*, 1998.

[4] R.J.M. den Hollander and Alan Hanjalic, "Logo recognition in video stills by string matching," in *ICIP*, 2003.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," in *IJCV*, 2004.

[6] A. D. Bagdanov, L. Ballan, M. Bertini, and A. Del Bimbo, "Trademark matching and retrieval in sports video databases," in *MIR*, 2007.

[7] S. Sanyal and S. H. Sengamedu, "Logoseeker: a system for detecting and matching logos in natural images," in *Proc. of ACM Multimedia*, 2007.

[8] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Addison Wesley, 2005.

[9] J. Tesic, S. Newsam, and B. S. Manjunath, "Mining image datasets using perceptual association rules," in *SIAM '03 Workshop on Mining Scientific and Engineering Datasets*.

[10] T. Quack, V. Ferrari, B. Liebe, and L. V. Gool, "Efficient mining of frequent and distinctive feature configurations," in *ICCV*, 1997.

[11] Krystian Mikolajczyk and Cordelia Schmid, "Scale & affine invariant interest point detectors," in *IJCV*, 2004.

[12] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *ICCV*, 2005.

[13] S. Lazebnik, C.Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006.

[14] J. Han and Y. Fu, "Discovery of multiple-level association rules from large databases," in *VLDB*, 1995.

[15] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *KDD*, 1996.