Are Buildings Only Instances?

Exploration in Architectural Style Categories

Abhinav Goel
abhinavgoel@students.iiit.ac.in

Mayank Juneja
juneja@students.iiit.ac.in

C.V. Jawahar
jawahar@iiit.ac.in

Center for Visual Information Technology, IIIT-Hyderabad, India

ABSTRACT

Instance retrieval has emerged as a promising research area with buildings as the popular test subject. Given a query image or region, the objective is to find images in the database containing the same object or scene. There has been a recent surge in efforts in finding instances of the same building in challenging datasets such as the Oxford 5k dataset[19], Oxford 100k dataset and the Paris dataset[20].

We ascend one level higher and pose the question: *Are Buildings Only Instances?* Buildings located in the same geographical region or constructed in a certain time period in history often follow a specific method of construction. These architectural styles are characterized by certain features which distinguish them from other styles of architecture. We explore, beyond the idea of buildings as instances, the possibility that buildings can be categorized based on the architectural style. Certain characteristic features distinguish an architectural style from others. We perform experiments to evaluate how characteristic information obtained from low-level feature configurations can help in classification of buildings into architectural style categories. Encouraged by our observations, we mine characteristic features with semantic utility for different architectural styles from our dataset of European monuments. These mined features are of various scales, and provide an insight into what makes a particular architectural style category distinct. The utility of the mined characteristics is verified from Wikipedia.

Keywords

Discovery, Mining, Classification, Architecture

1. INTRODUCTION

In recent years, the dedicated efforts of a considerable section of the computer vision community have been directed towards solving the problem of instance retrieval. Given a query image or region, the goal is to find and retrieve identical instances of the query object from a large collection of image/videos. The work on Oxford5k dataset[19] has contributed significantly in this direction. Google Goggles is a widely popular product of research in instance retrieval. Sivic *et. al* [25] introduced the Bag-of-Visual-Words (BoW) method to search for objects and scenes in feature films. Images are represented using a histogram of visual words, obtained by clustering high dimensional descriptors obtained from images. Similar images are retrieved using techniques such as inverted file index [25] or min-hash based methods. A key addition to this pipeline, was the introduction of spatial verification for incorporating geometric information into the orderless Bag of Words method. Initially used as a post-processing step [25, 19], it is now an integral part of the retrieval pipeline [31, 12] for matching multiple views of the same object across images. Since then, the BoW technique, in its many forms, has become a mainframe of several instance retrieval techniques. For the purpose of smooth object retrieval, Arandjelović and Zisserman [1] proposed a Bag-of-Boundaries (BoB) method by vector quantizing HOG descriptors computed on regularly sampled points from object contours. Moving on from objects, several works [19, 20, 31] have used the Bag of Words method for searching on building facades and architectural features. In all these cases, the authors find instances of the same building, albeit, from multiple viewpoints and with visual ambiguities. We ascend one level higher, and pose the question - Are buildings only instances?

According to Wikipedia, an architectural style is a specific method of construction. This may include elements

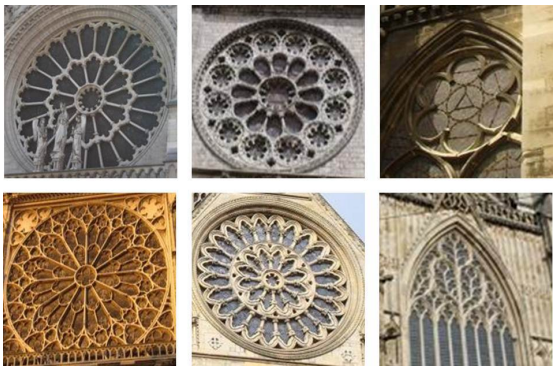


Figure 1: Monuments of an Architectural style have visually similar structures in common. These are some commonly occurring structures mined from our dataset of European Architectural Styles for Gothic Architecture.

ICVGIP '12, December 16-19, 2012, Mumbai, India

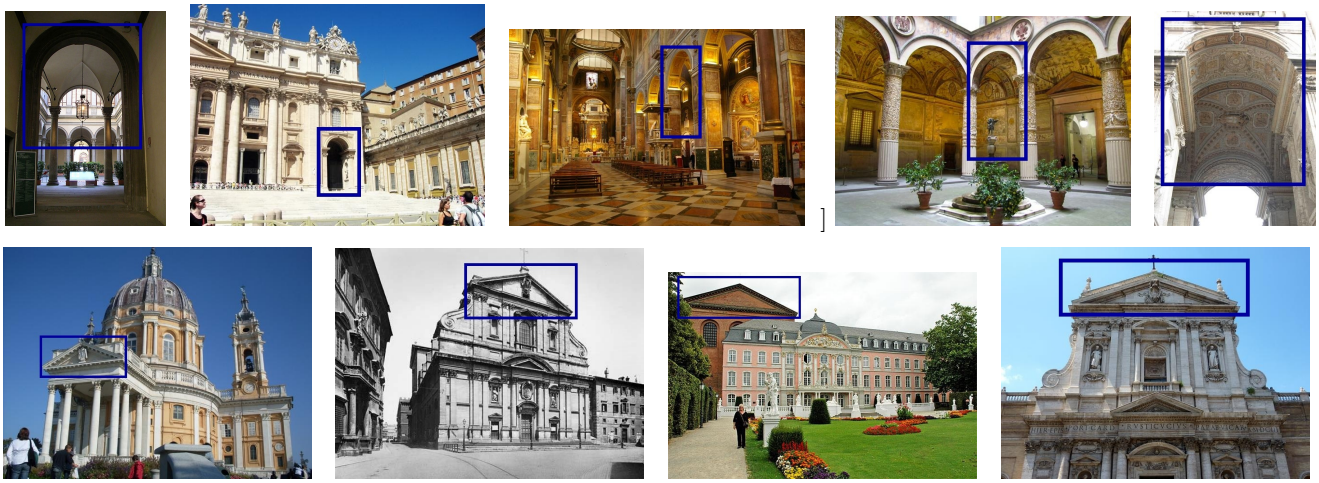


Figure 2: Characteristic Features for two categories of Architectural Styles. Row 1: (Left to Right) Palazzo Strozzi (Florence), San Pietro, Sant Agostino, Santa Maria (Novella), and St. Peter's Basilica. Semi-circular arches are common features of monuments of Renaissance architecture. Row 2: (Left to Right) Basilica di Superga, Church of the Gesu, Elector's Palace, and Santa Susanne (Rome). Monuments in Baroque Architecture often have a triangular pediment above the main entrance. These and many other structures are common across different buildings, belonging to similar architectural styles. Unlike instances, these structures are found in different buildings and are characteristic features for that class of architectural style.

such as form, materials, arrangement, and regional characters. Just like fashion trends, architectural styles vary with time, as well as geographical region. We believe that buildings, apart from being instances, can also be categorized by architectural styles. Each architectural style has certain features that make it distinguishable from other forms of architecture. These characteristic features play a key role in the identification of such monuments. For instance, Rose windows[1] are often found on monuments of the Gothic architectural style. Figure 1 shows images of different varieties of Rose Windows mined as part of our results from various Gothic monuments. An expert in this field provided with information about the characteristics of a monument can easily identify the category of architectural style to which this monument belongs, if not the monument itself. Given a large collection of images of various monuments, we are interested in automating the task of identifying such features and finding an answer to the possibility of buildings being something more than instances. We have no apriori information about (a) which monuments have been captured in the images that we possess, and (b) what features are we trying to discover.

We start by exploring the utility of configurations of low-level discriminative features in categorizing buildings into separate architectural style categories. Each style of architecture is characterized by several features. The color and texture of buildings comes from the materials used for construction. Structures such as windows and arches, vaults and domes can vary in shape across monuments. Various engravings and decorations found on monuments are often characteristic to the time period when they were built, and may reflect the lifestyle and ideas of the people who were around when the monument was being constructed. Based on these observations, we experiment with multiple features to capture color, texture, shape and appearance information, and create strong baselines.

Being able to automatically identify visually characteristic elements in architectural scenes can open up the possibility of a whole new area of research. Such characteristics can be used to assist tasks in classification and retrieval of architectural style categories. As a stepping stone towards solving this problem, we propose a simple, yet effective approach to mine characteristic pairs of features which occur frequently across buildings in the same category and use them for improving classification performance. In all our experiments, we ensure that images of the same monument cannot be used for training as well as testing. This ensures that the results we obtain are not instances, but different buildings with similar architectural styles. The improvement in categorization further motivates us to explore higher-level, semantically meaningful features.

Research on characteristic features for architectural scenes is a fresh research direction. A new type of local descriptor, computed at symmetric points in images, has been proposed for matching architectural scenes in [10]. Doersch et al. [9] find visually repeating elements (windows, balconies and street signs) from Geo-tagged imagery of the city of Paris obtained from the Internet. Similar characteristics are obtained for different cities and those which are specific to Paris are identified. Using a patch-based method for finding visually-informative elements as in [9] restricts the features that are discoverable. The problem of discovering characteristic features across monuments is much more challenging due to (a) Multiple scales of characteristic features (b) Significant variation in appearance of the same visual characteristic across monuments (c) Multiple viewpoints and occlusion. Unlike [9], where discriminative patches containing small elements such as windows and street signs were mined, we mine characteristic features at multiple scales which are semantically meaningful and informative of a particular ar-

[1] http://en.wikipedia.org/wiki/Rose_window

Feature	K	Accuracy (in %)
Shape Context	1000	27.91
Geometric Blur 1	4000	35.935
Geometric Blur 2	4000	36.073
HOG	4000	32.90
HOG	1000	35.60
DoG + SIFT	4000	31.31
PHOW[27]	**4000**	**46.74**
Spatial Pyramid + DSIFT	84000	42.25
Dense SIFT	**4000**	**46.22**

Table 2: Baseline Experiments using multiple feature descriptors and detectors. In Geometric Blur 1 and 2, 4000 and 10,000 features were extracted from each image respectively.

chitectural style category. These characteristics can range from capitals, found in Renaissance Architecture[2] to huge windows with pointed arches found on Gothic monuments[3]. Characteristic features, represented by frequent sub-graphs, were mined using in [7], and used for improving classification of architectural scenes and product images. We categorize images into one of several categories of architectural styles. Our dataset comprises of large number of high resolution images of various monuments obtained from the Internet. The extent of characteristic features of architectural styles, if they do exist, is unknown.

2. IMAGE CATEGORIZATION

Being able to correctly discover characteristics which distinguish an architectural style category has several applications. These features can be used to assist classification and retrieval tasks in Computer Vision. We perform a set of experiments to show how characteristic features can be used to improve classification of an architectural style category. This can be a stepping stone towards using semantically meaningful characteristic features for categorizing buildings into one of several categories of architecture.

The Dataset of European Architectural Styles: To evaluate the performance while discovering characteristics of architectural styles, we have collected images of 25 different European monuments, which belong to one of five architectural styles. The membership of each monument into architectural style category was validated from Wikipedia. The images for each monument were downloaded from Flickr and Google Image Search. For each monument, its name in English and also in the language spoken in the region where it is located was used as a query while downloading images from the Internet. For example, both "Florence Cathedral" and "Basilica di Santa Maria del Fiore" were used as query to download images for this monument. From the downloaded images, those which provide an external view of the monument were retained and the rest discarded. Table 1 shows the categorization of the dataset into different architectural styles, further broken down into 5 different monuments for each architectural style. There are a total of 6713 high resolution images of size 640×480 pixels in the dataset. Classification experiments have been performed on the complete dataset.

[2]http://en.wikipedia.org/wiki/Renaissance_architecture
[3]http://en.wikipedia.org/wiki/Gothic_architecture

	Gothic	Korean	Georgian	Islamic
Visual Pattern [7]	93	76	79	77
Affine+SIFT	93.73	96.88	94.64	78.89
Dense SIFT	97.27	97.50	100	87.64

Table 3: Comparison of our baseline methods and our method of improving classification using Word Mining with the Visual Pattern Discovery method on their dataset for architectural image classification. [7].

Baseline Methods: To examine the utility of features such as shape, color and appearance for the purpose of categorization of architectural scenes, we experimented with multiple feature detectors and descriptors. To capture color information, color descriptors extracted over local regions and quantized into visual words were implemented. For capturing shape information, we used the HOG descriptor [8] and Geometric Blur [4]. HOG blocks were computed at regular intervals on the images and combined with a Bag of Words approach. Different sizes of the visual codebook (K=1000 and K=4000) were investigated in the case of HOG.

The Shape Context feature was first introduced by Belongie et. al [3] for matching silhouettes. For use in representing architectural images, we first extracted contours of all the images in the database. Small contours (based on the number of points) were rejected. Each contour was represented by the Shape Context feature computed over that contour at regularly sampled points. Each image was represented using a histogram of Shape Context descriptors, obtained by assigning Shape Context features to Visual Words (K=4000).

To capture appearance information, we used the SIFT descriptor [14]. Several interest point detectors (Hessian-Affine [16], Difference of Gaussian [14], MSER [15], Dense Sampling) were coupled with SIFT and evaluated. Table 2 shows the classification results using the baseline methods. SIFT descriptors computed on a dense grid outperformed the other features and were used for the representation of monument images in our dataset of architectural style categories. The SIFT descriptors were assigned to visual words from a pre-trained vocabulary on a randomly sampled subset of features from the database. The visual vocabulary is obtained using K-means algorithm (K=4000) run 8 times with different random initialization, and keeping the cluster with the lowest energy. SVM classifier was used to perform the classification experiments. We ensure that images of the monument from which the candidate window was generated is absent in the database. This shows how well monuments can be categorized based on architectural styles.

We compare our method of classification with the approach of [7], who also work on the classification of architectural scenes. For this we use their dataset [4] of architectural images. The dataset consists of 423 images, including 111 Gothic images, 156 Korean images, 75 Georgian images, and 81 Islamic images. A 10-fold cross-validation scheme is used. For each fold, 30 images are randomly selected from each class as the training images, and the remaining is for testing. We show results using multiple methods of image representation. Table 3 summarizes the results. We can

[4]http://www.cs.ccu.edu.tw/ wtchu/projects/VP/index.html

Table 1: European Monuments Dataset

Art Noveau	Baroque	Gothic	Renaissance	Romanesque
Casa Batllo	Basilica de Superga	Abbey of St. Denis	Florence Cathedral	Mainz Cathedral
Casa Lleo Morera	Church of the Gesu	Chartres Cathedral	San Pietro	Mosiac Abbey
Casa Mila	Elector's Palace	Notre dame de Paris	Sant Agostino	Notre dame de Puy
Elisabeth's Church	Queluz National Palace	Salisbury Cathedral	Santa Maria, Novella	Peterborough Cathedral
Sagrada Familia	Santa Susanne,Rome	York Minster	St. Peter's Basilica	Sant Ambrogio

see that we have created strong baselines, which outperform their classification performance. The average accuracy obtained using Visual Pattern Discovery [7] is 81%. Using our baseline of Dense SIFT features, we achieve an average accuracy of 95.6%. This demonstrates the superior performance of our method.

2.1 Mining Pairs of Visual Words Occurrences

We propose a method to mine characteristic pairs of visual words for each architectural category and improve classification. Using higher order feature groups for image classification and object discovery is not new [17, 18, 29, 13, 30]. The utility of doublets and triplets of visual words has been explored for the task of object categorization in [13]. Higher order feature configurations were mined [21] which occurred frequently on instances of a given object class. However, an accurate bounding box enclosing the object was required in the images, from which discriminative feature configurations are mined. We propose a simple, yet effective method to mine doublets of visual words discriminative for each class. Our method works with complete images rather than bounding boxes localizing the buildings in each image.

Using a combination of local features preserves spatial context, which is lost in traditional Bag of Words method. We first create a 2-d histogram of visual word occurrences for each category. The size of the histogram is thus $K \times K$, where K is the number of visual words in the vocabulary. This histogram captures, for each visual word in the codebook, it's co-occurrence with neighboring visual word in in images of the same category. This is done by moving a sliding window of fixed size over each image. The 2-d class histogram is updated with the counts of all pairs of visual words that occur in the sliding window.

The counts of bins, across the diagonal, corresponding to the same pair of visual words are summed. Thus, the upper triangular histogram contains complete co-occurrence information for the class.

$$H_c(i,j) = H_c(i,j) + H_c(j,i), i < j \quad (1)$$

We now have a 2-d histogram H_c which encapsulates the co-occurrence of visual words for category c. The discriminative power of each pair of visual words $d(v_{c,i}, v_{c,j})$ for class c is computed as

$$d(v_{c,i}, v_{c,j}) = |H_c(i,j) - \frac{1}{(NC-1)} \sum_{C=1, C \neq c}^{NC} H_C(i,j)| \quad (2)$$

This can be computed for each class in a single step, since the frequency of all pairs of visual words for a given class are stored in a single histogram.

$$dH_c = |dH_c - \frac{1}{(NC-1)} \sum_{C=1, C \neq c}^{NC} H_C| \quad (3)$$

Method	Accuracy
Zhang et. al[30]	78.3%
Word Mining	80%
QPC[17]	81.8%
QPC+Sel[17]	80.8%
Sgl[17]	81.7%
LPC[17]	83.9%

Table 4: Comparison of our method with recent approaches in improving classification using mined word pairs.

2.2 Utility of Pairs of Visual Words for Image Categorization

The top P visual word pairs with the highest discriminative score are retained for each architectural style category. The initial histogram representation for each image in the database is augmented with an extended histogram of length $P \times NC$, where NC is the number of categories in our dataset. For each image, this extended histogram contains the frequency of the mined pairs in that image. Classification is performed using the augmented image representations.

Comparison with Recent approaches: Our approach of mining visual word pairs is comparable with recent approaches which use mined higher order features to improve classification accuracy. The effectiveness of 2^{nd} and 10^{th} order features was explored by Zhang et. al [30]. The correspondence between the same n^{th} order feature across two images is computed. This is done by transforming the local features into offset space. For a visual word (w, r_1, r_2), where w is the visual word, r_1 is the location of the word in one image, and r_2 is the location of the word in the second image, the position of the word in the offset space is computed as

$$\Delta r = (\Delta x, \Delta y) = (x_1 - x_2, y_1 - y_2) \quad (4)$$

Based on the location of words in the offset space, a kernel is computed which is then used for classification using KNN and SVM. We performed experiments on the MSRC v2 dataset[28]. The experimental setup was the same as used in [30]. The size of the sliding window was set to 24×24 pixels, and $P(= 200)$ most discriminative word pairs were used for each class. Table 4 shows how our approach compares with some of the recent approaches. The length of our histogram representation is 5000 $(3200 + 9 \times 200)$. Using this, we obtain better performance than [30], and comparable performance to QPC[17], and QPC + Sel[17], where the number of pairwise feature clusters used is much larger.

Comparison with Visual Pattern Discovery [7] [7] propose a method to discover visual patterns in the data and mine sub-graphs of frequently occurring patterns using

Figure 3: Top 5 retrieval results for a randomly picked image of San Pietro (Renaissance Architecture). The image outlined in black (left) is the query image. The images outlined in green are the top 5 retrieved images of monuments other than San Pietro, but of the same architectural style. The image outlined in red is a false positive (Romanesque Architecture)

	Gothic	Korean	Georgian	Islamic
Visual Pattern [7]	93	76	79	77
Dense SIFT	97.27	97.50	100	87.64
Word Mining	98.18	99.38	98.57	97.50

Table 5: Comparison of our method of improving classification using Word Pair Mining with the Visual Pattern Discovery method [7].

Method	Accuracy
Word Mining	32.64%
SVM	46.22%
SVM + Word Mining	48.25%

Table 6: Classification Performance on Dataset of European Architectural Styles

a graph mining approach. These visual patterns are then used to improve classification performance. We improve our baselines on their dataset using our method of mining discriminative word pairs for each category. The results are summarized in Table 5. The size of the sliding window was set to 36×36 pixels, and $P(= 150)$ most discriminative word pairs were mined for each class. Using our approach of mining discriminative pairs of visual words, we are able to bypass our baseline results using Dense SIFT features. We obtain an average accuracy of 98.41% over four classes.

Classification Results The dataset used in [7] is not sub-categorized into monuments. Hence, different images of the same monument can lie in the training and testing sets. Also, the characteristic features to be discovered for a particular architectural style category are present in all the images in the architectural style category for this dataset [7]. We create a much more challenging dataset to evaluate the importance of discovering characteristic features for architectural styles. We ensure that images of the monument from which the candidate window was generated is absent in the database. This ensures that the discovered structures are characteristic to a particular architectural style, and not mere instances of the same building. Table 6 shows the classification performance on our dataset. The parameters are the same as used in our previous experiment for comparison with [7]. An interesting observation is the classification accuracy obtained using only the 1000 length histogram of visual word pairs for image representation. The high accuracy (32.64%) shows the utility of doublets of visual words for image categorization. The low classification performance as compared to [7] is mainly due to two reasons - the challenging nature of our dataset, and ensuring that images of the same monument cannot lie in both training and testing sets simultaneously.

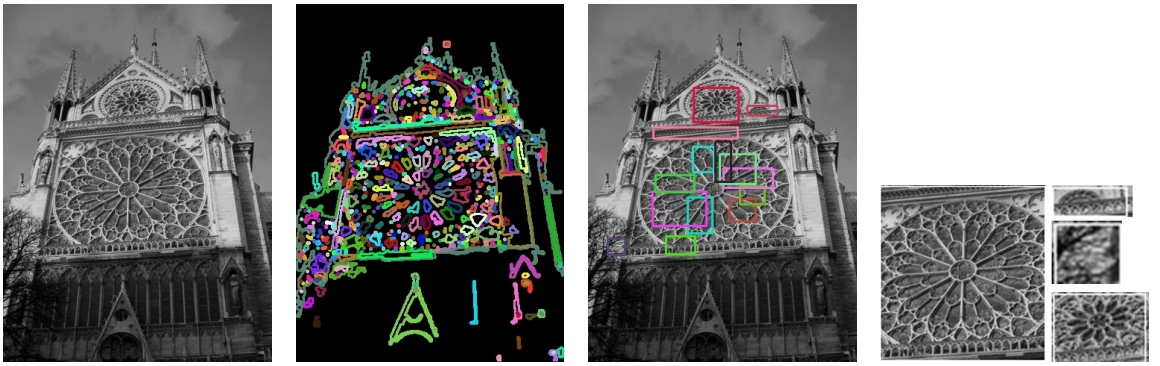
2.3 Word Pairs for Image Retrieval

In this section, we evaluate the utility of the mined pairs of visual words for a particular architectural style in retrieving images of the same architectural style category. Similar to image categorization, a Bag-of-Words representation using SIFT keypoints computed on a densely sampled grid are used for image representation. The final representation of the image is a 4000 length L1-normalized histogram. For each architectural style category, 10 samples are randomly selected for one monument for querying. The images of the rest of the monuments are put in the database. A simple nearest neighbor method is used for retrieval. The similarity between two image histograms is computed using the Hellinger Kernel [2], which has shown to give superior performance in image retrieval. To evaluate the retrieval performance, the Mean Average Precision (MAP) is computed over all query samples. The baseline MAP is 22.39%, which increases to 22.95% when the histograms are appended with the mined word pairs for each category. Figure 3 shows the top 5 retrieved results using pairs of visual word occurrences for a random image from Renaissance Architecture. The improvement in both image categorization and retrieval tasks, however small, is encouraging. We believe that the key to recognition of architectural style categories lies in the larger and semantically meaningful features. We now propose a method to discover such characteristics.

3. DISCOVERING SEMANTIC PATTERNS

Starting with a large collection of architectural images, we wish to discover the characteristics that make monuments built according to a particular architectural style distinguishable from other monuments which might follow a different style. This is similar to the problem of Object Category Discovery, which has been previously addressed by several researchers in Computer Vision. Given a large dataset of unlabeled images, the objective is to automatically determine the visually similar categories. Borrowed from the text mining literature, techniques such as probabilistic Latent Semantic Analysis (pLSA) [11] and Latent Dirichlet Allocation (LDA) [5] for Topic Discovery have been investigated in [23, 24].

We first generate a large set of candidates which serve as potential features for an architectural style category. A possible approach to obtain such a candidate set would be to randomly sample all the images in the database at multiple spatial scales, and exhaustively search through this huge



(a) Original Image (b) Contour Map (c) Minimum Area Bounding Boxes around medium-sized contours (d) Final seeds obtained after clustering

Figure 4: Generating candidate seeds from an image of the monument - Notre dame de Paris

set. However, the number of windows generated in such a manner can be in the order of millions! Our method efficiently filters most of the irrelevant and redundant candidate windows. Inspired by the method used in [22], we extract multiple segmentations from the dataset images guided by segmentation cues. For our database of 2001 images, used for discovery, we obtained a candidate set of 18,127 seeds. The "seeds" are then used for mining clusters of visually similar elements from the database, using a Bag of Words retrieval pipeline. Restricting the size of each cluster to be 100, we still have a huge set of $18,127 \times 100$ patches divided into clusters. The clusters are pruned using a spatial verification step. Based on the number of elements in each cluster, and their SIFT [14] matches with the cluster seed, visually informative clusters are ranked higher. The top few clusters are further refined using LBP features.

A "seed" is a potential architectural feature. We propose a method to generate a reasonable number of seeds guided by segmentation cues. Characteristic features can be of varying shapes and sizes. Candidate windows should be able to completely capture a particular feature in a building. First, edge maps are computed for all images in the dataset using a Canny Edge detector [6]. Closed contours are discovered from the binary edge map using the method of [26]. Each contour consists of the end points of the vertical, horizontal and diagonal line segments it contains. Contours are filtered such that

$$\theta_1 < N < \theta_2 \tag{5}$$

where N is the number of end points for the contour. The value of θ_1 and θ_2 has been set to 20 and 500, respectively, in our experiments. This removes most of the noise from the binary image, which may be in the form of people or trees or vehicles. These contours are represented by a bounding box of minimum area enclosing the contour. Any candidate windows generated using these bounding boxes are bound to completely capture a given structure in the image. Often large structures consisting of several smaller sub-structures may not be completely captured. Overlapping bounding boxes are iteratively grouped which results in a single large bounding box for the set of smaller boxes. The clustering procedure is terminated once no overlapping bounding boxes are present. The seeds obtained using the above method are extracted for all images in the dataset, with a slack of 10 pixels in the length and width of the boxes. Figure 2.2 gives an overview of the pipeline of generation of candidate windows for one image from the database. The final bounding boxes are segmented from the images. These potential features are now used as seeds for mining similar features across monuments. These features should occur uniformly over all monuments of the same architectural category, and rarely on monuments of different categories.

Image Representation Images in our architectural image database, as well as the potential features generated in the previous section, are represented using SIFT [14] descriptors computed at affine-invariant Hessian regions [16]. Randomly sampled SIFT descriptors from a subset of the images in the database are used to construct a visual vocabulary, which is then used to assign visual word ids to the descriptors. We build an inverted index from the database images which is used to compare vector representations of database images with the potential features. A standard tf-idf weighting scheme is employed.

Mining Characteristic Features Each potential feature generated earlier is used as a seed for mining characteristic features. Visually similar characteristics are searched for in the database of images using an inverted index. Due to the large number of generated "seeds", we allow for soft-assignment of database images into more than one clusters. In this phase of clustering, we allow for no more than 100 images per cluster of characteristic feature. The clusters obtained are refined by geometrically verifying the mined characteristics and the initial seeds.

Spatial Verification: An affine transformation with 3 degrees of freedom (dof) is fitted between the cluster seed and images falling into the cluster. First, a set of correspondences between SIFT descriptors from the cluster seed and a cluster image are obtained. Similar to [19, 1], hypotheses are generated from only a single pair of correspondences, which has shown to speed up matching and reduce the number of hypotheses generated. While evaluating the generated hypotheses, we allow for large re-projection errors. This is because the appearance of characteristic features can vary significantly across images of different monuments, even though they correspond to the same semantic category of characteristic feature. Images with very few number of inliers obtained from matching with the cluster

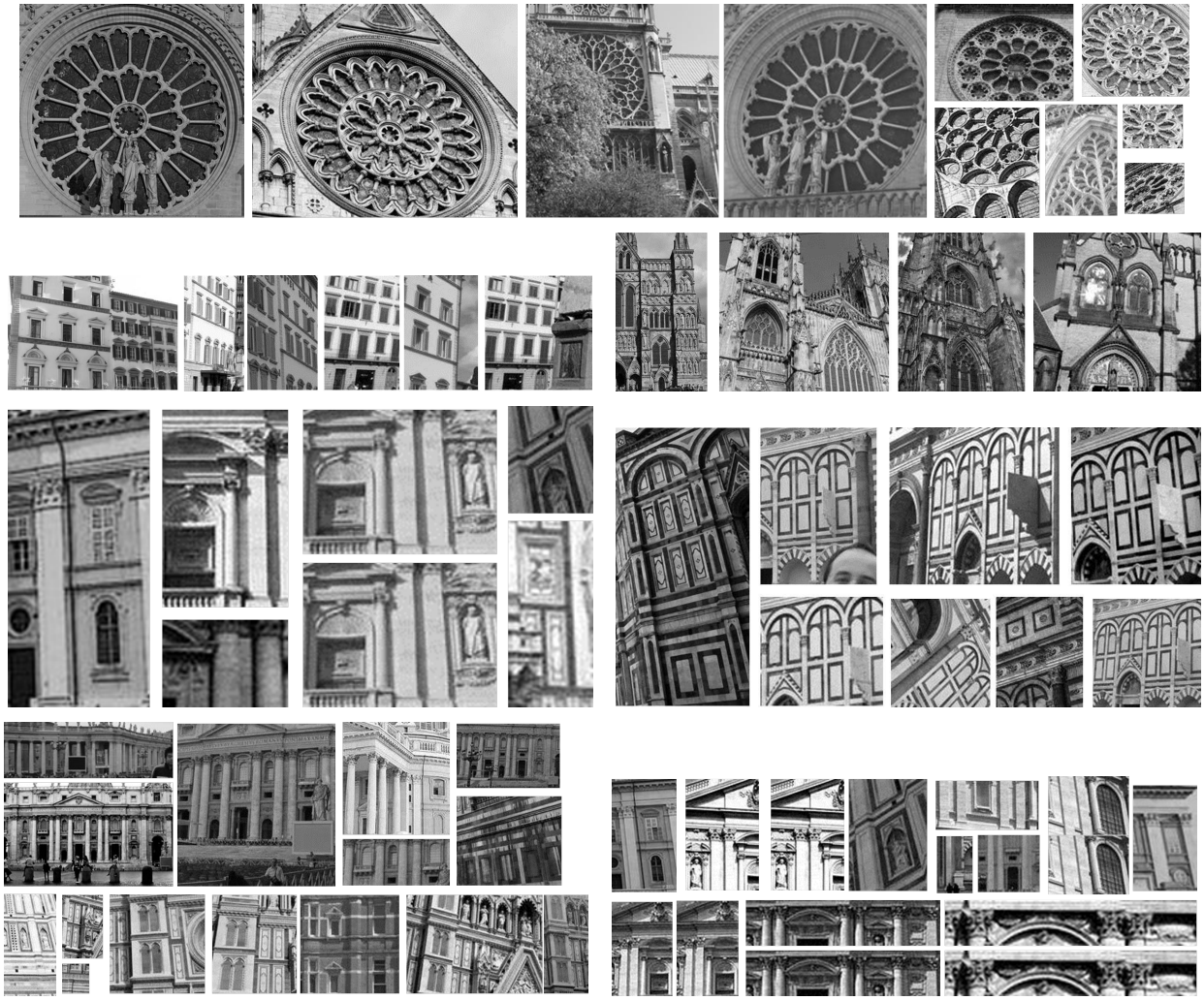


Figure 5: Characteristic features mined by our method for various Architectural styles. Row 1: Rose windows in Gothic architecture. The cluster seed is from the monument, "Notre Dame de Paris". Similar windows were found in images of other Gothic monument - York Minster, Notre Dame de Chartres and Abbey of St. Denis. Row 3: Windows with semi-circular arch and semi-circular arches in monuments of Renaissance architecture. Row 4: Regular spaced columns and colonnades of two columns in Baroque Architecture

seed are removed from that cluster.

Refinement: The clusters obtained are further refined to remove semantically irrelevant images. For this, we use the LBP feature descriptor. The LBP for a location (x, y) is a string of eight bits, Each bit corresponds to one of the 8-neighbors and is equal to one if it is brighter than the central location. For example, the first bit is one if, and only if,

$$I(x+1, y) > I(x, y) \tag{6}$$

The patches are represented using a histogram of LBP features. The distance between the cluster seed and other elements in the cluster is computed. The score of the cluster is computed as a sum of rankings obtained by (a) SIFT matching, and (b) LBP distance.

Results: The clusters obtained as a result of mining and pruning are analyzed. The top few clusters obtained are visualized in Figure 3. Rose windows found across various Gothic monuments by our algorithm have been shown in Figure 5(a). Another characteristic of Gothic monuments are windows with pointed arches as shown in Figure 5(c). Figure 5(d) and Figure 5(e) show semi-circular windows and arches, which are often found on the facades of monuments of Renaissance architecture. Monuments in Baroque architecture are characterized by a dynamic rhythm of columns (Figure 5(f)) and colonnades of two columns at regular intervals (Figure 5(g)). All these results have been verified from Wikipedia.

4. CONCLUSIONS

In this work, we explore architectural style categories. We start by identifying the nature of the problem of categorizing buildings into architectural styles categories, and the challenges one might face while solving it. This has been achieved using a set of comprehensive baseline experiments using multiple features. We have evaluated the utility of low-level features in improving the classification performance.

Figure 6: Characteristic Features in Romanesque Architecture which cannot be captured by our method.

The results are encouraging, and motivate us to look for larger and semantically meaningful characteristics. We have proposed a method to identify characteristic features for architectural style categories in an unsupervised fashion. The mined characteristic features are visually informative, and representative of architectural style categories, as verified from Wikipedia. We are, however, limited by the features we employ. There are several characteristic features such as the height of the monument, the plan of the building, or internal features which are difficult to capture. Figure 6 shows arched vaults which are commonly found in Romanesque monuments[5], but does not show in our results. Many many more such characteristics is a promising research direction which can be explored by new researchers, who can make the transition from buildings as instances to buildings as categories.

5. ACKNOWLEDGMENTS

This work was partly supported by Indian Digital Heritage project of DST

6. REFERENCES

[1] R. Arandjelović and A. Zisserman. Smooth object retrieval using a bag of boundaries. In *ICCV*, 2011.

[2] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012.

[3] S. Belongie and J. Malik. Matching with shape contexts. In *IEEE Workshop on Content-based Access of Image and Video Libraries*, 2000.

[4] A. Berg, T. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *CVPR*, 2005.

[5] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.

[6] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1986.

[7] W.-T. Chu and M.-H. Tsai. Visual pattern discovery for architecture image classification and product image search. In *ACM ICMR*, 2012.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[9] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 2012.

[10] D. Hauagge and N. Snavely. Image matching using local symmetry features. In *CVPR*, 2012.

[5]http://en.wikipedia.org/wiki/Romanesque_Architecture

[11] T. Hofmann. Probabilistic latent semantic indexing. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999.

[12] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008.

[13] D. Liu, G. Hua, P. Viola, and T. Chen. Integrated feature selection and higher-order spatial feature extraction for object categorization. In *CVPR*, 2008.

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 2004.

[15] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image Vision Comput.*, 2004.

[16] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *ECCV*, 2002.

[17] N. Morioka and S. Satoh. Building compact local pairwise codebook with joint feature space clustering. In *ECCV*, 2010.

[18] N. Morioka and S. Satoh. Compact correlation coding for visual object categorization. In *ICCV*, 2011.

[19] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[20] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008.

[21] T. Quack, V. Ferrari, B. Leibe, and L. V. Gool. Efficient mining of frequent and distinctive feature configurations. In *ICCV*, 2007.

[22] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006.

[23] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering objects and their location in images. In *ICCV*, 2005.

[24] J. Sivic, B. Russell, A. Zisserman, W. Freeman, and A. Efros. Unsupervised discovery of visual object class hierarchies. In *CVPR*, 2008.

[25] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[26] S. Suzuki and K. Abe. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 1985.

[27] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, 2008.

[28] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, 2005.

[29] L. Yu, J. Liu, and C. Xu. Descriptive local feature groups for image classification. In *ICIP*, 2011.

[30] Y. Zhang and T. Chen. Efficient kernels for identifying unbounded-order spatial features. In *CVPR*, 2009.

[31] Y. Zhang, Z. Jia, and T. Chen. Image retrieval with geometry-preserving visual phrases. In *CVPR*, 2011.