Home Interior Classification using SIFT Keypoint Histograms

Brian Ayers, Matthew Boutell
Rose-Hulman Institute of Technology
Terre Haute, IN 47803
{ayersbw, boutell}@rose-hulman.edu

Abstract

Semantic scene classification, the process of categorizing photographs into a discrete set of classes using pattern recognition techniques, is a useful ability for image annotation, organization and retrieval. The literature has focused on classifying outdoor scenes such as beaches and sunsets. Here, we focus on a much more difficult problem, that of differentiating between typical rooms in home interiors, such as bedrooms or kitchens. This requires robust image feature extraction and classification techniques, such as SIFT (Scale-Invariant Feature Transform) features and Adaboost classifiers. To this end, we derived SIFT keypoint histograms*, an efficient image representation that utilizes variance information from linear discriminant analysis. We compare SIFT keypoint histograms with other features such as spatial color moments and compare Adaboost with Support Vector Machine classifiers. We outline the various techniques used, show their advantages, disadvantages, and actual performance, and determine the most effective algorithm of those tested for home interior classification. Furthermore, we present results of pairwise classification of 7 rooms typically found in homes.*

1. Introduction and related work

Automatically annotating photographs based on their contents is a useful ability, leading to the ability to organize people's digital shoeboxes and retrieve images from them. Thus far, the literature has focused on outdoor scenes. Distinguishing between common rooms in *indoor* environments such as homes is a much more challenging problem. Features typically used for outdoor scenes in the literature, such as spatial color moments [2][10], often fail in indoor environments because colors are a weak predictor of room type. Also, home interiors contain much extraneous data common to most or all room types that is useless for classification: homogeneous areas corresponding to blank walls, floors, and ceilings and generic "objects" such as doors or corners. What is needed is a feature that captures those features that are present in some rooms and not others.

In this paper, we demonstrate a technique that uses SIFT features [7] in an attempt to obtain the most useful information from the images. We present *SIFT keypoint histograms*, a feature derived by first applying linear discriminant analysis (LDA) and then populating a histogram which uses the variance information obtained from LDA to create a low-dimensional representation of each image. It is important to summarize keypoints using frequency information because of the many extraneous features.

Johnson and Cipolla devised a method for pruning extraneous keypoints, but utilize information from successive frames in a video sequence [5]. Ke and Sukthankar also successfully applied dimensionality reduction techniques (principal components analysis, or PCA) to SIFT features, but for the purpose of individual object detection, not scene classification [5]. SIFT features have been used to classify indoor scenes [8], but for navigation within a constrained environment (rooms within a single building). The novelty of this work, by contrast, the attempt to distinguish among *classes* of rooms in general, unconstrained indoor photographs.

We experimented with the problem of distinguishing between 7 room types typically found in many homes: bathrooms, bedrooms, dining rooms, hallways, kitchens, living rooms, and studies. On a set of 295 such images, SIFT keypoint histograms proved to be the most effective feature of all those tested. In an experiment comparing all possible pairs of rooms, accuracies varied greatly, with a maximum of approximately 94% when distinguishing between hallways and dining rooms.

2. Raw feature extraction

SIFT (Scale-Invariant Feature Transform) keypoints [7] are invariant to rotation, scale, and translation, and are used to detect distinctive edges and textures in an image. Each keypoint has 132 dimensions: 128 spatial orientation bins, plus the coordinates, scale, and rotation of the keypoint. Figure 1 is an example of the keypoints

1-4244-1180-7/07/$25.00 ©2007 IEEE

extracted from a photograph of a hallway using SIFT. SIFT features do not take color into account, but for the purposes of this problem, color is not very useful.

Spatial color moments are a state-of-the-art feature used to distinguish outdoor scenes [2][10]; we use them as a baseline feature for comparison, even though color is expected to be more salient for outdoor scenes than indoor ones. As in [2], we convert the image to LST space and split the image into blocks formed by an NxN grid. We then compute the mean and variance of each block's color band. We use a 4x4 grid in our tests, yielding a 16 x 3 x 2 = 96-dimension feature.

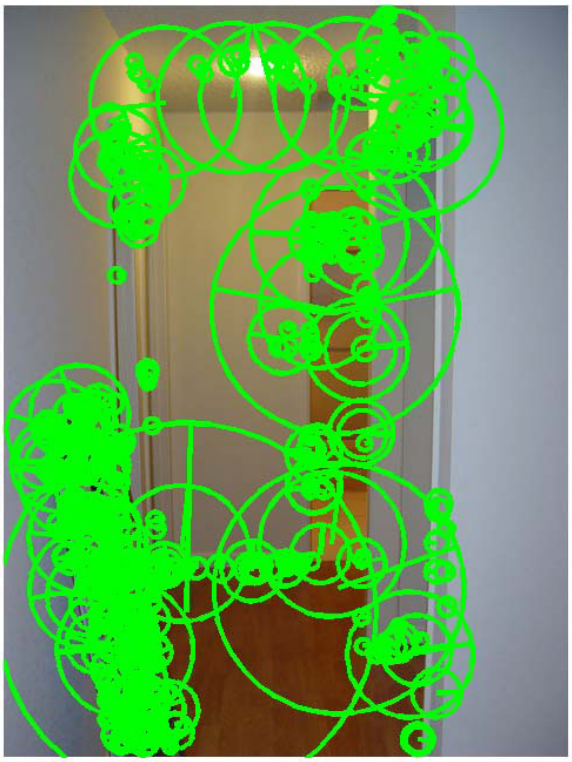

Figure 1: SIFT keypoints for a photo of a hallway. Note how most of the SIFT keypoints are located in the bottom left corner, where they cover a piece of cardboard. Unfortunately, they do not help classify the room.

3. Data reduction

Each photograph in our collection typically contains over 1000 SIFT keypoints, each with 132 dimensions. One way to use these keypoints for scene classification is to classify each keypoint independently, and then classify the image with the same class assigned to the majority of its keypoints. In our set of approximately 300 images, this means over 300,000 keypoints will be classified, many of which do not vary much per image. Therefore, we needed to reduce dimensionality or prune the number of detected keypoints per image before using them for classification. We tried multiple approaches to this: PCA and LDA for dimensionality reduction, and distance-based pruning, class-based pruning, k-means clustering, and histograms to reduce the number of keypoints. We discuss the most successful approaches in this section, reserving discussion of the other choices for the Appendix.

3.1 Dimensionality reduction

Linear discriminant analysis (LDA) [3] exploits class information, maximizing between-class scatter, measured by scatter matrix S_B, while minimizing within-class scatter, measured by scatter matrix S_W. This is done by maximizing the ratio:

$$J(w) = \frac{w^t S_B w}{w^t S_W w} \quad (1)$$

over projection vectors w.

LDA reduces the data set to a dimensionality equal to one fewer than the number of classes, C. The (C-1) dimensions are sorted by the amount of the data set's variance they capture. As expected, LDA improved classification accuracy significantly on this problem (Figure 2).

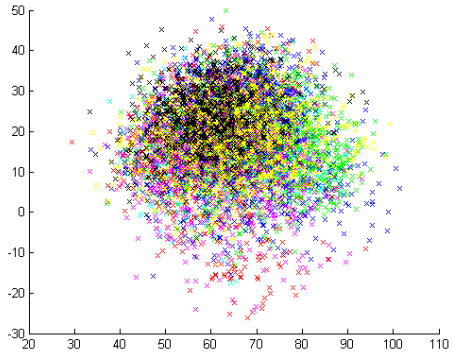

Figure 2: The first 2 dimensions of the data after performing LDA on it, color coded by class.

3.2 Keypoint histograms

Dimensionality reduction does not take into account the many extraneous features common to all of the images. Our attempts to prune keypoints, documented in the Appendix, were mostly unsuccessful. An alternative to classifying each keypoint independently is to create a *keypoint histogram* for each image from the 1000+ keypoints detected in that image. This technique allows the relationships between numbers and types of keypoints in an image to be extrapolated.

We started by applying LDA to each keypoint, as discussed in Section 3.1. For the 7-class problem described in the introduction, this yielded 6 dimensions for each keypoint. Histograms typically have high dimension to preserve granularity of the input, which

causes classification, particularly on small data sets, to suffer from the so-called "curse of dimensionality". To maximize granularity while keeping a small number of dimensions, we optimized the allotment of bins by data variance, allocating a higher number of bins to those dimensions that exhibited higher variance on the data set, as calculated by the LDA algorithm. Specifically, the number of bins for each dimension was chosen to be approximately proportional to the data variance in that dimension, since dimensions with a higher variance required more bins to faithfully represent the data. We found experimentally that performance was highest when the dimension with the highest variance had an even number of bins, as this likely helped to split the data apart in its densest area, the center. We found that a 48-dimension histogram with 4, 3, 2, and 2 bins for dimensions 1-4 was the most effective. Note we ignored the 2 least important dimensions due to the small size of our data set. For larger data sets, one can increase the number of dimensions some while avoiding the curse of dimensionality. But for this set, our choice of histogram allowed for the maximum preservation of information in the histogram without sacrificing performance.

4. Classifiers

We experimented with three classifiers: a nearest mean classifier, support vector machines (SVM), and Adaboost.

Of these three, the nearest mean classifier is the simplest and is also the only one that does not use keypoint histograms. We start by breaking the SIFT keypoints found for each class into a number of clusters using an adaptive *k*-means algorithm described in the Appendix. We then represent each class by the means of its set of clusters. The nearest mean classifier then lets every SIFT keypoint in a test image cast a vote for the class that has the closest mean to it. The test image is classified as the class with the most votes in the image.

Figure 3 shows an example of the clusters formed from SIFT keypoints extracted from the bathroom images. Support vector machines are more robust than the previous classifier. SVMs find a hyperplane that maximizes the margin between the classes. Kernel functions are used to project the data to a higher dimension to make the data more easily separable in that dimension; we used the radial basis function (RBF) kernel.

The final classifier used was Adaboost (Adaptive Boosting) [4]. Adaboost works by consulting and weighting various weak learners (which are any type of simple classifier that has a classification rate above 50%); we used decision stumps. It iteratively trains each weak learner on randomly-chosen subsets of the training data and then compares their accuracy for classification. The weak learners that are more accurate are weighted higher than others. We used the ModestBoost variant [11] in our tests.

Figure 3: The keypoints of a bathroom image are clustered into 3 different groups by using *K*-means segmentation. The original image is shown (top left) along with the 3 different clusters of keypoints, each in their own image.

Adaboost seems to be the most useful classifier for this problem. It is particularly resilient to over-fitting, and consumes less memory than SVM, another important point when SIFT keypoints tend to be numerous.

5. Experimental results

The test set consists of 295 images (30 bathrooms, 85 bedrooms, 41 dining rooms, 23 hallways, 50 kitchens, 43 living rooms, and 23 offices or studies) collected from a variety of online sources, such as realtor websites. Images with people in them or with other anomalies such as black borders around the image were filtered out. We attempted to acquire a wide-range of images from each room type. All tests used leave-one-out cross validation.

The first test compared feature and classifier combinations, namely (1) color moments with SVM, (2) keypoints reduced by LDA with nearest means, (3) keypoint histograms with SVM, and (4) keypoint histograms with Adaboost. In general, keypoint histograms with Adaboost performed best. For the full 7 class problem, they correctly classified the images 38.1% of the time. This is much higher than the baseline obtained by using color moments, 23.8%, and the theoretical random, 14.3%. It is also significantly higher than the accuracy obtained using SIFT features with other classifiers: SVM (33.7%) and nearest-means (32.3%)

We then performed the same test while varying the number of rooms, *n*, between which to distinguish (Figure 4). For each value of *n*, we repeated a test in we chose *n* rooms and then trained a classifier to distinguish between those *n* rooms. When the number of classes is reduced, all

features and classifiers had a higher-than-random accuracy in every trial, and keypoint histograms with Adaboost always had the highest accuracy. For the 2 class problems, they had a 77.7% classification accuracy overall.

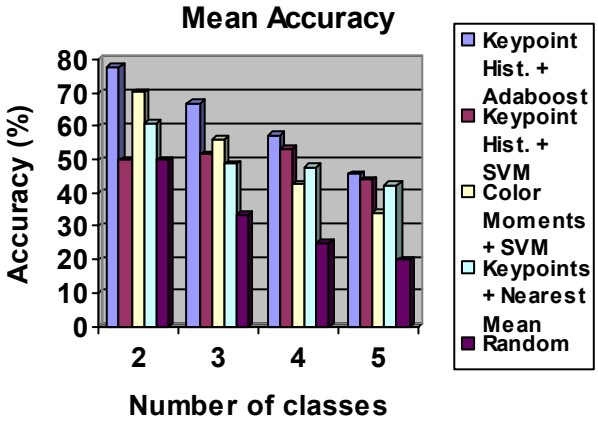

Figure 4: Mean accuracy of each type of classification technique (along with the theoretical random classification rate for comparison) for different numbers of randomly selected classes.

Once determining which features and classifiers work best, an interesting question to ask is which rooms were easiest and most difficult to classify. We therefore ran a second experiment in which we performed binary classification on all pairs of classes. We first precomputed the keypoint histograms for all images. We then selected two rooms and trained an Adaboost classifier on all the images from those two rooms. We found a great deal of variance in the classification accuracy depending on the two room types being compared (Table 1). The rooms with the most similar SIFT keypoints were studies and living rooms; this was the only pair that performed worse than random. Hallways and dining rooms proved to be the easiest for the classifier to tell apart, having almost 94% classification accuracy. This makes sense, given the similarity between the objects in living rooms and studies (both have chairs, tables, etc) versus the extreme difference between hallways and living rooms (only flat walls and doors in hallways).

Figure 5 contains examples of images that were successfully or unsuccessfully classified. The first image contains all of the classic features of a living room and would be expected to be correctly classified, although the second correct living room image classification is surprising, since the only entirely visible object is the fireplace. We conjecture that the failed living room classification is because the many shadows on the floor add too much noise to the keypoint histogram. The bedroom image on the left is clearly dominated by the bed, probably explaining its successful classification. The image on the right, classified incorrectly, shows a great deal of variance from the typical bedroom in its details. The many extra objects and frills likely made the keypoint histogram too noisy for a correct classification, much like the shadows in the living room example. Although the classifier has a difficult time with extravagant or poorly photographed rooms that vary too far from the "average" room of that class, it does allow for a fair deal of variance, as shown in the middle bedroom image.

7. Conclusions and future work

Indoor room classification was shown to be extremely difficult using low-level features. SIFT keypoint histograms classified using Adaboost were to be the most robust classification algorithm of those tested, achieving accuracy approximately twice that obtained by chance. While still unable to perform highly accurately on the full 7 class problem, they achieve a respectable accuracy with 2 and 3 classes.

Future work involves using spatial information explicitly. An attempt to spatially cluster SIFT features in an image could allow for finding objects in a room; for instance, a chair in front of a blank wall would be spatially clustered together as an object. These clusters are a type of semi-supervised object detector [9]; some clusters of similar types within each class would correspond to the objects that are in most images of that room type. Finding these in a novel image would then be evidence for that room type, particularly in rooms that tend to be defined by the presence of a few objects. Alternately, an explicit object-finding approach (using *domain-specific* detectors for tables, beds, sinks, etc), followed by probabilistic scene modeling, as was done with outdoor scenes [1], may prove be more powerful.

References

[1] M. Boutell, J. Luo, and C. Brown. Factor-graphs for region-based whole-scene classification. In *Proc International Workshop on Semantic Learning Applications in Multimedia (in conjunction with CVPR2006)*, New York, NY, 2006.

[2] M. Boutell, A. Choudhury, J. Luo, and C. Brown. Using semantic features for scene classification: How good do they need to be? *In Proc IEEE International Conference on Multimedia and Expo*, Toronto. 2006.

[3] R. Duda, R. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, Inc., New York, 2001.

[4] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *In Proc. Second European Conference on Computational Learning Theory*, 1995. p. 23-37.

	Bedroom	Dining Room	Hallway	Kitchen	Living Room	Office or Study
Bathroom	74.8	80.0	X	60.0	83.6	X
Bedroom		85.6	85.2	85.2	70.3	85.2
Dining Room			93.7	80.0	69.9	79.4
Hallway				82.2	83.3	X
Kitchen					66.7	71.2
Living Room						62.1
Office/Study						

Table 1: Classification accuracy (%) using SIFT keypoint histograms with Adaboost for each pair of rooms. The three pairs that had too few images to test are marked with an X.

Figure 5: Samples of successful and unsuccessful classifications on the full 7-class problem. The first row contains living room photos, while the second row contains bedroom photos.

[5] M. Johnson and R. Cipolla, Stable interest points for improved image retrieval and matching. Tech. Report TR20060916, University of Cambridge, Cambridge, UK, 2006.

[6] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *Proc IEEE Internation Conference on Computer Vision and Pattern Recognition*, Washington, D.C., 2004.

[7] Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60, 2: 91-110 (2004).

[8] L. Ledwich and S. Williams. Reduced SIFT features for image retrieval and indoor localisation. In *Proc. Australasian Conf. on Robotics and Automation (ACRA)*, 2004.

[9] C. Rosenberg and M Hebert. Training object detection models with weakly labeled data. In *Proc. British Machine Vision Conference*, 2002.

[10] A. Vailaya, M. Figueiredo, A. Jain, and H. J. Zhang. Content-based hierarchical classification of vacation images. In *Proc. IEEE Multimedia Systems*, 1999.

[11] A. Vezhnevets, V. Vezhnevets Modest AdaBoost - Teaching AdaBoost to generalize better. Graphicon-2005, Novosibirsk Akademgorodok, Russia, 2005.

Appendix: Other data reduction options

Here we briefly document a number of alternative dimensionality-reducing and clustering algorithms that we tried.

Principal components analysis (PCA) [3] can be used to reduce the dimensionality of the data, as done by Ke and Sukthankar [5]. PCA uses the eigenvectors of the covariance matrix of all the feature vectors to re-orient the coordinate system to maximize the variance using the fewest number of dimensions. The dimensions are sorted by the amount of variance exhibited in that dimension. Applying PCA on our data set reduced the dimensionality to under 50% of the original while still preserving approximately 90% of the data variance. (It could further reduce the dimensionality to under 20% of the original while keeping about 80% of the variance.) However, since PCA does not take class information into account, it was not very effective.

As stated in Section 3.2, dimensionality reduction does not take into account the many extraneous features common to all of the images. In an attempt to remove as many of those as possible, we first employed a simple distance-based data pruning algorithm. Any pair of SIFT features with distances in the feature space below a threshold were both removed completely. This caused any large clusters of SIFT features, which we assumed to be common to all classes, to be removed (Figure 6). Although this slightly helped classification, it didn't solve the problem.

To incorporate class information into the pruning process, we improved the pruning algorithm using a heuristic scoring system. Each point is assigned a score based on the points surrounding it. Each neighboring point that belongs to the same class contributes +1 to its score, while those that belong to a different class contribute -1. Once this is complete, only points that have a higher score than all their neighbors of a different class are retained; the rest are pruned. This allowed for classes to cluster their features in regions where they are most predominate. However, this was also insufficient, as it tended to allow one class to "take over" (Figure 7).

We also used *K*-means segmentation to find useful clusters of data, so that clusters could be represented by their means. *K*-means segmentation is an algorithm that divides the data into *k* clusters. We used an adaptive version of the algorithm that determines the optimal *k* value using distance thresholds for the minimum and maximum sizes of clusters, creating or destroying clusters as needed. In general, 2 to 7 clusters were found in each class. Although clustering proved to be ineffective for enhancing the keypoint data-set, it was useful as part of the nearest-mean classifier described in Section 4.

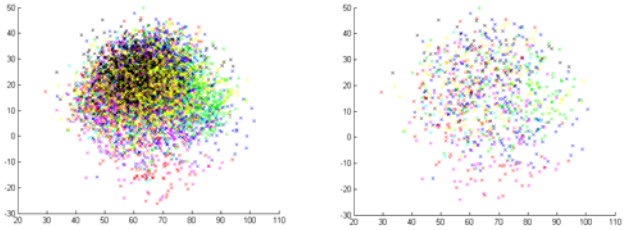

Figure 6: The first 2 dimensions of the data after performing LDA on it. (Left) The original data, repeated from Figure 2, color coded by class. (Right) After distance-based pruning.

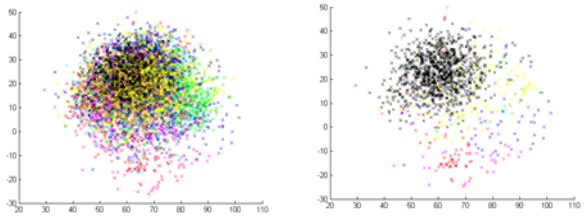

Figure 7: The original data (left) is pruned using the class-based pruning system, yielding the image on the right. Note how the class colored as black has taken over a significant part of the space.