Large-scale Indoor/Outdoor Image Classification via Expert Decision Fusion (EDF)

Chen Chen Yuzhuo Ren C.-C. Jay Kuo

Department of Electrical Engineering
University of Southern California, Los Angeles, CA 90089, U.S.A.
{chen80,yuzhuore}@usc.edu,cckuo@sipi.usc.edu

Abstract. In this work, we propose an Expert Decision Fusion (EDF) system to tackle the large-scale indoor/outdoor image classification problem using two key ideas, namely, data grouping and decision stacking. By data grouping, we partition the entire data space into multiple disjoint sub-spaces so that a more accurate prediction model can be trained in each sub-space. After data grouping, the EDF system integrates soft decisions from multiple classifiers (called experts here) through stacking so that multiple experts can compensate each other's weakness. The EDF system offers more accurate and robust classification performance since it can handle data diversity effectively while benefiting from data abundance in large-scale datasets. The advantages of data grouping and decision stacking are explained and demonstrated in detail. We conduct experiments on the SUN dataset and show that the EDF system outperforms all existing methods by a significant margin with a correct classification rate of 91%.

1 Introduction

Indoor/outdoor scene classification is one of the basic scene classification problems in computer vision. Its solutions contribute to general scene classification [1–6], image tagging [7–9], and many other applications [10–13]. As compared to general scene classification problems, the indoor/outdoor scene classification problem has a clearer definition, namely, whether the scene is inside or outside a man-made structure with enclosed roofs and walls. Since the man-made structure is well-defined, the decision is unambiguous under various circumstances.

Indoor/outdoor classification allows a precise characterization of a wide range of images with diversified semantic meanings. For example, images from ① kitchen to ⑨ green house in the left column of Fig. 1 should all be classified as indoor images. In contrast with other scene classification problems [14–16], semantic objects in the scene may not help much in the decision. For example, indoor ⑤ swimming pool and outdoor ⑤ swimming pool in Fig. 1 share the same salient semantic object (*i.e.*, the pool), yet they should be classified differently from the aspect of indoor/outdoor scene classification. The same observation occurs in quite a few real-world images.

Fig. 1. Exemplary indoor and outdoor scene images from the test dataset are given in the left and right of the dash line, respectively.

Millions of images have been created every day due to the popularity of smart phones. Due to the huge size and great diversity of image data, applications such as large-scale image search [17] and tagging [18] will benefit from accurate indoor/outdoor classification results. Several methods, including SP [19], VFJ [20], SSL [21], PY [22], KPK [23] and XHE [24], were proposed to tackle this problem based on image datasets consisting of about 1,000 images. It is not clear whether the reported performance of these methods is scalable to large-scale datasets consisting of more than 100,000 images. This is the main focus of our current research.

To address the large-scale indoor/outdoor scene classification problem, we propose an Expert Decision Fusion (EDF) system that consists of two key ideas – data grouping and decision stacking. In contrast with prior art, the proposed EDF system is less concerned with the search of new features but on a meaningful way to partition the dataset and organize basic indoor/outdoor classifiers in an effective way to lead to a more accurate and robust classification system. For convenience, each basic indoor/outdoor classifier is called an "expert" in this paper.

The design of the EDF system can be described as follows. We select a set of experts as the constituent members of the EDF system. After evaluating a few existing indoor/outdoor image classifiers [19–33], we choose 6 experts. They are SP [19], VFJ [20], SSL [21], PY [22], KPK [23] and XHE [24]. Furthermore, we developed three new experts (namely, HSH, TN and HDH) on our own. To han-

dle the problem of data diversity, we propose an effective way to partition data samples into multiple groups, where data in one group are more homogeneous to model and predict. Furthermore, the EDF system integrates soft decisions of constituent experts via stacking [34–36] to offer a better classification performance than each individual expert in each partitioned sub-space. To illustrate the advantage of the EDF system, we label all images in the SUN [24] dataset (consisting of 108,754 images in total) with the indoor/outdoor ground truth, and compare the performance of a set of methods.

There are several contributions of this work. First, to the best of our knowledge, this is the first study on the large-scale indoor/outdoor scene classification problem with a dataset exceeding 100,000 images. The developed methodology and learned experience contribute to the fundamentals of "big data" science and engineering. Second, three new indoor/outdoor image classifiers (or experts) are proposed as constituent members in the EDF system. Third, we demonstrate the power of data grouping and decision stacking in the design of the EDF system. Finally, the proposed EDF system reaches a correct classification rate of 91% against the SUN dataset, which offers 6-26% performance improvement over other benchmarking methods. Besides, we show that it provides a scalable solution by examining its performance as a function of different sizes of the SUN dataset.

One important lesson learned from our current study is that, as the data size becomes larger, there are two competing factors that have a high impact on the performance of a classification system – data diversity and data abundance. The former demands a better classifier design. In this work, we propose the use of data grouping and decision stacking to achieve this goal. Once the data diversity problem is addressed, data abundance actually helps improve the performance of a robust classifier.

The rest of this paper is organized as follows. We describe constituent experts in the EDF system in Sec. 2, which include existing indoor/outdoor scene classifiers as well as three newly developed classifiers. Then, the design of the EDF system is detailed in Sec. 3. Experimental results are reported and discussion is given in Sec. 4. Finally, concluding remarks and possible future extensions are presented in Sec. 5.

2 Description of Constituent Experts

2.1 Six Experts from Existing Work

Many indoor/outdoor scene classification solutions have been proposed in the past 15 years. The main focus has been on the selection of discriminant features. Low-level features such as color, texture and shape have been examined. For example, color histograms [19, 21, 23] and color moments [20] are two popular features. The global color pattern of tiny images [24] offers another color descriptor. Besides the RGB color space, other color spaces such as Ohta [37], LST and HSV, were studied [19, 21, 23]. Texture features were applied to indoor/outdoor

scene classification [19, 21]. The MSAR [38] and the multi-scale wavelet [39] are used as local texture descriptors.

Features such as edge angle histograms [23] and responses of Gabor filters (GIST, [22, 29]) were used in recent works. KPK [23] partitions an image into one horizontal block in the top portion and four vertical blocks in the middle and lower portions and assigns different weights to features in these five blocks for further processing. Rather than partitioning an image into blocks, PY [22] computes the GIST [40] features from the original image and its edge map separately and cascades the two responses into a feature vector.

The performance of these classifiers approaches to their limits quickly as the image dataset becomes larger. We choose six of them as constituent experts of the EDF system, denoted by SP [19], VFJ [20], SSL [21], PY [22], KPK [23] and XHE [24]. We implement all of them by ourselves in the experimental section since none of the source codes is available.

Feature extraction and classifier training are two basic steps in developing an expert. Machine learning has been widely used in classifier training. The K-Nearest Neighbor (KNN) algorithm and the Learning Vector Quantization (LVQ) [41]) were considered in [19, 29] and [20], respectively, where the choice of a good distance measure was the main issue. Later, the Support Vector Machine (SVM) [42] was used in [21–23] and the Probability Neural Network (PNN) [43] became popular due to their good performance and the availability of open source codes.

2.2 Three New Experts

We propose three new experts based on the features of Thermal Noise (TN), the Hue-Saturation Histogram (HSH) and the Hue-Dark Histogram (HDH). Their justification and implementation are detailed below.

The TN Expert. Thermal noise [44] arises in the image acquisition process due to poor illumination, high temperature, etc. Typically, indoor scenes have weak lighting sources and lower temperatures while outdoor scenes have stronger natural light and higher temperatures. For this reason, we propose to use TN to differentiate indoor/outdoor scenes. In the feature extraction step, noise levels in different color channels are calculated as a descriptor. First, we adopt a bilateral filter approach [45] to denoise each channel of the RGB, HSV and YUV color representations of an input image. Then, absolute differences between the original and the denoised image channels are computed to yield 9 noise maps for a single color image. Finally, standard deviations of all noise maps are concatenated to form a feature vector. In the model training step, we adopt linear SVM in the package [42] and the 5-fold cross validation process for performance evaluation.

The HSH Expert. The HSV color space is strongly linked to human visual perception. Here, we use a "modified" Hue-Saturation Histogram (HSH) to characterize the global color distribution of an image based on the following observation. Image pixels with low value (V) and low saturation (S) components do not contribute to the discrimination of indoor/outdoor scenes since too dark or

bright pixels are not reliable in the decision. Hence, the hue-saturation histogram is only calculated in a partial volume of the HSV color space by excluding dark and bright pixels in our implementation. That is, for a pixel with its HSV color coordinates (h, s, v) we will include this pixel in the histogram calculation only if $v \geq T_v$, $s \geq T_s$. Otherwise, it is abandoned. T_v and T_s are empirically set to 0.2 and 0.1, respectively. We quantize the hue values into 16 bins and adopt a 5-bin saturation histogram for each hue bin. Consequently, we obtain an 80-bin hue-saturation histogram of an image. This HSH descriptor is used to train a linear SVM classifier to yield the HSH expert.

The HDH Expert. The dark channel was introduced in [46]. For a given pixel, its dark channel value, denoted by D, is the lowest one among its R, G, B three channel values. We have an interesting observation, namely, the dark channel patterns are different for indoor and outdoor scenes. Bright red objects in indoor scenes, such as the carpet in ② parlor and sheet in ③ bedroom in Fig. 1, usually have small dark channel values since they have small values in green and blue channels. In contrast, due to lighting conditions, red objects in outdoor scenes, such as walls in ① urban and the sunset halo in ③ beach in Fig. 1 have larger dark channel values. Furthermore, we observe different relations between certain colors and their dark channel values in indoor and outdoor images. To model this relationship and design a suitable feature, we partition an image into $4 \times 4 = 16$ sub-images and calculate the hue-dark histogram in each sub-image. For the hue-dark histogram, we quantize the hue channel values into 16 bins and compute a 5-bin dark value histogram in each hue bin to result in a 80-bin hue-dark histogram (HDH). Then, we concatenate the HDH descriptors of 16 sub-images to yield the final HDH descriptor, which is a 1280-dimensional feature vector. Again, a SVM model is used to train the HDH expert.

2.3 Feature Selection via ANOVA

A correct classification rate of 90% for several experts such as SP, VFJ and SSL was reported before on experiments with around 1,000 images. However, when the data size becomes much larger, we see a significant performance drop between the training stage and the testing stage for experts with a high dimensional feature vector. This phenomenon is attributed to the over-fitting of high-dimensional feature vector in the training stage. We list the feature vector dimension of each expert in Table 1. To avoid the overfitting problem in classifier training and reduce the training-testing performance gap, one solution

Table 1. The feature dimension numbers of experts before (labeled as "original") and after (labeled as "selected") the ANOVA feature selection process.

Experts	SP	VFJ	SSL	PY	KPK	XHE	TN	HSH	HDH
Original	1536	600	880	1024	80	768	9	80	1280
Selected	360	40	280	300	80	500	9	80	240

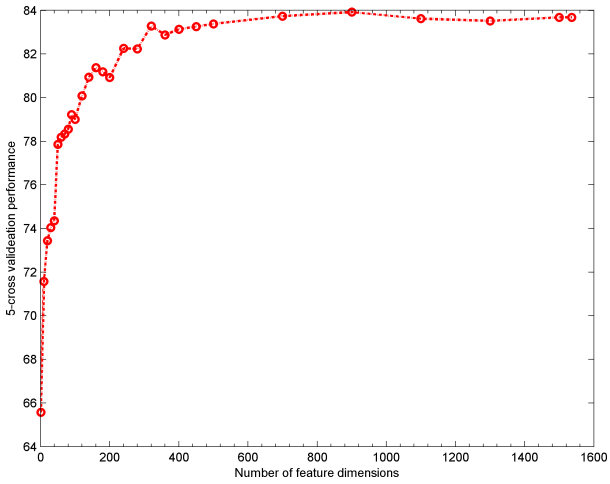

Fig. 2. The correct classification rate curve of SP with top D dimensions of the SP feature vector used as the training features and being evaluated by 5-fold cross validation.

is to select a smaller set of discriminant features. In the following, we use the well-known analysis of variance (ANOVA [47]) method for feature selection.

One of the most widely used tools in ANOVA is the F-test [48]. For a single feature dimension, its F-value is defined as the ratio of the between-group variance and the within-group variance. Let $\bar{Y}$ be the mean of all data in this feature dimension, and K is the number of groups. We have indoor/outdoor two groups so that $K = 2$. We use $\bar{Y}_i$ and n_i to denote the sample mean and the observation number over this dimension in the i^{th} group. Then, the between-group variance of a single feature dimension can be written as

$$Var_{bg} = \sum_{i=1}^{K} n_i(\bar{Y}_i - \bar{Y})^2/(K-1). \tag{1}$$

Furthermore, its within-group variance can be expressed as

$$Var_{wg} = \sum_{ij} n_i(Y_{ij} - \bar{Y}_i)^2/(N-K), \tag{2}$$

where Y_{ij} is the j^{th} observation in the i^{th} out of K groups and N is the overall sample size. Finally, the F-value can be written mathematically as

$$F = \frac{Var_{bg}}{Var_{wg}}. \tag{3}$$

In classification, the larger F value is, the more discriminative this feature dimension is. After computing the F values of all feature dimensions of an expert's feature vector, we rank them from top to bottom and select the top D as

desired features for classification. Fig. 2 shows the performance curve of the SP expert as a function of the D value. We select the 360 dimensions with larger F values for SP based on this figure since the performance becomes saturated after the use of 360 feature dimensions. In our implementation, different experts have different numbers of selected feature dimensions. This result is listed in Table 1.

3 Design and Analysis of EDF System

Instead of adopting a single classifier, the idea of using a system of classifiers to improve the overall classification performance has been investigated before. For example, SP [19], SSL [21] and the work in [29] all adopt a two-stage classification system. At the first stage, they partition images into $4 \times 4 = 16$ sub-images and determine indoor or outdoor labels for each sub-image individually. Then, the decisions of these 16 sub-images are integrated by either voting or training a second-level classifier to make the final decision for the whole image. There is a major difference between EDF and the above idea. That is, the EDF system does not integrate decisions from sub-images but decisions from multiple experts made for the whole image. In the following, we first conduct the analysis on a single expert's decision in Sec. 3.1. Then, we explain the "diversity gain" of any two experts in Sec. 3.2. Finally, we discuss the structure of the EDF system in Sec. 3.3.

3.1 Analysis of Single Expert Decision

Before considering the collaboration of experts, we first analyze the decision behavior of a single expert. Without loss of generality, we use expert KPK as an illustrative example. For the j^{th} image sample, denoted by I_j, KPK can generate a soft decision score, d_j^{kpk}, for it using its sample-to-boundary distance normalized to the range $[0, 1]$, where 0 and 1 indicate the indoor and outdoor scenes with complete confidence, respectively. When there is only one expert, we need to quantize the soft decision score into a binary decision. That is, we divide interval $[0, 1]$ into two subintervals $S_1 = [0, T)$ and $S_2 = [T, 1]$, where $0 < T < 1$ is a proper threshold value (typically, $T = 0.5$). If $d_j^{kpk} \in S_1$, I_j is classified to an indoor image. Otherwise, $d_j^{kpk} \in S_2$ and I_j is classified to an outdoor image.

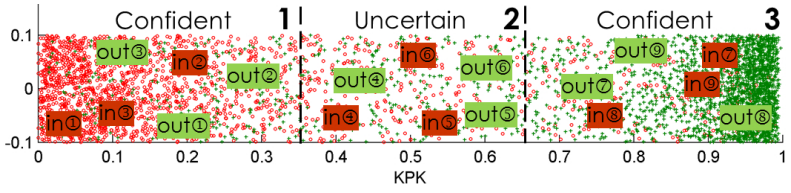

Fig. 3. The distribution of soft KPK decision scores d^{kpk} from 5,000 random samples.

When soft decision score d_j^{kpk} is closer to threshold T, expert KPK is less confident about its decision. To take this into account, we may partition the entire decision interval into 3 subintervals $S_1 = [0, T_1)$, $S_2 = [T_1, T_2)$, and $S_3 = [T_2, 1]$, where $0 < T_1 < T_2 < 1$ are two thresholds. Parameters T_1 and T_2 are set to 0.35 and 0.65 in our implementation. Subintervals S_1 and S_3 are called the confident regions while subinterval S_2 is called the uncertain region.

We show the distribution of soft KPK decision scores collected from 5000 sample images randomly selected from the SUN database in Fig. 3, where red circles and green crosses denote indoor and outdoor image samples, respectively. To avoid the overlap of cluttered samples along the x-axis, we generate a vertical random shift between -0.1 to 0.1 for each sample, which is purely for the visualization purpose and has no practical meaning. We see that most red circles are in S_1 while most green crosses are in S_3. They can be correctly classified by KPK. On the other hand, there are few red circles in S_3 and few green crosses in S_1, and they will be misclassified by KPK. There are some red circles and green crosses in S_2, which are difficult to set apart using the soft KPK decision scores. It is apparent that the criteria of a good expert can be stated as:

1. it has a larger ratio of correct versus incorrect decision samples in S_1 and S_3; and
2. it has a smaller percentage of samples in S_2.

We will discuss ways to achieve the above goal by inviting the second expert to join the decision-making process in Sec. 3.2.

To gain more sights, we show the KPK soft scores of all 18 images in Fig. 1 in Fig. 3. Note that we select the images in Fig. 1 carefully so that there are three representative indoor and outdoor images in each sub-interval in Fig. 3. Visual inspection of these sample images will help us understand the strength and weakness of KPK.

Images in the Uncertain Region. For images in S_2, KPK cannot make a firm decision. Indoor images ④-⑥ and outdoor images ④-⑥ lie in this region. The two swimming pool images, images ⑤ in both indoor and outdoor categories, have similar elements such as blue water and dark tops. In addition, indoor and outdoor images ⑥ also share similar color patterns.

For images in S_1 and S_3, KPK has confident soft scores. They can be further divided into two cases.

Correctly Classified Images. Indoor images ①-③ and outdoor images ⑦-⑨ are correctly classified. Recall that KPK partitions an image into 5 blocks (namely, one horizon block in the top and four parallel vertical blocks in the lower portion). Indoor images ①-③ have small d^{kpk} values since they all have red and wooden objects at the bottom part of the images and shell-white ceilings or walls, which are easy to classify with KPK's block-based color and edge descriptors. Similarly, the top horizontal block carries the valuable sky information for outdoor images ⑦-⑨.

Misclassified Images. Outdoor images ①-③ and indoor images ⑦-⑨ are misclassified although their scores fall in the confident regions. They are called outliers. Outdoor images ①-③ all have dark colors and clear edge structures

over the entire image, which misleads KPK. The blue top part of indoor image ⑦ is also misleading. Indoor image ⑧ is difficult since its wall contains the outdoor view and painting. Indoor image ⑨ can be even challenging to human being since one may make a different decision depending on the existence of the ceiling and the wall.

For the outlying images, low-level features mislead KPK to draw a confident yet wrong conclusion. Human can make a correct decision by understanding the semantic meaning of the scenes such as the river, the street and the ocean in outdoor images ①-③, respectively. Furthermore, indoor images ③ and ⑦ have the same semantic theme (bedroom) but different low-level features (color and texture patterns).

It is well known that there exists a gap between low-level features and high-level semantics of an image, which explains the fundamental limits of experts that rely purely on features in decision-making. Despite the semantic gap, a well-designed feature-based classifier can offer a reasonable classification performance due to the strong correlation between good low-level features and high-level semantics in a great majority of images.

3.2 Diversity Gain of Two Experts

As the size of image data becomes larger and their contents become more diversified, it is challenging to design a single expert that can handle all image types effectively. It is a natural idea to get the opinions of multiple experts and combine their opinions to form one final decision. In this subsection, we consider the simplest two-expert case. Intuitively, such a system may not work well under the

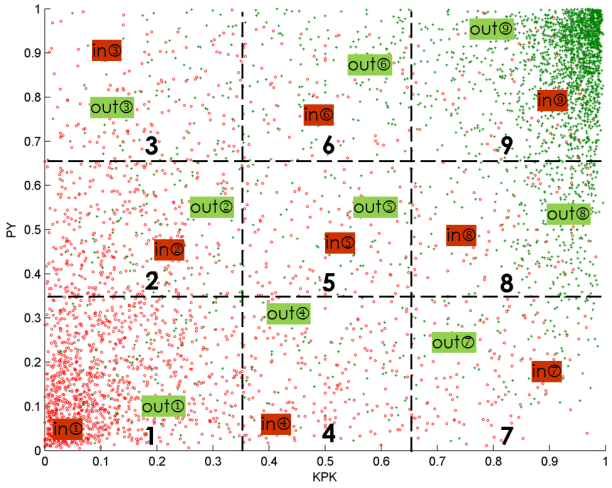

Fig. 4. The distribution of KPK-PY soft decision scores for the same 5,000 randomly picked samples shown in Fig. 3.

following two scenarios: 1) if the opinions of two experts are too similar to each other; or 2) if one expert is significantly better than the other. In both scenarios, we do not benefit much by inviting the second expert in the decision process. Scenario (2) is self-evident. We will focus on scenario (1) by investigating the diversity gain of the two-expert system.

Without loss of generality, we choose KPK and PY as the two experts. The soft decision scores of expert PY, denoted by d^{py}, for the same 5,000 samples are plotted along the vertical axis in Fig. 4. The j^{th} sample image represented by a red circle (indoor) or a green cross (outdoor) has a 2-D coordinate, (d_j^{kpk}, d_j^{py}), whose coordinate domain is called the KPK-PY soft decision map. With different combinations of soft decisions from the two experts, we can divide the 2-D decision space into 9 regions. KPK and PY have consistent opinions in their soft decisions in regions 1, 5 and 9, complementary decisions in regions 2, 4, 6 and 8, and contradictory opinions in regions 3 and 7.

By comparing Figs. 3 and 4, we see that PY can help KPK in resolving some decision ambiguities in regions 4-6. That is, PY can offer more confident scores for images in regions 4 and 5 than KPK. Similarly, KPK can help PY in resolving some decision ambiguity in regions 2, 5 and 8 since KPK is confident for images in regions 2 and 8. KPK and PY offer complementary strength since they examine different low-level features in evaluating an input image. KPK focuses on the local color and edge distributions while PY focuses on global scene structures. Indoor/outdoor images ②, ④, ⑥ and ⑧ are exemplary images in regions 2, 4, 6 and 8, respectively.

We focus on indoor/outdoor images ④ and ⑥, for which KPK does not have a confident score. Recall that PY [22] does not partition an image into multiple sub-images but computes the GIST features from the original image and its edge map separately and cascades the two responses into a feature vector. As a result, PY can make a more confident decision. PY's decisions on indoor image ④ (complicated scene structure for the whole image) and outdoor image ⑥ (textures of grass and leaves) are correct, yet PY's decisions on indoor image ⑥ (similar to outdoor image ⑥) and outdoor image ④ (consisting of many straight vertical lines similar to the view observed inside a church building) are not accurate. Since there are more indoor images than outdoor images in region 4 and more outdoor images than indoor images in region 6, PY does contribute to the correct classification rate in regions 4 and 6.

The same discussion applies to regions 2 and 8, where KPK helps PY in resolving ambiguity in a positive way. Region 5 remains to be ambiguous in the two-expert system. If the two experts share very similar opinions, most samples will fall in regions 1, 5 and 9 so that the two-expert system does not offer a clear advantage. On the other hand, if the two experts have good but diversified opinions, we will observe more samples in the four complementary regions and, as a result, the overall classification performance can be improved.

Finally, PY and KPK have conflicting opinions in regions 3 and 7. To resolve the conflict, we can invite another expert in the decision making process as detailed in the next subsection.

3.3 Structure of EDF System

We use Fig. 5 to explain the design methodology of the EDF system. It consists of the following two stages.

First, we perform data grouping by considering a two-expert system (say, KPK and PY shown in this figure). Given the KPK-PY soft decision map, we partition the data sample space into 9 regions. Generally speaking, the data grouping technique is a powerful pre-processing step in machine learning. Its main purpose is to enhance the correlation between training and testing samples. A good grouping strategy can contribute to the overall performance of the learning-based system significantly. We have tried different combinations of two experts from nine experts introduced in Sec. 2, and found that KPK and PY provide the best results due to their excellent individual performance and good complementary property. After grouping, the diversity of data samples in each region is reduced.

Second, we fuse the soft decisions of all nine experts in each region. We compare two methods in Sec. 4 – voting and stacking. For voting, we binarize the soft decision of each expert and use the simple majority voting rule to fuse expert's decisions. For stacking, we build a meta-level classification model that takes soft scores of all experts as the input features and make a final binary system decision. Since the training data in each region are different, different meta-level data models are built for different regions. The meta-level classifier is trained by linear SVM using samples with known binary outputs and, then, the trained model is used to predict samples with unknown binary outputs in the test. The correct classification rates of voting-based and stacking-based EDF

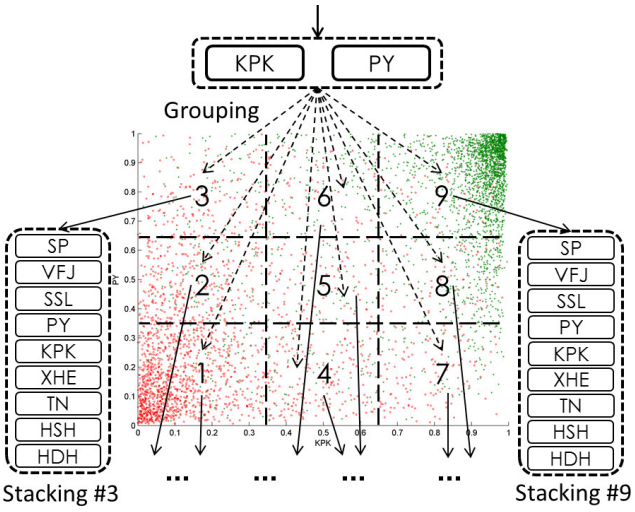

Fig. 5. The structure of the EDF system, where Stacking #3 indicates the stacking method in Region 3 of the joint KPK-PY soft decision map.

systems are compared in Sec. 4. We will show that the stacking approach provides a better result.

Both grouping and stacking provide powerful tools to handle the problem of data diversity. Through grouping, we have more and smaller homogeneous datasets rather than one large highly heterogenous dataset. Through stacking, we can improve the robustness of the final decision in each region by leveraging the complementary strength of multiple experts.

4 Experimental Results

In the indoor/outdoor scene image classification literature, datasets used by other research groups are either too small or not available to the public. For example, the Kodak consumer image dataset, tested by SP [19] and SSL [21], contains 1343 images. Coral, used by VFJ [20], is not available to the public. A benchmark of 1000 images used in [29] is also not available. KPK [23] collects around 1200 images from the Internet, yet they are not released to the public. Two datasets consisting of 390 and 968 images, respectively, and used in [43] are accessible from their websites. Recently, a very large dataset, SUN, was published by [24] for the general scene classification benchmark. It consists of 397 well-sampled scene category indexes and 108,754 images. We labeled the whole SUN dataset into 47,260 indoor images and 61,494 outdoor images. Our experiments are conducted with respect to this dataset.

First, we show the correct classification rates of nine experts against the full SUN dataset in Table 2 without data grouping. The 5-fold cross validation is adopted in the experiment and the averaged performance is listed. There are one, three and five experts with correct classification rates between 60-69%, 70-79% and 80-89%, respectively. Expert KPK has the best performance with a correct classification rate of 85.30%.

Next, we show the correct classification rate achieved by nine experts and the EDF system in Regions 1-9 in Table 3, where the 5-fold cross validation is conducted in each region and the averaged performance is reported. For the last row, we perform the weighted sum of correct rates in nine regions based on their sample population to derive the results with respect to "all" data samples. For EDF_V, we binarize the soft decision of each expert and use the voting scheme to fuse expert's decisions. The majority rule is used to select the final system decision. For EDF_S, we adopt the stacking scheme to fuse experts' decisions. That is, we build a meta-level on top of all soft decisions which learns the fusion rule with an SVM classifier that treats experts' soft decisions as features. We

Table 2. Correct classification rate of nine experts conducted on the full dataset (in the unit of %).

SP	VFJ	SSL	PY	KPK	XHE	TN	HSH	HDH
82.80	79.01	84.69	82.28	85.30	78.68	64.52	77.84	82.23

Table 3. Classification performance comparison of nine experts and EDF in Regions 1-9 on the full dataset (in the unit of %).

	SP	VFJ	SSL	PY	KPK	XHE	TN	HSH	HDH	EDF_V	EDF_S	EDF
1	92.22	92.22	92.22	92.22	92.22	92.22	92.22	92.22	92.22	92.22	92.43	92.43
2	76.91	75.99	78.24	80.07	75.96	75.96	75.92	75.94	77.34	75.96	86.32	86.32
3	71.63	61.18	73.18	70.07	66.17	73.18	61.47	66.75	70.49	77.48	81.36	81.36
4	83.46	83.22	83.67	83.22	83.22	83.22	83.22	83.32	83.41	83.22	85.58	85.58
5	68.96	56.87	70.50	65.21	60.23	70.16	61.57	66.14	67.79	68.23	79.84	79.84
6	81.17	81.07	81.72	80.90	81.09	80.99	80.87	81.06	81.70	80.90	86.32	86.32
7	70.04	63.06	71.72	68.68	66.14	71.06	63.06	64.62	68.80	65.60	80.12	80.12
8	72.51	70.56	72.58	75.35	70.37	72.44	70.35	70.58	72.05	70.51	81.54	81.54
9	96.70	96.74	96.74	96.75	96.74	96.74	96.74	96.74	96.74	96.74	96.74	96.74
All	88.64	87.44	88.96	88.64	87.87	88.73	87.56	88.00	88.56	88.52	91.15	91.15

see that the performance of EDF_S is no worse than EDF_V in all regions. Thus, it is chosen to be the final EDF solution.

By comparing results in Table 2 and Table 3, we see clearly that the performance of each expert has improved a lot (ranging from 4-11%) due to data grouping. After data grouping, the performance gap among different experts narrows down significantly. Their correct classification rates now are in the range of 87.44-88.96%. The EDF system can achieve a correct classification rate of 91.15% by stacking all experts in each region. With the combination of grouping and stacking, the EDF system can outperform traditional experts (without grouping) by a margin of 6-26%.

Finally, we plot the performance of each individual expert (without data grouping) and the EDF system as a function of the size of the dataset in Fig. 6. We select subsets of increasing sizes randomly from the SUN dataset and list the size in the x-axis while the averaged correct classification rate using the 5-fold cross validation is shown in the y-axis. The vertical segment on each marker indicates the standard deviation of a particular test. We see that the performance of some individual experts stay flat while others drop as the data size becomes large. In contrast, the performance of the EDF system improves as the data size becomes larger.

When the data size becomes larger, there are two competing factors that influence the performance of a classification system in two opposite directions. On one hand, the data become more diversified and the performance of an expert may go down if its model cannot handle a diversified data set. On the other hand, the number of similar data (*i.e.*, belonging to the same data type) becomes more. Data abundance helps improve the performance of learning-based classifiers. Fig. 6 implies that the data diversity problem is under control by the robust EDF system so that the EDF system benefits more from data abundance. We expect the EDF system performance to be level-off at a certain data size although we have not yet observed such a phenomenon in Fig. 6. This is because that the EDF system does not address the semantic gap issue.

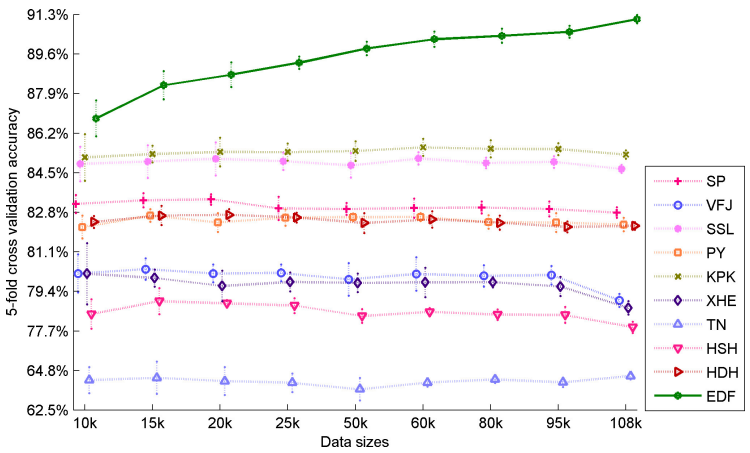

Fig. 6. Comparison of classification performance of nine experts and EDF as a function of the size of dataset.

5 Conclusion and Future Work

An Expert Decision Fusion (EDF) system was developed to address the large-scale indoor/outdoor image classification problem in this work. As compared with the traditional classifiers (or experts), the EDF system consists of two key ideas: 1) grouping of data samples based on the soft decisions of two experts into 9 regions; and 2) stacking of soft decisions from all constituent experts to enhance the classification performance in each region. It was shown by experimental results that the proposed EDF system outperforms all traditional classifiers in the classification accuracy by a margin of 6-26% on the large-scale SUN image dataset. The classification performance of EDF improves as the size of the dataset grows, which can be explained by its capability of handling data diversity. With this capability in place, as the dataset grows to a very large size, data abundance becomes a more dominant factor than data diversity. Thus, the EDF system offers a robust and scalable solution.

As discussed in Sec. 3, there is some fundamental limits in the feature-based classifiers since they do not take the image semantics into account. We expect to see a point where the performance of EDF becomes saturated, which will be the true upper performance bound of EDF. To achieve this goal, we need to look for some dataset even larger than SUN. To improve the performance of EDF furthermore beyond the saturation point, we need to look for semantic-based experts. This is clearly a very challenging problem since it involves object and scene recognition. Finally, a good indoor/outdoor image classifier is an important pre-processing step to scene analysis. It is desirable to leverage our current results to obtain better methods for scene classification and recognition.

References

1. Bosch, A., Zisserman, A., Munoz, X.: Scene classification via plsa. In: Computer Vision–ECCV 2006. Springer (2006) 517–530
2. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Volume 2., IEEE (2005) 524–531
3. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: Computer vision and pattern recognition (CVPR), 2009 IEEE conference. (2009)
4. Li, L.J., Su, H., Xing, E.P., Li, F.F.: Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: NIPS. Volume 2. (2010) 5
5. Wu, J., Rehg, J.M.: Centrist: A visual descriptor for scene categorization. Pattern Analysis and Machine Intelligence, IEEE Transactions on **33** (2011) 1489–1501
6. Vailaya, A., Figueiredo, M., Jain, A., Zhang, H.: Content-based hierarchical classification of vacation images. In: Multimedia Computing and Systems, 1999. IEEE International Conference on. Volume 1., IEEE (1999) 518–523
7. Lim, J.H., Jin, J.S.: A structured learning framework for content-based image indexing and visual query. Multimedia Systems **10** (2005) 317–331
8. Chatzichristofis, S.A., Boutalis, Y.S.: Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In: Computer Vision Systems. Springer (2008) 312–322
9. Vailaya, A., Jain, A., Zhang, H.J.: On image classification: City images vs. landscapes. Pattern Recognition **31** (1998) 1921–1935
10. Zhang, L., Li, M., Zhang, H.J.: Boosting image orientation detection with indoor vs. outdoor classification. In: Applications of Computer Vision, 2002.(WACV 2002). Proceedings. Sixth IEEE Workshop on, IEEE (2002) 95–99
11. Battiato, S., Curti, S., La Cascia, M., Tortora, M., Scordato, E.: Depth map generation by image classification. In: Proceedings of SPIE. Volume 5302. (2004) 95–104
12. Bianco, S., Ciocca, G., Cusano, C., Schettini, R.: Improving color constancy using indoor–outdoor image classification. Image Processing, IEEE Transactions on **17** (2008) 2381–2392
13. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern recognition **37** (2004) 1757–1771
14. Vailaya, A., Figueiredo, M., Jain, A., Zhang, H.J.: Content-based hierarchical classification of vacation images. In: Multimedia Computing and Systems, IEEE International Conference on. Volume 1. (1999) 518–523
15. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference. Volume 2. (2006) 2169–2178
16. Vogel, J., Schiele, B.: Semantic modeling of natural scenes for content-based image retrieval. International Journal of Computer Vision **72** (2007) 133–157
17. Jégou, H., Douze, M., Schmid, C.: Improving bag-of-features for large scale image search. International Journal of Computer Vision **87** (2010) 316–336
18. Li, Y., Crandall, D.J., Huttenlocher, D.P.: Landmark classification in large-scale image collections. In: Computer Vision, 2009 IEEE 12th International Conference on, IEEE (2009) 1957–1964

19. Szummer, M., Picard, R.W.: Indoor-outdoor image classification. In: Content-Based Access of Image and Video Database on, IEEE (1998) 42–51
20. Vailaya, A., Figueiredo, M.A., Jain, A.K., Zhang, H.J.: Image classification for content-based indexing. Image Processing, IEEE Transactions on **10** (2001) 117–130
21. Serrano, N., Savakis, A., Luo, A.: A computationally efficient approach to indoor/outdoor scene classification. In: Pattern Recognition, Proceedings. 16th International Conference on. Volume 4. (2002) 146–149
22. Pavlopoulou, C., Yu, S.: Indoor-outdoor classification with human accuracies: Image or edge gist? In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. (2010) 41–47
23. Kim, W., Park, J., Kim, C.: A novel method for efficient indoor–outdoor image classification. Journal of Signal Processing Systems **61** (2010) 251–258
24. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: Computer vision and pattern recognition (CVPR), 2010 IEEE conference, IEEE (2010) 3485–3492
25. Luo, J., Savakis, A.: Indoor vs outdoor classification of consumer photographs using low-level and semantic features. In: Image Processing, 2001. Proceedings. 2001 International Conference on. Volume 2. (2001) 745–748
26. Kane, M.J., Savakis, A.: Bayesian network structure learning and inference in indoor vs. outdoor image classification. In: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. Volume 2., IEEE (2004) 479–482
27. Traherne, M., Singh, S.: An integrated approach to automatic indoor outdoor scene classification in digital images. In: Intelligent Data Engineering and Automated Learning–IDEAL 2004. Springer Berlin Heidelberg (2004) 511–516
28. Deng, D., Zhang, J.: Combining multiple precision-boosted classifiers for indoor-outdoor scene classification. In: Information Technology and Applications, 2005. ICITA 2005. Third International Conference on. Volume 1., IEEE (2005) 720–725
29. Payne, A., Singh, S.: Indoor vs. outdoor scene classification in digital photographs. Pattern Recognition **38** (2005) 1533–1545
30. Payne, A., Singh, S.: A benchmark for indoor/outdoor scene classification. In: Pattern Recognition and Image Analysis. Springer (2005) 711–718
31. Hu, G.H., Bu, J.J., Chen, C.: A novel bayesian framework for indoor-outdoor image classification. In: Machine Learning and Cybernetics, 2003 International Conference on. Volume 5., IEEE (2003) 3028–3032
32. Efimov, S., Nefyodov, A., Rychagov, M.: Block-based image exposure assessment and indoor/outdoor classification. In: Proc. of 17th Conf. on Computer Graphics GraphiCon. (2007)
33. Tao, L., Kim, Y.H., Kim, Y.T.: An efficient neural network based indoor-outdoor scene classification algorithm. In: Consumer Electronics (ICCE), 2010 Digest of Technical Papers International Conference, IEEE (2010) 317–318
34. Wolpert, D.H.: Stacked generalization. Neural networks **5** (1992) 241–259
35. Deng, L., Yu, D., Platt, J.: Scalable stacking and learning for building deep architectures. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, IEEE (2012) 2133–2136
36. Džeroski, S., Ženko, B.: Is combining classifiers with stacking better than selecting the best one? Machine learning **54** (2004) 255–273
37. Ohta, Y.I., Kanade, T., Sakai, T.: Color information for region segmentation. Computer graphics and image processing **13** (1980) 222–241

38. Mao, J., Jain, A.K.: Texture classification and segmentation using multiresolution simultaneous autoregressive models. Pattern recognition **25** (1992) 173–188
39. Daubechies, I., et al.: Ten lectures on wavelets. Volume 61. SIAM (1992)
40. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision **42** (2001) 145–175
41. Kohonen, T., Kangas, J., Laaksonen, J., Torkkola, K.: Lvq_pak: A software package for the correct application of learning vector quantization algorithms. In: Neural Networks, 1992. IJCNN., International Joint Conference on. Volume 1., IEEE (1992) 725–730
42. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) **2** (2011) 27
43. Gupta, L., Pathangay, V., Patra, A., Dyana, A., Das, S.: Indoor versus outdoor scene classification using probabilistic neural network. EURASIP Journal on Applied Signal Processing **2007** (2007) 123–123
44. Johnson, J.B.: Thermal agitation of electricity in conductors. Physical review **32** (1928) 97
45. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Computer Vision, 1998. Sixth International Conference on, IEEE (1998) 839–846
46. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. Pattern Analysis and Machine Intelligence, IEEE Transactions on **33** (2011) 2341–2353
47. Iversen, G.R., Norpoth, H.: Analysis of variance. Sage (1987)
48. Lomax, R.G., Hahs-Vaughn, D.L.: Statistical concepts: a second course. Routledge (2013)