To appear, Proceedings of the International Conference on Computer Vision (ICCV), 2011.

Annotator Rationales for Visual Recognition

Jeff Donahue and Kristen Grauman
Dept. of Computer Science, University of Texas at Austin
{jdd,grauman}@cs.utexas.edu

Abstract

Traditional supervised visual learning simply asks annotators "what" label an image should have. We propose an approach for image classification problems requiring subjective judgment that also asks "why", and uses that information to enrich the learned model. We develop two forms of visual annotator rationales*: in the first, the annotator highlights the spatial region of interest he found most influential to the label selected, and in the second, he comments on the visual attributes that were most important. For either case, we show how to map the response to synthetic contrast examples, and then exploit an existing large-margin learning technique to refine the decision boundary accordingly. Results on multiple scene categorization and human attractiveness tasks show the promise of our approach, which can more accurately learn complex categories with the explanations behind the label choices.*

Figure 1. **Main Premise:** Subjective or complex image classification tasks such as those depicted above may require deeper insight from human annotators than the usual category labels. We propose to ask for *spatial* or *attribute-based rationales* for the labels chosen, and augment a large-margin classifier objective to exploit both the labels and these explanations.

1. Introduction

Image classification is an important challenge in computer vision, and has a variety of applications such as automating content-based retrieval, analyzing medical imagery, or recognizing locations in photos. Much progress over the last decade shows that supervised learning algorithms coupled with effective image descriptors can yield very good scene, object, and attribute predictions, e.g., [3, 13, 14]. The standard training process entails gathering category-labeled image exemplars, essentially asking human annotators to say "what" is present (and possibly "where" in the image it is). In this respect, current approaches give a rather restricted channel of input to the human viewer, who undoubtedly has a much richer understanding than a simple label can convey.

Thus, our goal is to capture deeper cues from annotators. We are particularly interested in complex visual recognition problems that require subjective judgment (e.g., saying whether a face is attractive, rating an athletic performance) or else lack clear-cut semantic boundaries (e.g., describing a scene category, categorizing by approximate age). See Figure 1. Can we really expect to learn such subtle concepts purely by training SVMs with HOG descriptors and category names? We instead propose to allow annotators to give a *rationale* for the label they choose, and then directly use those explanations to strengthen a discriminative classifier. Their insight about "why" should not only enable more accurate models, but potentially also do so with less total human effort, since we could amortize the time spent analyzing the image to determine the label itself.

How can an annotator give an explanation? We propose two modes. In the first, the annotators indicate *which regions of the image* most influenced their label choice by drawing polygons. That is, they highlight what seemed most telling for the classification task at hand: "I can tell it's class X, mainly due to this region here." In the second mode, the annotators indicate *which visual attributes* were deemed most influential, where an attribute refers to some nameable property or part. For example, assuming we have intermediate detectors for attributes like size, color, and specific textures, they can state: "It's too round to be an X", or "She's attractive because she's fit."

In either case, the rationale should help focus the classi-

fier on the low- or mid-level image features that can best be used to discriminate between the desired image categories. To that end, we directly leverage an idea originally developed by Zaidan and colleagues for document classification [28]; it generates synthetic "contrast examples" that lack the features in the rationales, and then adds constraints to a classifier objective that require the contrast examples to be considered "less positive" (or less negative) than the original examples. In this way, the contrast examples can refine the decision boundary in the target label space.

While recent work explores various issues in collecting useful labeled datasets [1, 4, 12, 19, 21, 24], we are the first to propose asking annotators for explanations of their labels to directly improve visual category learning. Without injecting into a classifier knowledge of *why* a given label was chosen, traditional discriminative feature selection techniques risk overfitting to inadequate or biased training examples. In contrast, our strategy stands to benefit more immediately from complex human insight (and thus, potentially, with less total training data).

We demonstrate our approach with both scene and human attractiveness categorization tasks, and report results on the 15 Scenes [10] and Public Figures Face [12] datasets, as well as a new dataset of "Hot or Not" images. We show that both proposed visual rationales can improve absolute recognition accuracy. We also analyze their impact relative to several baselines, including foreground-segmented images and a standard mutual information feature selection approach. Overall, we find that human intuition can be captured in a new way with the proposed technique.

2. Related Work

Much work in visual recognition employs standard "image + label" supervision, as exhibited by benchmark collection efforts [4, 9, 19]. Annotations are generally uniform across examples, and the goal is to obtain object names, and optionally sub-image segmentations. Training discriminative classifiers with such data to distinguish basic categories can be quite effective [3, 14]. In this work, however, we tackle more subtle categorization tasks for which an annotator's rationales are expected to be valuable if not necessary.

Recent work offers ways to improve the efficiency of collecting image annotations. Active learning methods predict which images would be most useful to label next given the current category models, and can reduce total annotation time (e.g., [24],[20]). Image labeling games can entice people to contribute useful data for free [26]. A substantial shift in the scale of image datasets annotated today is in part due to the emergence of online services like Mechanical Turk; one can post jobs to get a large number of labels more quickly [4, 21], and consider ways to manage quality control during collection [8, 21, 27]. Such ideas could potentially be paired with ours to make more efficient use of annotator time. However, in contrast to any previous attempts to improve image annotation effectiveness, our approach requests the rationale behind a target label.

In addition to efficiency issues, researchers are exploring the impact of requesting *deeper*, more complete annotations. This includes gathering fully segmented [19], pose-annotated [1], or "attribute"-labeled images [8, 12, 13, 17]. Attribute labels add cost, but can reveal useful mid-level cues [12], or enable novel tasks like zero-shot learning [13]. Our work can be seen as another way to enrich annotations, and we specifically make use of attributes within one form of rationale. Compared to previous uses of attributes, our idea is that human describable properties offer a new way for the annotator to communicate to the learning algorithm, and better teach it to recognize a complex visual category.

Some work in both language and vision studies how to capture (or predict) those elements most "important" to a human. The information can be explicitly gathered through classic iterative relevance feedback (e.g., [2]). However, more implicit measures are also possible, such as by learning what people mention first in a natural scene [11, 22], or what they deem a foreground object [16]. Whereas such methods use these cues to predict important regions in novel images, our goal is to use what a human deems influential so as to better predict the category label for novel images. Work in natural language processing (NLP) explores whether humans can pick out words relevant for a given document category as a form of human feature selection [6, 18, 28]. In particular, the NLP method of [28] proposes rationales to better predict sentiment in written movie reviews, and inspires our approach; we adapt the authors' basic idea to create two new forms of contrast examples for the visual domain.

We use human attractiveness as one of our testbeds. Previous work shows that supervised learning can predict human judgements about facial beauty with about 0.6 correlation [7], and a graphics technique synthesizes "more attractive" faces [15]. Whereas the latter aims to create new pictures, our method adapts descriptors in a visual feature space and aims to better learn the categories.

3. Approach

The main idea is to gather visual rationales alongside traditional image labels, in order to use human insight more fully to better train a discriminative classifier. We particularly target subjective, perceptual labeling tasks. We first explain the notion of contrast examples as used in prior NLP rationales work [28] (Section 3.1), and then define our two proposed forms of visual rationales (Sections 3.2 and 3.3).

3.1. Rationales as Contrast Examples in an SVM

Zaidan *et al*. [28] propose a technique for document sentiment analysis based on a modified support vector machine

(SVM) objective using "rationales" that is also relevant for our problem. We briefly review it here; see [28] for details.

Suppose we have a labeled set of n instances $\mathcal{L} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, where each $\boldsymbol{x}_i \in \Re^d$ and $y_i \in \{-1, 1\}$. To train a traditional SVM classifier, one seeks a large-margin separating hyperplane for the positive and negative exemplars that satisfies their label constraints, with some slack variables to enable a soft margin. The intuition behind rationales is to further require that a set of corresponding synthetically generated contrast examples $\mathcal{C} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ are treated as *less confidently labeled* examples by that classifier.[1]

More specifically, for some positively-labeled training example $\boldsymbol{x}_i$, one must use its human-provided rationale to modify the example in some way to form $\boldsymbol{v}_i \in \Re^d$, such that $\boldsymbol{v}_i$ resembles $\boldsymbol{x}_i$, but *lacks* the features most critical to defining it as positive. Then, beyond the usual label constraints to satisfy each $(\boldsymbol{x}_i, y_i) \in \mathcal{L}$, the classifier objective also includes constraints requiring that there be a secondary margin μ between any pair $(\boldsymbol{x}_i, \boldsymbol{v}_i) \in \mathcal{C}$. See Figure 2.

Formally, this leads to the following objective for hyperplane $\boldsymbol{w}$:

$$\text{minimize} \quad \left(\frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i \xi_i + C_c \sum_i \gamma_i\right) \tag{1}$$

$$s.t. \quad y_i \boldsymbol{w}^T \boldsymbol{x_i} \geq 1 - \xi_i; \quad \forall i \in \mathcal{L} \tag{2}$$

$$y_i(\boldsymbol{w}^T \boldsymbol{x_i} - \boldsymbol{w}^T \boldsymbol{v}_i) \geq \mu(1 - \gamma_i); \quad \forall i \in \mathcal{C} \tag{3}$$

$$\xi_i \geq 0; \gamma_i \geq 0, \tag{4}$$

where (2) enforces the examples' label constraints, and the important addition (3) lists the *contrast constraints*. They enforce a margin $\mu \geq 0$ separating the positive and "less positive" (or negative and "less negative") example pairs $(\boldsymbol{x}_i, \boldsymbol{v}_i)$ as well. Intuitively, larger values of μ increase the secondary margin, giving greater influence to the rationales. The slack variables ξ_i and γ_j allow soft margins, and the parameters $C > 0$ and $C_c > 0$ denote the associated penalty costs. Setting $C_c < C$ reflects the contrast constraints' secondary importance. The bias terms are omitted above, but accounted for by appending a 1-element to each training example.

The optimization problem can be solved using standard algorithms used to train SVMs, by rewriting (3) as $\forall i \in \mathcal{C}, y_i \boldsymbol{w}^T \hat{\boldsymbol{x}}_i \geq 1 - \gamma_i$, where $\hat{\boldsymbol{x}}_i = \frac{\boldsymbol{x}_i - \boldsymbol{v}_i}{\mu}$. Though not done in [28], it is also straightforward to kernelize this approach.

To use this classification approach, then, the key is to define (1) how an annotator should specify a visual rationale, and (2) how to map their input into a contrast example for the original example. We introduce two variants for these definitions in the following.

[1] In reality, not every training example need have a contrast example (i.e., we can have $|\mathcal{C}| < |\mathcal{L}|$), but we index them as such for notation simplicity.

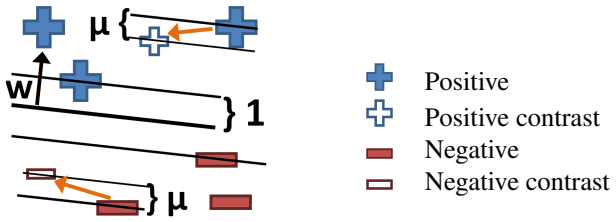


Figure 2. **SVM with Contrast Examples** In addition to maximizing the margin between true positive and negatives, the rationales SVM objective [28] also requires a secondary margin μ between the original examples and their synthetically generated contrast examples. This can usefully alter the ultimate hyperplane $\boldsymbol{w}$.

3.2. Spatial Image Region Rationales

First we consider how contrast examples may be generated based on those *spatial regions* in an image an annotator found most important to his/her label choice. Using an annotation interface with a polygon-drawing tool, we can ask the annotator to mark regions of interest in the image. Then, we will use those regions and their complement to generate a contrast example. For example, in the scene categorization task we explore in this work, an annotator might mark certain telling objects that indicate a scene type (a sink in a 'kitchen' image, or a cash register in a 'store'). Similarly, for the human attractiveness task, s/he might mark facial features or body parts that were found most appealing.

For this idea to work, we require an image representation that consists of spatially localized components. A segment-based or local feature-based representation is appropriate, since we can directly map features located within the region selected into a contrast example.

In particular, suppose for each training image $\boldsymbol{x}_i$ we have an associated *set* of m_i local descriptors $\boldsymbol{x}_i = \{(\boldsymbol{f}_1^i, x_1^i, y_1^i), \ldots, (\boldsymbol{f}_{m_i}^i, x_{m_i}^i, y_{m_i}^i)\}$, where each $\boldsymbol{f}_j^i$ denotes an appearance feature of some kind extracted at position (x_j^i, y_j^i) in image i. In our implementation, we use feature sets composed of local SIFT features, detailed below. Now let $\boldsymbol{r}_i = \{(\boldsymbol{f}_{r_1}^i, x_{r_1}^i, y_{r_1}^i), \ldots, (\boldsymbol{f}_{r_k}^i, x_{r_k}^i, y_{r_k}^i)\}$ denote the subset of those points falling within the annotator-drawn region of interest—the rationale. To generate a contrast example, we simply take the complement of the rationale in the image:

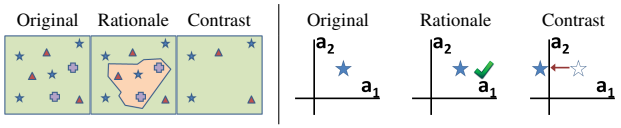
$$\boldsymbol{v}_i = \boldsymbol{x}_i \setminus \boldsymbol{r}_i. \tag{5}$$

See Figure 3(a). Quite intuitively, the spatial rationale makes a contrasting image example that *lacks* the visual features the annotator found most important to deciding on the class. However, rather than simply mask out its image pixels directly (which would likely have unintended consequences by altering the feature space artificially) we manipulate the rationale in the intermediate local descriptor space.

We generate one contrast example per polygon set on an image. More generally, one can introduce multiple ratio-

Attribute	Rationale?
$\mathbf{a}_1$: **pointed toes**	✓
$\mathbf{a}_2$: **on ground**	
$\mathbf{a}_3$: **balanced**	
$\mathbf{a}_4$: **falling**	
$\mathbf{a}_5$: **knee angled**	✓

Annotation task: *Is the skater's form good?*



(a) (b)

Figure 3. **Rationale Modes** We explore two modes of visual rationales. (a): The annotator draws polygons to indicate the region(s) most influential in the class choice. In this mode, a contrast example is formed by masking out the features falling inside a rationale polygon (bottom left). (b): The annotator comments on attributes most influential in the class choice. In this mode, a contrast example is formed by weakening the predicted presence of an attribute, which is a human-nameable visual property (bottom right). Note that other attributes that are present need not play a role in the rationale. In either case, constraints in the classifier objective enforce that the contrast examples be considered "less confident" exemplars for the class chosen.

nales per polygon annotation, sampling subsets of the encompassed points in order to isolate their impact.

We stress that for subjective image classification problems, such spatial image rationales will differ from traditional foreground segmentation. The annotator is not telling us where the category is, but rather, what aspects most push it into the selected category ('kitchen', 'store', 'attractive', etc.). We confirm this important difference in our results.

3.3. Nameable Visual Attribute Rationales

As a second form of rationale, we allow the annotator to comment on the *nameable visual attributes* that most influenced his/her label choice. Suppose we have a set of interpretable visual properties—such as "small", "spotted", "dark", "bushy eyebrows", "smiling"—and that we can train predictive models for them using external image data (i.e., as in the attribute vocabularies used in previous work to describe animals, faces, or scenes [8, 12, 13]). We can use this vocabulary to allow an annotator to communicate about the image aspects that appear most important in determining the category label. For example, in the human attractiveness task, s/he might comment that a person is attractive in a certain photo because of her "happy" demeanor; or that they find a beach scene aesthetically pleasing because it looks "calm".

Let $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_V\}$ denote a vocabulary of V binary visual attributes. For each visual attribute, we train a binary classifier to predict its presence in novel images, using a separate set of training data from that involved in the target image categorization task. Then, when working with attribute rationales, we represent each training image $\boldsymbol{x}_i$ as a V-dimensional descriptor, $\boldsymbol{x}_i = [a_1^i, \ldots, a_V^i]$, where each a_j^i denotes the raw classifier output value for attribute j on image i. We use a linear SVM per attribute.

Given a rationale $\boldsymbol{r}_i = \{a_{r_1}^i, \ldots, a_{r_k}^i\}$ for $\boldsymbol{x}_i$ stating that some k attributes present in the image were most influential to the label choice, we generate a contrast example that reduces those k attribute values by a factor $\delta > 0$ of the standard deviation σ_{a_j} over all classifier outputs for a_j:

$$\boldsymbol{v}_i = [a_1^i, \ldots, a_{r_1}^i - \delta\sigma_{a_{r_1}}, \ldots, a_{r_k}^i - \delta\sigma_{a_{r_k}}, \ldots, a_V^i], \quad (6)$$

See Figure 3(b). This rationale reflects that had the image lacked the propert(ies) named as most relevant, it would have been less likely to be categorized as it was.

In our results, we consider two forms of attribute-based rationales, which differ in how the supervision is provided. In the first "homogeneous" form, we allow a single annotator to specify the attributes most relevant for the target binary categorization task; the contrast examples are generated as described thus far, only we use the same influential attribute dimensions for all training points. In the second "individual" form, we obtain annotator responses on individual images. The former is very inexpensive, and reasonable when human insight is possible at the top-down class level, while the latter should be more effective when the attributes' significance is more variable per example.

While in some cases attributes could be localized (like our spatial rationales above), they may also refer to global properties of the image that one could not possibly indicate with a polygon. Hence, our two rationale modalities can be complementary. Furthermore, note that just as the spatial rationales are distinct from foreground extraction, the attribute rationales are distinct from an attribute-level labeling. When asking the annotator for input, we request those attributes they deemed most important—not a complete binary indicator over all V properties.

Finally, we note that the proposed attribute-based rationale exploits the descriptive nature of attributes to allow an annotator to better intervene in training, which has not been explored in any prior work, to our knowledge. Though a simple implementation, it captures the power of attributes in a novel way.

Given rationales of either form (at least on some portion of the training images), we prepare the corresponding contrast examples $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$, and solve for the SVM hyperplane minimizing (1).

4. Data and Annotations

We explore the utility of our approach with three datasets. Central to our approach are the human annotators, who provide rationales that should give insight into

	Scenes	Hot or Not		PubFig
		MTurk	Ours	
# Annotations	8055	1845	426	247
# Unique Workers	545	104	2	8
Mean Annotations/Worker	15	18	213	31
Mean Seconds/Annotation	21.7	77.5	N/A	45

Table 1. **Annotation Statistics** - Summary from our rationale collection on Scenes [10], Hot or Not (both MTurk annotations and our own), and Public Figures [12].

their class decision for an image. We use Mechanical Turk (MTurk) to gather most of our annotations in order to create large datasets with a variety of annotation styles.

See Table 1 and Figure 4 for a summary of all datasets, and our project page[2] for screenshots of the interfaces we built to collect the rationales.

4.1. Scene Categories

We first consider the **15 Scene Categories** dataset [10], which consists of 15 indoor and outdoor scene classes, with 200-400 images per class. We select this dataset since (1) scene types often lack clear-cut semantic boundaries (e.g., "inside city" vs. "street scene"), requiring some thought by an annotator, and (2) scenes are loose collections of isolated objects with varying degrees of relevance, making them a good testbed for spatial rationales.

When gathering the scene rationales on MTurk, we imposed only two requirements on the 545 annotators: that they draw at least one polygon, and that they select one scene category label per image. See Table 1. The majority of rationales seemed intelligible (see Fig. 4, top left), some were quite specific (top center), and a minority took some artistic license (top right). Due to the subjective nature inherent in rationales, we did not remove any such cases. We also kept any incorrect scene labels (relative to the dataset ground truth), to maintain consistency with the rationales. In fact, annotators specified the wrong label 21.5% of the time, reinforcing our claim that scene categorization is not clear-cut and warrants rationales.

4.2. Hot or Not

We introduce a new dataset using images from **Hot or Not**[3], a once popular website where users rate images of one another for attractiveness on a 1-10 scale. We collected 1000 image/rating pairs of both men and women, requiring each to have been rated at least 100 times (for a robust "hotness" label). This dataset is an excellent testbed, since we expect rationales are valuable to better learn such subjective classification tasks.

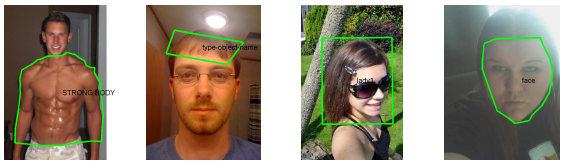
We annotated a small subset of this data ourselves for a preliminary experiment (>100 images in each of the classes), and then crowdsourced a larger portion on MTurk

[2] http://vision.cs.utexas.edu/projects/rationales/

[3] http://www.hotornot.com/

Typical Tight "Artistic"

Scene Categories



Hot, Male Not, Male Hot, Female Not, Female

Hot or Not

Youth
Smiling
Straight Hair
Narrow Eyes

Youth
Black Hair
Goatee
Square Face
Shiny Skin
High Cheekbones

Public Figure Attribute Annotations

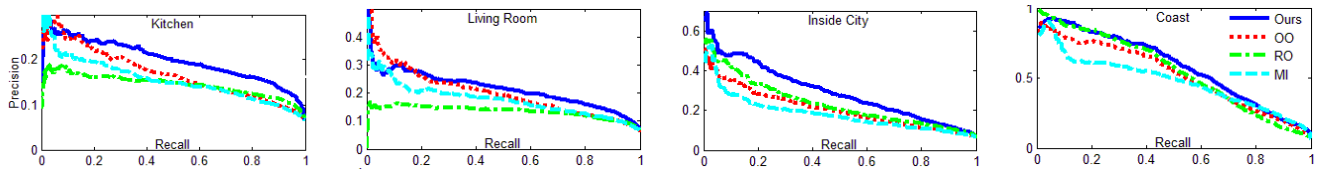
Figure 4. **Sample Annotations** - **Top:** For Scenes, annotators were shown an image of a scene, and were asked to classify it into one of the 15 categories and annotate it with one or more spatial rationales indicating the part of the image that most influenced their decision. **Middle:** For Hot or Not, annotators were shown an image of a man or a woman rated in either the top or bottom 25% of the websites' user votes for hotness, and were asked for a spatial rationale indicating what they found especially attractive/unatttractive. **Bottom:** For Public Figs, annotators were shown an image of a public figure and asked to select 3-7 attributes most relevant to an attractive vs. unattractive decision.

(see Table 1, middle). We leverage the 100+ ratings from hotornot.com as the image labels (a more robust estimate of "ground truth" than a single person's opinion), and then ask the annotators to simply answer, "Why is this person attractive?" (or unattractive). In general, rationale quality seemed quite high (see Figure 4, middle row).

4.3. Public Figures

Finally, we consider the **Public Figures** dataset [12], which contains 58,797 images of 200 people, and 73 attributes. We use it to test our attribute rationales for the attractiveness test, and leverage the binary attribute classifier outputs kindly shared by the authors to define $\mathbf{a}_1, \ldots, \mathbf{a}_V$. Given that most people in the dataset are well-known public figures, we simply divided them by identity into the attractive/unattractive categories ourselves. Of the 116 men, we took 74 as attractive, 42 unattractive; of the 84 women, we took 76 as attractive and just 8 as unattractive.

We used 51 (of the 73) attributes from [12] that we expected were possibly relevant rationales for attractiveness, such as *Smiling*, *Receding Hairline*, and *Pale Skin*.



Kitchen Living Room Inside City Coast

Figure 5. **Scene Categories Accuracy** - Precision-recall for the four scene categories most improved by our approach. Key: **Ours** = our approach; **OO** = Originals Only baseline; **RO** = Rationales Only baseline; **MI** = Mutual Information baseline.

For ease of explaining the slightly more complex and time-consuming task, we asked friends unfamiliar with the project to provide rationales (rather than MTurk; see Table 1). We showed all annotators the 51 attribute names by checkboxes, along with both the full image, and a clip of the face of interest (see project page URL). They were asked to select $\sim 3-7$ attributes *most* relevant to the given attractiveness label. From that data, we selected all attributes referenced at least 20 times to include in all experiments.

5. Results

We now present results to demonstrate that annotator rationales can be of value to build classifiers for complex visual tasks. Throughout, all methods use the same base image features and linear SVMs, and we set parameters C, C_c, and μ on validation data.

5.1. Spatial Rationales

We first evaluate the effectiveness of spatial rationales (from Section 3.2), using SIFT features with bag-of-words as the image descriptors $\boldsymbol{x}_i$, and a vocabulary of 500 visual words (we use DoG on Scenes, and dense sampling on Hot or Not). We compare to **three baselines**:

- **Originals Only**: standard image classification, no rationales.
- **Rationales Only**: uses only the spatial region given as a rationale to build the bag-of-words (BoW).
- **Mutual Information**: uses feature selection to automatically select the $k = 100$ most discriminative visual words to use as a refined bag-of-words, following the model in [5].

Because these baselines use standard BoW descriptors and SVMs as in many state-of-the-art systems, they can be considered strong baselines.

Scene Categories We begin evaluation with the Scene Categories. We pose 15 one-vs-all (OVA) classification tasks. We perform 100 trial runs per class, each time selecting a random split of 375 train/1500 test images. We present per-scene results since the rationales use a binary classifier objective, making this the most direct study of their value; it also allows us to understand the impact on different scene types separately.

Class Name	Ours	Originals Only	Rationales Only	Mutual Information
Kitchen	**0.1395**	0.1196	*0.1277*	0.1202
Living Room	**0.1238**	0.1142	0.1131	*0.1159*
Inside City	**0.1487**	0.1299	*0.1394*	0.1245
Coast	**0.4513**	*0.4243*	0.4205	0.4129
Highway	**0.2379**	*0.2240*	0.2221	0.2112
Bedroom	**0.3167**	*0.3011*	0.2611	0.2927
Street	**0.0790**	*0.0778*	0.0766	0.0775
Open Country	**0.0950**	0.0926	*0.0946*	0.0941
Mountain	**0.1158**	0.1154	0.1151	*0.1154*
Office	**0.1052**	*0.1051*	0.1051	0.1048
Tall Building	**0.0689**	0.0688	*0.0689*	0.0686
Store	**0.0867**	*0.0866*	0.0857	0.0866
Forest	**0.4006**	0.3956	*0.4004*	0.3897
Suburb	0.0735	0.0735	**0.0737**	0.0733
Industrial	0.1046	**0.1056**	0.0911	0.0981

Table 2. **Scene Categories Accuracy** - Mean average precision for rationales (Ours) and 3 baselines. Our approach is most accurate for 13/15 classes. Best overall is in bold, best baseline is italicized.

Table 2 summarizes the results. Our method improves the MAP over the *best* baseline in 13 of the 15 categories. The gain is statistically significant for 11 out of 15 classes ($\alpha = 0.1$). Figure 5 shows precision-recall curves for the 4 categories most improved by our approach. The gains on Living Room and Inside City are intuitive, since these are scenes particularly defined by loose configurations of rationale-friendly objects. Gains on Coast are more modest; upon inspection, we found that rationales for this class often encompassed most of the image, hence our similar accuracy to Originals and Rationales Only.

The fact that rationales outperform the Originals baseline shows the clear value in asking annotators for spatial explanations. Moreover, outperforming the Rationales Only baseline shows that a rationale is *not* equivalent to foreground segmentation; we do not see as strong of results by simply cropping out the parts of the image that do not lie in a rationale polygon. Finally, our advantage over Mutual Information shows that human insight can be competitive and even more useful than an automated feature selection technique, if utilized appropriately.

Hot or Not Next we evaluate the human attractiveness classification task on the Hot or Not data. We densely sample SIFT at every 2 pixels at a single scale of 8 pixels, using the VLFeat library [23], to ensure good coverage in all ar-

	Male		Female	
Training Examples per Class	$N = 25$	100	25	100
Ours (Our Annotations)	**55.40%**	**60.01%**	53.13%	**57.07%**
Ours (MTurk Annotations)	53.73%	54.92%	53.83%	56.57%
Originals Only	52.64%	54.86%	**54.02%**	55.99%
Rationales Only	51.07%	54.01%	50.06%	50.00%
Mutual Information	52.51%	54.50%	52.58%	53.94%
Faces as Rationales	52.17%	53.40%	53.39%	56.11%

Table 3. **Hot or Not Accuracy** - Spatial rationales on the Hot or Not dataset for 2 training set sizes, compared to several baselines.

eas of the image, including the smoother face regions. We build separate classifiers per gender. Since the raw data has real-valued rankings of hotness, we map them to binary labels by taking the top 25% of ratings as Hot and the bottom 25% as Not. We show results for training sets of size 25 and 100 per class, randomly selected across 100 trials.

We compare results to the 3 baselines outlined above, plus a new face-specific baseline that uses Viola-Jones [25] to detect the face in the image, and only takes those features. We call this the **Faces as Rationales** baseline, since it will show how well one could do by simply focusing on the face *a priori*.

Table 3 shows the results. Our rationales clearly outperform the several baselines for Males. The improvement is smaller when we use the MTurk rationales from 104 workers. While we did not find an obvious difference in quality of the rationales, a likely explanation is that the particular tastes of the single annotator providing "our annotations" were more consistent and thus more accurately learnable.

The Males result also illustrates the potential for rationales to reduce total human effort. With just 25 training examples per class, our approach outperforms the traditional supervised approach trained with 100 examples per class. Although rationales do cost more time to obtain, we estimate it adds less than a factor of four in annotator time. Thus, beating the baseline with just a quarter of the training data (25 vs. 100) is a win for total annotation effort.

For Females, improvements are much less significant, very close to the Originals baseline; the Face Rationale baseline suggests that the female ratings were largely dependent on facial appearance. We leave as future work to explore even richer low-level features, and to exploit the face detector for our own method's advantage.

5.2. Attribute Rationales

Finally, we demonstrate our second form of visual rationales, using attributes for the human attractiveness classification task. We use the Public Figures dataset instead of Hot or Not for the attributes rationales, since we can leverage the collection effort of Kumar *et al*. [12], who publicly share the classifier outputs of their attribute models. (Refer back to Secs. 3.3 and 4.3 for algorithm and data details.)

We compare our attribute rationales to the Originals Only baseline, which learns an SVM attractiveness classifier using the same original attribute feature space. Note, this setting does not lend itself to a parallel Rationales Only baseline, as we showed above.

	Homogeneous		Individual	
	Ours	**Originals**	**Ours**	**Originals**
Male	**68.14%**	64.60%	**62.35%**	59.02%
Female	**55.65%**	51.74%	51.86%	52.36%

Table 4. **Attribute Rationales Performance** - Average accuracy across 100 trials for homogeneous and individual attribute rationales. Our homogeneous rationales attribute approach performs significantly better ($\alpha = 0.0001$ in a one-sided t-test) than the baseline for both Males and Females. Our individual rationales attribute approach performs significantly better than the baseline for Males, but the data is inconclusive for Females ($\alpha = 0.001$).

Homogeneous Rationales We first test the homogeneous attribute rationales, where rationales are propagated down for the entire category based on human knowledge. For this test, we selected attributes *Chubby*, *Big Nose*, and *Senior* as indicative of the unattractive category, and left *Mouth Closed*, *Square Face*, *Round Face*, and *No Beard* as "neutral" in the contrast examples.[4] We use a training set of 15 images per class, and 500 test images per class. We split the train/test sets so as to ensure that the same person (e.g., Meg Ryan) never appears in *both* the training and test set, and run 100 randomly selected splits.

Table 4, left, shows the results. For both men and women, our approach performs statistically significantly better than the baseline. With just 7 attributes and 15 training examples per class, these results show that homogeneous rationales can give useful insight to the classifier. The minimal training cost entailed by homogeneous attribute-based rationales is also a clear strength.

Individual Rationales Next we make use of the rationales collected on the same data from annotators on individual images. We use a training set of 15 examples per class, and 500 test examples per class, and perform 50 trials over different splits (again ensuring no single person is in both the training and test set on a given trial), setting $\delta = 2$ for Males and $\delta = 1$ for Females (Equation 6).

Table 4, right, shows the results. For Males, our results show a statistically significant performance increase over the baseline. For Females, however, they are comparable. This may be due to several factors, including the the "annotatability" of these attributes (since we wanted to use existing attribute data, the attribute vocabulary is not necessarily the most amenable to attractiveness explanations), and consistency across annotators. In addition, overall performance was likely negatively impacted by the small training set size

[4]The authors cringe at putting such offensive judgments in this paper; this classification task is not very PC.

and, in the case of the women, the desperate brutality with which the authors deemed unattractive a few of the quite attractive women in this dataset, due to the need for *some* unattractive Female examples.

Overall, these attractiveness results both show the real difficulty of learning such a complex visual property, as well as the promise of allowing human insight to be more fully transferred to the learning algorithms. It is natural to imagine further applications of our approach in which spatial and attribute rationales could be combined to strengthen classification performance.

6. Conclusions

We presented a new way to look at supervised learning of image classes: by using not only the "what" of an annotator's classification, but also the "why". Our results from thorough experiments and multiple datasets are very encouraging. They show that asking an annotator to not only label an image with its class but also with the regions or attributes that were most influential in his or her class choice can be useful in multiple domains, including recognizing scene categories and classifying images of humans as attractive or unattractive.

The results indicate some potential for rationales to reduce total annotator effort; once the annotator decides the subjective class label, s/he has already invested some time in the task, and so further requesting a rationale amortizes that analysis. Nonetheless, we stress that rationales are not only useful for scenarios with small training sets. They give a classifier useful knowledge about the human's opinion that cannot be replicated simply with more labels.

This general strategy for image classification may be useful in many other domains as well, especially for recognition tasks that require perceptual or subjective judgment. In all, this study suggests new possibilities to allow systems to communicate more fully with human annotators, for better ultimate performance.

Acknowledgements We thank all the annotators who contributed their rationales. This research is supported in part by ONR ATL N00014-11-1-0105 and NSF CAREER IIS-0747356.

References

[1] L. Bourdev and J.Malik. Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations. In *ICCV*, 2009.

[2] E. Chang, S. Tong, K. Goh, and C.-W. Chang. Support Vector Machine Concept-Dependent Active Learning for Image Retrieval. *IEEE Trans. on Multimedia*, 2005.

[3] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, 2005.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[5] G. Dorko and C. Schmid. Selection of Scale-Invariant Parts for Object Class Recognition. In *ICCV*, 2003.

[6] G. Druck, B. Settles, and A. McCallum. Active Learning by Labeling Features. In *EMNLP*, 2009.

[7] Y. Eisenthal, G. Dror, and E. Ruppin. Facial Attractiveness: Beauty and the Machine. *Neural Computation*, 2006.

[8] I. Endres, A. Farhadi, D. Hoiem, and D. Forsyth. The Benefits and Challenges of Collecting Richer Object Annotations. In *Wkshp on Adv Comp Vis with Humans in the Loop*, 2010.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 2010.

[10] L. Fei-Fei and P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In *CVPR*, 2005.

[11] S. J. Hwang and K. Grauman. Reading Between the Lines: Object Localization Using Implicit Cues from Image Tags. In *CVPR*, 2010.

[12] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifier for Face Verification. In *ICCV*, 2009.

[13] C. Lampert, H. Nickisch, and S. Harmeling. Learning to Detect Unseen Object Classes by Between-Class Attribute Transfer. In *CVPR*, 2009.

[14] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*, 2006.

[15] T. Leyvand, D. Cohen-Or, G. Dror, and D. Lischinski. Data-Driven Enhancement of Facial Attractiveness. In *SIGGRAPH*, 2008.

[16] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum. Learning to Detect a Salient Object. In *CVPR*, 2007.

[17] D. Parikh and K. Grauman. Interactively Building a Discriminative Vocabulary of Nameable Attributes. In *CVPR*, 2011.

[18] H. Raghavan, O. Madani, and R. Jones. Interactive Feature Selection. In *IJCAI*, 2005.

[19] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: a Database and Web-Based Tool for Image Annotation. Technical report, MIT, 2005.

[20] B. Siddiquie and A. Gupta. Beyond Active Noun Tagging: Modeling Contextual Interactions for Multi-Class Active Learning. In *CVPR*, 2010.

[21] A. Sorokin and D. Forsyth. Utility Data Annotation with Amazon Mechanical Turk. In *CVPR*, 2008.

[22] M. Spain and P. Perona. Some Objects Are More Equal Than Others: Measuring and Predicting Importance. In *ECCV*, 2008.

[23] A. Vedaldi and B. Fulkerson. VLFeat, 2008.

[24] S. Vijayanarasimhan and K. Grauman. Multi-Level Active Prediction of Useful Image Annotations for Recognition. In *NIPS*, 2008.

[25] P. Viola and M. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. In *CVPR*, 2001.

[26] L. von Ahn, R. Liu, and M. Blum. Peekaboom: A Game for Locating Objects in Images. In *CHI*, 2006.

[27] P. Welinder and P. Perona. Online Crowdsourcing: Rating Annotators and Obtaining Cost-effective Labels. In *Wkshp on Adv Comp Vis with Humans in the Loop*, 2010.

[28] O. Zaidan, J. Eisner, and C. Piatko. Using Annotator Rationales to Improve Machine Learning for Text Categorization. In *NAACL - HLT*, 2007.