Learning Visual Shape Lexicon for Document Image Content Recognition

Guangyu Zhu, Xiaodong Yu, Yi Li, and David Doermann

University of Maryland, College Park, MD 20742, USA

Abstract. Developing effective content recognition methods for diverse imagery continues to challenge computer vision researchers. We present a new approach for document image content categorization using a lexicon of shape features. Each lexical word corresponds to a scale and rotation invariant shape feature that is generic enough to be detected repeatably and segmentation free. We learn a concise, structurally indexed shape lexicon from training by clustering and partitioning feature types through graph cuts. We demonstrate our approach on two challenging document image content recognition problems: 1) The classification of $4,500$ Web images crawled from Google Image Search into three content categories — pure image, image with text, and document image, and 2) Language identification of 8 languages (Arabic, Chinese, English, Hindi, Japanese, Korean, Russian, and Thai) on a $1,512$ complex document image database composed of mixed machine printed text and handwriting. Our approach is capable to handle high intra-class variability and shows results that exceed other state-of-the-art approaches, allowing it to be used as a content recognizer in image indexing and retrieval systems.

1 Introduction

Image content categorization has become a pressing problem in computer vision as we are facing phenomenal increase in the diversity of visual content. Content category recognition aims to reduce the semantic gap for ensuing tasks by providing usage-oriented content description that can be exploitable in individual applications. For vision systems involving high-volume, complex, and heterogeneous multimedia data, effective high-level content interpretation is essential prior to object detection or object category recognition at a finer level.

One pervasive form of information content is text. Once text content and the language are recognized, images containing text can be processed by an optical character recognition (OCR) system and indexed. Text-oriented content recognition provides a reliable alternative to object detection and recognition in a wide range of applications. Towards this end, however, there are a number of challenges to image content category recognition that are largely unsolved. In this paper, we focus on two important problems that are central to heterogeneous document image collections.

First, we consider the recognition of primary content of a general image as one of three content categories — pure image (*e.g.*natural image and human photos),

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part II, LNCS 5303, pp. 745–758, 2008.

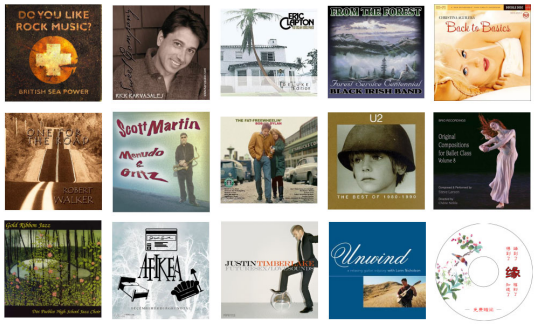


Fig. 1. Examples of images returned by Google Image using the keyword "CD cover"

image with text (see examples in Fig. 1), or document image. Automated content categorization like this has a big impact in image search, and content-based image indexing and retrieval.

Second, we address the problem of recognizing the primary language of a document image in an unconstrained setting. This is a fundamental research challenge currently facing systems that need to automatically process diverse multilingual document images, such as Google Book Search [30] or an automated global expense reimbursement application [32], because almost all existing work on OCR requires that the script and/or language of the processed document be known [23]. The performance of language identification is crucial for the success of a broad range of tasks — from determining the correct OCR engine for text extraction to document indexing, translation, and search [33]. Progress in the field of language identification has focused almost exclusively on machine printed text. Document collections, however, often contain a diverse and complex mixture of machine printed and unconstrained handwritten content, and vary tremendously in font and style. Language identification on document images involving diverse content types, including unconstrained handwriting, is still an open research area [22] and to our best knowledge, no reasonable solutions have been presented in the literature.

We propose a novel approach for document image content recognition using image descriptors built from a lexicon of generic low-level shape features that are translation, scale, and rotation invariant. To construct a structural index among large number of features extracted from diverse content, we dynamically partition the space of shape primitives by clustering similar feature types. We formulate feature partitioning as a graph cuts problem with the objective to obtain a concise and globally balanced lexicon index by sampling from training data. Each cluster in the lexicon is represented by an exemplary lexical word, making association of feature type efficient. We obtain very competitive document image content categorization performance using a multi-class SVM classifier.

The structure of this paper is as follows. Section 2 reviews related work. In Section 3, we describe the algorithm for learning the shape lexicon and present a document image content recognition approach using the shape lexicon. We discuss experimental results in Section 4 and conclude in Section 5.

2 Related Work

In this following, we first review contour learning approaches, and then point out work related to us on whole-image categorization. We present a comprehensive overview of existing on script and language identification techniques and discuss their limitations on document images with unconstrained content.

2.1 Contour-Based Learning

Topological relationships among adjacent contours is an important aspect in many vision problems and have been exploited in work on object detection [1,9,24], and on perceptual grouping [12,17]. Ferrari *et al*. [8] proposed scale-invariant adjacent segments (kAS) features extracted from the contour segment network of image tiles, and use them in a sliding window scheme for object detection. By explicitly encoding both geometric and spatial arrangement among the segments, kAS descriptor demonstrates state-of-the-art performance in shape-based object detection, and outperforms descriptors based on interest points and histograms of gradient orientations [6]. However, kAS descriptor is not rotation invariant, because segments are rigidly ordered from left to right. This limits the repeatability of high-order kAS, and the best performance in [8] is reported when using 2AS.

2.2 Image Content Categorization

There has been little computer vision and image processing literature directly addressing the problem of content recognition for heterogeneous image repositories. However, several approaches based on different motivations have demonstrated good performance in tasks that involve diverse objects. Oliva and Torralba [21] developed a holistic image representation called Spatial Envelope for scene recognition using a set of discriminative energy spectrum templates that characterizes the dominant spatial structure of a scene. Another fairly intuitive approach is to treat blocks of text as texture [29]. One widely used rotation invariant feature for texture analysis is the Local Binary Patterns (LBP) proposed by Ojala *et al*. [20], which effectively captures spatial structure of local image texture in circular neighborhoods across angular space and resolution.

2.3 Language Identification

Prior research on script and language identification has largely focused on the domain of machine printed document images. These works can be broadly classified into three categories — statistical analysis of text lines [7, 14, 18, 27, 28], texture analysis [4, 29], and template matching [11].

Statistical analysis using discriminating features extracted from text lines, including distribution of upward concavities [14, 27], horizontal projection profile [7, 28], and vertical cuts of connected components [18], has shown to be effective on homogeneous collection of printed documents. These approaches, however, do have a few major limitations. First, they are based on the assumption of uniformity among printed text, and require precise baseline alignment and word segmentation. Freestyle handwritten text lines are curvilinear, and in

general, there are no well-defined baselines, even by linear or piecewise-linear approximation [15]. Second, it is difficult to extend these methods to a new language, because they employ a combination of hand-picked and trainable features and a variety of decision rules. In fact, most of these approaches require effective script identification to discriminate between selected subset of languages, and use different feature sets for script and language identification, respectively.

Script identification using rotation invariant multi-channel Gabor filters [29] and wavelet log co-occurrence features [4] were demonstrated on small blocks of printed text with similar characteristics. However, no results were reported on full-page documents that involve variations in layouts and fonts.

The current state-of-the-art script/language identification system was developed by Hochberg *et al.* [11] at Los Alamos National Laboratory, and is able to process 13 machine printed scripts without explicit assumptions of distinguishing characteristics for a selected subset of languages. The system determines the most likely script on a document page by probabilistic voting on matched templates. Each template pattern is of fixed size and is rescaled from a cluster of connected components. Template matching is intuitive and delivers impressive performance when the content is constrained (*i.e.*printed text in similar fonts). However, templates are not flexible enough to generalize across large variations in fonts or handwriting styles that are typical in diverse datasets [11]. From a practical point of view, the system also has to learn the discriminability of each template through labor-intensive training to achieve the best performance. This requires tremendous amount of supervision and further limits applicability.

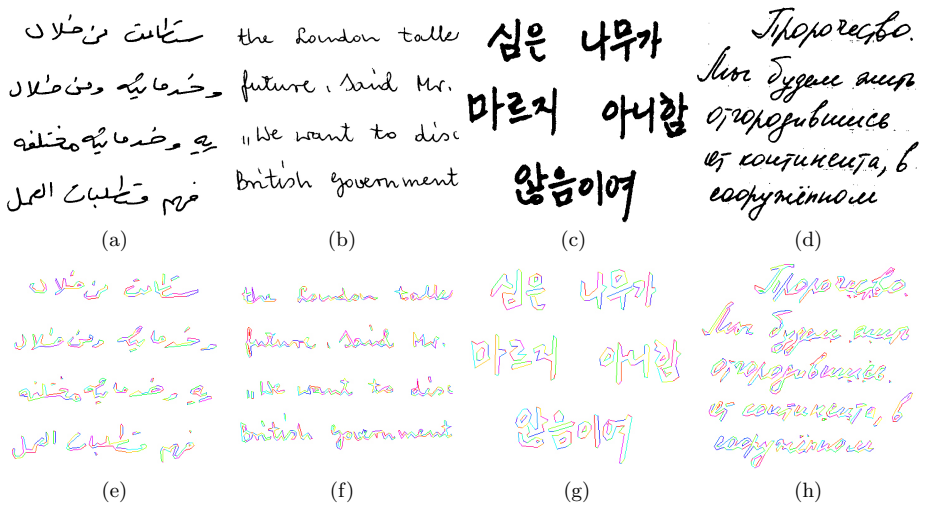
3 Recognizing Image Content Using Shape Lexicon

Recognition of diverse visual content needs to account for large variations, because content appears in many forms and in many different contexts today. The scope of the problem is intuitively in favor of low-level shape primitives that can be detected repeatably. Rather than focusing on selection of class-specific features, our approach aims to distinguish intricate differences between content types collectively using the statistics of a large variety of generic, geometrically invariant feature types (lexical words) that are structurally indexed. Our emphasis on the generic nature of lexical words provide a different perspective to recognition that has traditionally focused on finding sophisticated features or visual selection models, which may limit generalization performance.

We explore the kAS contour feature recently introduced by Ferrari *et al.* [8], which consists of a chain of k roughly straight, connected contour segments. Specifically, we focus on the case of triple contour segments, which strike a balance between lower-order contour features that are not discriminative enough and higher-order ones that are less likely to be detected robustly.

3.1 Extraction of Contour Feature

Feature detection in our approach is very efficient since we perform computation locally. First, we compute edges using the Canny edge detector [5], which



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 2. Visual shape differences are captured locally by a large variety of neighboring contour features. (a)-(d) Examples of handwriting from four different languages. (e)-(h) Detected contour features by our approach, each shown in a random color.

consistently demonstrates good performance on text content and gives precise localization and unique response. Second, we group contour segments by connected components and fit them locally into line segments. Then, within each connected component, every triplet of connected line segments that starts from the current segment is extracted. Fig. 2 provides visualization of the quality of detected contour features by our approach using random colors.

Our feature detection scheme requires only linear time and space in the number of contour fragments n, and is highly parallelizable. It is much more efficient and stable than [8], which requires construction of contour segment network and depth first search from each segment, leading to $O(n \log(n))$ time on average and $O(n^2)$ in the worst case.

We encode object contours in a translation, scale, and rotation invariant fashion by computing orientations and lengths with reference to the first detected line segment. A contour feature C can be compactly represented by an ordered set of lengths and orientations of c_i for $i \in \{1, 2, 3\}$, where c_i denotes line segment i in C. This is distinct from the motivation of kAS descriptor that attempts to enumerate spatial arrangements of contours within local regions. Furthermore, kAS descriptor does not take into account of rotation invariance.

3.2 Measure of Dissimilarity

The overall dissimilarity between two contour features can be quantified by the weighted sum of the distances in lengths and orientations. We use the following generalized measure of dissimilarity between two contour features C_a and C_b

$$d^{(C_a, C_b, \boldsymbol{\lambda})} = \boldsymbol{\lambda}_{\text{length}}^T \mathbf{d}_{\text{length}} + \boldsymbol{\lambda}_{\text{orient}}^T \mathbf{d}_{\text{orient}}, \tag{1}$$

where vectors $\mathbf{d}_{\text{length}}$ and $\mathbf{d}_{\text{orient}}$ are composed of the distances between contour lengths and orientations, respectively. $\boldsymbol{\lambda}_{\text{length}}$ and $\boldsymbol{\lambda}_{\text{orient}}$ are their corresponding weight vectors, providing sensitivity control over the tolerance of line fitting. One natural measure of dissimilarity in lengths between two contour segments is their log ratio. We compute orientation difference between two segments by normalizing their absolute value of angle difference to π. In our experiments, we use a larger weighting factor for orientation to de-emphasize the difference in the lengths because they may be less accurate due to line fitting.

3.3 Learning the Shape Lexicon

We extract a large number of lexical words by sampling from training images, and construct an indexed shape lexicon by clustering and partitioning the lexical words. A lexicon provides a concise structural organization for associating large varieties of low-level features, and is efficient because it enables comparison to much fewer feature types.

Clustering Lexical Words. Prior to clustering, we compute the distance between each pair of lexical words and construct a weighted undirected graph $\mathcal{G} = (V, E)$, in which each node on the graph represents a word. The weight on an edge connecting two nodes C_a and C_b is defined as a function of their distance

$$w(C_a, C_b) = \exp(-\frac{d(C_a, C_b)^2}{\sigma_d^2}), \tag{2}$$

where we set parameter σ_d to 20 percent of the maximum distance among all pairs of nodes.

We formulate feature clustering as a spectral graph partitioning problem, for which we seek to group the set of vertices V into disjoint sets $\{V_1, V_2, \ldots, V_K\}$, such that by the measure defined in (1) the similarity among the vertices in a set is high and that across different sets is low.

More concretely, let the $N \times N$ symmetric weight matrix for all the vertices be W, where $N = |V|$. We define the degree matrix D as an $N \times N$ diagonal matrix, whose i-th element $d(i)$ along the diagonal satisfies $d(i) = \sum_j w(i, j)$. We use an $N \times K$ matrix X to represent a graph partition, *i.e.*$X = [X_1, X_2, \ldots, X_K]$, where each element of matrix X is either 0 or 1. We can show that the feature clustering formulation that seeks globally balanced graph partitions is equivalent to the normalized cuts criterion [26], and can be written as

$$\text{maximize } \epsilon(X) = \frac{1}{K} \sum_{l=1}^{K} \frac{X_l^T W X_l}{X_l^T D X_l}, \tag{3}$$

$$\text{subject to } X \in \{0, 1\}^{N \times K}, \text{ and } \sum_j X(i, j) = 1. \tag{4}$$

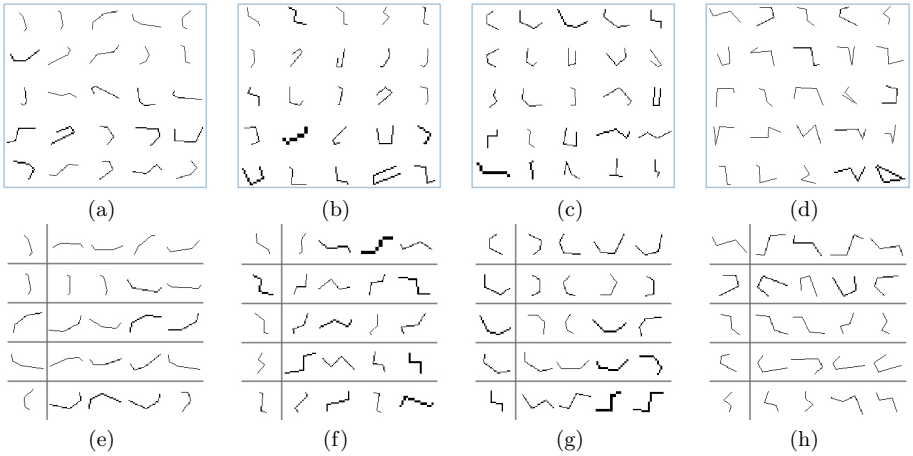


Fig. 3. The 25 most frequent exemplary lexical words in (a) Arabic, (b) Chinese, (c) English, and (d) Hindi document images, which very well capture the distinct features between languages. (e)-(h) show lexical words in the same cluster as the top 5 exemplary lexical words for each language, ordered by ascending distances to the center of their clusters. Scaled and rotated versions of feature types are clustered together.

Minimizing normalized cuts exactly is NP-complete. We use a fast algorithm [31] for finding its discrete near-global optima, which is robust to random initialization and converges faster than other clustering methods.

Organizing Features in the Lexicon. For each cluster, we select the feature instance closest to the center of the cluster as the exemplary lexical word. This ensures that an exemplary word has the smallest sum of squares distance to the other features within the cluster. In addition, each exemplary word is associated with a cluster radius, which is defined as the maximum distance from the cluster center to all the other features within the cluster. The constructed shape lexicon $\mathcal{L}$ is composed of all exemplary lexical words.

Fig. 3 shows the 25 most frequent exemplary lexical words for Arabic, Chinese, English, and Hindi, which are learned from 10 documents of each language. Distinguishing features between languages, including cursive style in Arabic, 45 and 90-degree transitions in Chinese, and various configurations due to long horizontal lines in Hindi, are learned automatically. Each row in Fig. 3(e)-(h) lists examples of lexical words in the same cluster, ordered by ascending distances to the center of their associated clusters. Through clustering, translated, scaled and rotated versions of feature types are grouped together.

Since each lexical word represents a generic local shape feature, intuitively a majority of lexical words should appear in images across content categories, even though their frequencies of occurrence deviate significantly. In our experiments, we find that 95.1% and 92.6% of lexical words in natural images also appear in images with text and document images, respectively. In addition, 86.3% of lexical word instances appear in document images across all 8 languages.

3.4 Constructing the Image Descriptor

We construct a shape descriptor for each image, which provides statistics of the frequency at which each feature type occurs. For each detected lexical word W from the test image, we compute the nearest exemplary word C_k in the shape lexicon. We increment the descriptor entry corresponding to C_k only if

$$d(W, C_k) < r_k, \tag{5}$$

where r_k is the cluster radius associated with the exemplary lexicon word C_k. This quantization step ensures that unseen features that deviate considerably from training features are not used for image description. In our experiments, we found that only less than 2% of the contour features cannot be found in the shape lexicon learned from the training data.

4 Experimental Results

Before we demonstrate our approach in image content categorization and language identification, we first quantitatively evaluate the discriminative power and geometrical invariance of shape lexicon in two object category recognition experiments using public shape databases.

4.1 Shape-Based Object Recognition

Kimia Database. One widely used shape database is the Kimia dataset [25], which contains 25 images from 6 categories. Several categories require rotation invariant matching for effective recognition. It has been tested in [25,10,2,16]. In this experiment, we use the χ^2 statistic as the measure of dissimilarity between two image descriptors, and the distance $D_{\chi^2}(h_a, h_b)$ between two normalized K-bin image descriptors h_a and h_b is defined as

$$D_{\chi^2}(h_a, h_b) = \frac{1}{2} \sum_{k=1}^{K} \frac{[h_a(k) - h_b(k)]^2}{h_a(k) - h_b(k)}. \tag{6}$$

Historically, results on Kimia database are reported as the number of 1st, 2nd, and 3rd nearest-neighbors that fall into the correct category. Our result is 25/25, 25/25, 24/25, which outperforms four other approaches shown in Table 1. This experiment demonstrates the discriminative power of shape lexicon descriptor, in addition to its rotation invariance.

MPEG-7 Shape Database. Our next experiment involves the MPEG-7 CE-Shape-1 database [13], which consists of 1,400 silhouette images: 70 shape categories, 20 images per category. The retrieval performance is measured using the "bullseye test", in which each image is used as a query and one counts the number of correct images in the top 40 matches. The retrieval score is the ratio of the number of correct hits to the best possible number of hits.

Table 1. Summary of results on Kimia database

Method	Top 1	Top 2	Top 3
Sharvit *et al.* [25]	23/25	21/25	20/25
Gdalyahu and Weinshall [10]	25/25	21/25	19/25
Belongie *et al.* [2]	25/25	24/25	22/25
Ling and Jacobs [16]	25/25	24/25	25/25
Our approach	**25/25**	**25/25**	**24/25**

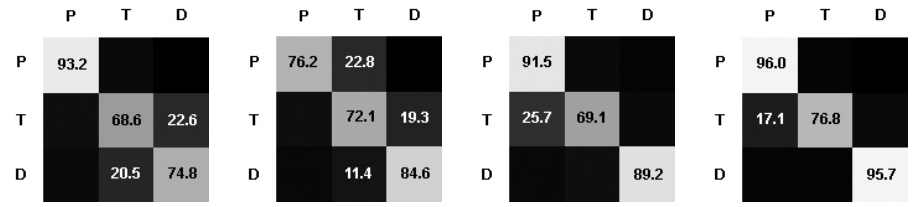
It is important to note that most well-known shape descriptors [2,16] and geometric hashing schemes [3] tested on MPEG-7 database have much higher complexity than the shape lexicon descriptor. For instance, shape contexts (SC) [2] and the articulation-invariant inner-distance shape contexts (IDSC) [16] both require attaching a spatial histogram for every point on the shape, which describes the distribution of the relative positions of all remaining points. The correspondences between two point sets are solved among $O(n^2)$ descriptor pairs, where n is the number of points sampled on object contours. The $O(n^3)$ complexity for each matching can be prohibitively expensive in most practical problems unless the object has been precisely localized and effectively segmented.

We quantitatively evaluate the discriminability of different descriptors, as this understanding is fundamental for image indexing and retrieval problems when it comes to the choice of either a local or global representation. In our experiment, the shape lexicon descriptor obtains a retrieval score of 62.10%. The bullseye score with SC is 64.59%, and the IDSC has the highest score of 68.83% as inner-distance can more robustly handle articulation than Euclidean distance.

To help understand the performance, we further verify the results and discover that a large number of errors by the shape lexicon descriptor come from object categories with low complexity, where the discriminative shape parts have been smoothed away after line fitting. This is less a problem for SC and IDSC as they directly sample along the object contour. The performance of the shape lexicon descriptor is comparable with SC and IDSC for object categories that involve complex local part structures. This demonstrates that a descriptor using nonparametric distribution of a set of indexed local shape primitives can have discriminability fairly close to a competitive global descriptor, such as shape contexts. In addition, our approach runs efficiently in linear time and space, which is critical in high-volume image indexing and retrieval.

4.2 Image Content Category Recognition

To evaluate our approach for image content category recognition, we construct a $4,500$-image dataset by crawling Web images from the Google Image search engine using a wide variety of keywords. Fig. 1 shows some examples of images with text returned by using the text keyword "CD cover". All the images are automatically downloaded by a script. Duplicate and junk images are manually inspected and removed to reduce the proportion of unrelated images.



(a) Mean diagonal = 78.9% (b) Mean diagonal = 77.6% (c) Mean diagonal = 83.3% (d) Mean diagonal = 89.6%

Fig. 4. Confusion tables for image content category recognition using (a) Spatial envelope [21], (b) LBP [20], (c) kAS [8], (d) Our approach. (P: Pure image, T: Image with text, D: Document image).

We compare our approach with spatial envelope [21], local binary patterns (LBP) [20], and the state-of-the-art kAS descriptor [8], which are well-known approaches based on different views of whole-image characterization. Spatial envelope uses a holistic image representation without attempting to exploit localized information such as shape, whereas LBP is based on rotation-invariant texture analysis. Since 2AS gives the best performance among different kAS [8], we use it as the benchmark for kAS.

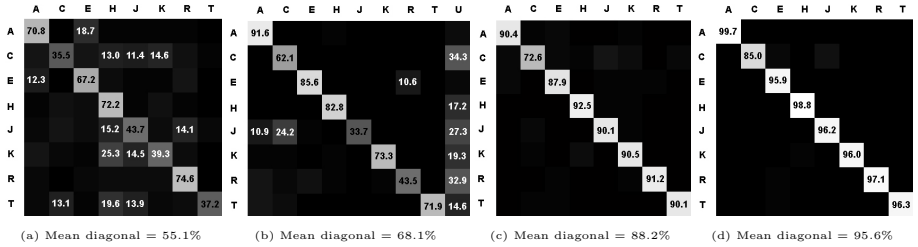
We use a multi-class SVM classifier trained with LIBSVM. The SVM classifier is trained using only 100 randomly selected images from each category, and used to test the rest images in the collection. For easy comparison, we set the dimensions of the image descriptor to 90 for both kAS and our approach in the following experiments.

The confusion tables for spatial envelope, LBP, kAS, and our approach are shown in Fig. 4. Spatial envelope demonstrates good performance in recognizing pure images, but it is not very effective for text content. Texture-based LBP gives balanced results for all the three content types. Our approach obtains the best performances for recognizing each content class, with a respectable mean diagonal of 89.6%. The only notable confusion occurs when distinguishing between image with text and pure image.

4.3 Language Identification

We use $1,512$ document images of 8 languages (Arabic, Chinese, English, Hindi, Japanese, Korean, Russian, and Thai) from the University of Maryland multilingual database [15] and the IAM handwriting database DB3.0 [19] (see Fig. 7) for evaluation on language identification. Both databases are large public real-world collections, containing the source identity of each image in the ground truth. This enables us to construct a diverse dataset that closely mirrors the true complexities of heterogeneous document image repositories in practice.

We compare our approach with the state-of-the-art language identification system [11], which is based on template matching. We also include LBP and kAS in this experiment since they have demonstrated reasonable performance on



(a) Mean diagonal = 55.1% (b) Mean diagonal = 68.1% (c) Mean diagonal = 88.2% (d) Mean diagonal = 95.6%

Fig. 5. Confusion tables for language identification using (a) LBP [20], (b) Template matching [11], (c) kAS [8], (d) Our approach. (A: Arabic, C: Chinese, E: English, H: Hindi, J: Japanese, K: Korean, R: Russian, T: Thai, U: Unknown).

Table 2. Confusion table of our approach for the 8 languages

	A	C	E	H	J	K	R	T
A	99.7	0.3	0	0	0	0	0	0
C	1.4	85.0	4.0	1.0	6.7	1.0	0.7	0.2
E	1.6	0	95.9	0.2	0	1.1	0.6	0.6
H	0.2	0.2	0	98.8	0.8	0	0	0
J	0	1.3	1.0	0.2	96.2	1.3	0	0
K	0	0.8	0.1	1.9	0.5	96.0	0.5	0.1
R	0.5	0	2.0	0	0	0	97.1	0.4
T	0	0.3	1.6	0.9	0.6	0.3	0	96.3

diverse text contents. In this experiment, the SVM classifier is trained using the same pool of 50 randomly selected images from each category.

The confusion tables for LBP, template matching, kAS, and our approach are shown in Fig. 5. The performance of template matching varies significantly across languages. One significant confusion of template matching is between Japanese and Chinese since a document in Japanese may contain varying amount of Kanji (Chinese characters). Rigid templates are not flexible for identifying discriminative partial features, and the bias in voting decision towards the dominant candidate causes less frequently matched templates to be ignored. Another source of error that lowers the performance of template matching (mean diagonal = 68.1%) is undetermined cases (see the *unknown* column in Fig. 5(b)), where probabilistic voting cannot decide between languages with roughly equal votes. Texture-based LBP could not effectively recognize differences between languages on a diverse dataset because distinctive layouts and unconstrained handwriting exhibit irregularities that are difficult to capture using texture, and its mean diagonal is only 55.1%.

Our approach gives excellent results on all 8 languages, with an impressive mean diagonal of 95.6% (see Table 2 for all entries in the confusion table of our approach). kAS, with a mean diagonal of 88.2%, is also shown to be very effective. Neither method has difficulty generalizing across large variations such

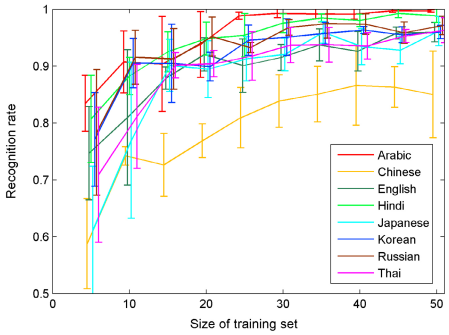


Fig. 6. Recognition rates of our approach for different languages as the size of training data varies. Our approach achieves excellent performance even using a small number of document images per language for training.

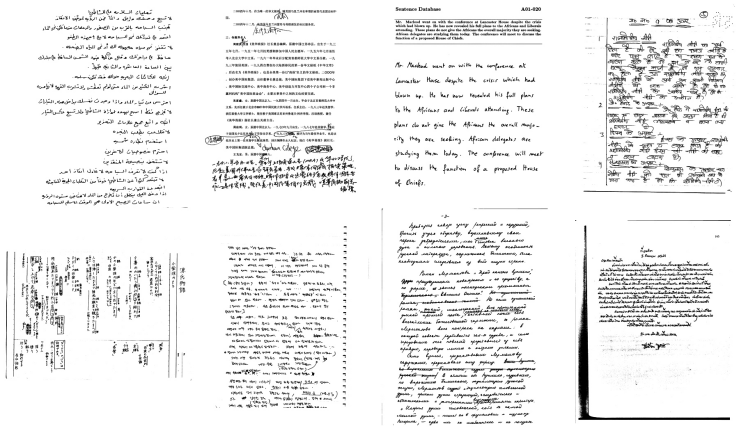


Fig. 7. Examples from the Maryland multilingual database [15] and the IAM handwriting DB3.0 database [19]. Languages in the top row are Arabic, Chinese, English, and Hindi, and those in the second row are Japanese, Korean, Russian, and Thai.

as font types or handwriting styles, which greatly impact the performances of LBP and template matching. This demonstrates the effectiveness of using generic low-level shape features when mid or high-level vision representations may not generalized or flexible enough for the task.

Fig. 6 shows the recognition rates of our approach as the size of training set varies. We observe very competitive language identification performance on this challenging dataset even when a small amount of training data per language class is used. In addition, our results on language identification are very encouraging from a practical point of view, as the training in our approach requires considerably less supervision than template matching. Our approach only needs

the class label of each training image, and does not require skew correction, scale normalization, or segmentation.

5 Conclusion

In this paper, we proposed a novel approach for document image content categorization using a lexicon composed of a wide variety of local shape features. Each lexical word represents a characteristic structure that is generic enough to be detected repeatably and segmentation free. The lexicon provides a principled approach to structurally indexing and associating a vast number of feature types, and is learned from training data with little supervision. Our approach is fully extensible and does not require constructing explicit content models. In two challenging real world document image content recognition problems involving large-scale, highly variable image collections, our approach demonstrated excellent results and outperformed other state-of-the-art approaches. Our future work will be directed towards refining and evaluating the approach by further incorporating spatial co-occurrences of lexical words using a secondary lexicon.

References

1. Amit, Y., Geman, D.: A computational model for visual selection. Neural Computation 11, 1691–1715 (1999)
2. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. 24(4), 509–522 (2002)
3. Biswas, S., Aggarwal, G., Chellappa, R.: Efficient indexing for articulation invariant shape matching and retrieval. In: Proc. CVPR, pp. 1–8 (2007)
4. Busch, A., Boles, W., Sridharan, S.: Texture for script identification. IEEE Trans. Pattern Anal. Mach. Intell. 27(11), 1720–1732 (2005)
5. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. 8(6), 679–697 (1986)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. CVPR, pp. 886–893 (2005)
7. Ding, J., Lam, L., Suen, C.: Classification of oriental and European scripts by using characteristic features. In: Proc. ICDAR, pp. 1023–1027 (1997)
8. Ferrari, V., Fevrier, L., Jurie, F., Schmid, C.: Groups of adjacent contour segments for object detection. IEEE Trans. Pattern Anal. Mach. Intell. 30(1), 1–16 (2008)
9. Fidler, S., Leonardis, A.: Towards scalable representations of object categories: Learning a hierarchy of parts. In: Proc. CVPR, pp. 1–8 (2007)
10. Gdalyahu, Y., Weinshall, D.: Flexible syntactic matching of curves and its application to automatic hierarchical classification of silhouettes. IEEE Trans. Pattern Anal. Mach. Intell. 21(12), 1312–1328 (1999)
11. Hochberg, J., Kelly, P., Thomas, T., Kerns, L.: Automatic script identification from document images using cluster-based templates. IEEE Trans. Pattern Anal. Mach. Intell. 19(2), 176–181 (1997)
12. Jacobs, D.: Robust and efficient detection of salient convex groups. IEEE Trans. Pattern Anal. Mach. Intell. 18(1), 23–37 (1996)

13. Latecki, L., Lakamper, R., Eckhardt, U.: Shape descriptors for non-rigid shapes with a single closed contour. In: Proc. CVPR, pp. 424–429 (2000)
14. Lee, D., Nohl, C., Baird, H.: Language Identification in Complex, Unoriented, and Degraded Document Images. Document Analysis Systems II (1998)
15. Li, Y., Zheng, Y., Doermann, D., Jaeger, S.: Script-independent text line segmentation in freestyle handwritten documents. IEEE Trans. Pattern Anal. Mach. Intell. 30(8), 1313–1329 (2008)
16. Ling, H., Jacobs, D.: Shape classification using the inner-distance. IEEE Trans. Pattern Anal. Mach. Intell. 29(2), 286–299 (2007)
17. Lowe, D.: Three-dimensional object recognition from single two-dimensional images. Artificial Intelligence 31(3), 355–395 (1987)
18. Lu, S., Tan, C.: Script and language identification in noisy and degraded document images. IEEE Trans. Pattern Anal. Mach. Intell. 30(2), 14–24 (2008)
19. Marti, U., Bunke, H.: The IAM-database: An English sentence database for offline handwriting recognition. Int. J. Document Analysis and Recognition 5, 39–46 (2006), http://www.iam.unibe.ch/~fki/iamDB/
20. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. 24(7), 971–987 (2002)
21. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. J. Computer Vision 42(3), 145–175 (2001)
22. Plamondon, R., Srihari, S.: On-line and off-line handwriting recognition: A comprehensive survey. IEEE Trans. Pattern Anal. Mach. Intell. 22(1), 63–84 (2000)
23. Rice, S., Nagy, G., Nartker, T.: Optical Character Recognition: An Illustrated Guide to the Frontier. Kluwer Academic Publishers, Dordrecht (1999)
24. Rothwell, C., Zisserman, A., Forsyth, D., Mundy, J.: Planar object recognition using projective shape representation. Int. J. Computer Vision 16(5), 57–99 (1995)
25. Sharvit, D., Chan, J., Tek, H., Kimia, B.: Symmetry-based indexing of image database. J. Visual Commun. and Image Representation 9(4), 366–380 (1998)
26. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 22(8), 888–905 (2000)
27. Spitz, A.: Determination of script and language content of document images. IEEE Trans. Pattern Anal. Mach. Intell. 19(3), 235–245 (1997)
28. Suen, C., Bergler, S., Nobile, N., Waked, B., Nadal, C., Bloch, A.: Categorizing document images into script and language classes. In: Proc. ICDAR, pp. 297–306 (1998)
29. Tan, T.: Rotation invariant texture features and their use in automatic script identification. IEEE Trans. Pattern Anal. Mach. Intell. 20(7), 751–756 (1998)
30. Vincent, L.: Google Book Search: Document understanding on a massive scale. In: Proc. ICDAR, pp. 819–823 (2007)
31. Yu, S., Shi, J.: Multiclass spectral clustering. In: Proc. ICCV, pp. 11–17 (2003)
32. Zhu, G., Bethea, T.J., Krishna, V.: Extracting relevant named entities for automated expense reimbursement. In: Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 1004–1012 (2007)
33. Zhu, G., Yu, X., Li, Y., Doermann, D.: Unconstrained language identification using a shape codebook. In: Proc. ICFHR, pp. 13–18 (2008)