Understanding and Predicting Importance in Images

Alexander C. Berg[1], Tamara L. Berg[1], Hal Daumé III[4],
Jesse Dodge [6], Amit Goyal [4], Xufeng Han[1], Alyssa Mensch [5],
Margaret Mitchell[2], Aneesh Sood[1], Karl Stratos [3], Kota Yamaguchi[1]

[1]Stony Brook University [2] University of Aberdeen [3] Columbia University
[4] University of Maryland [5] Massachusetts Institute of Technology [6] University of Washington

Abstract

What do people care about in an image? To drive computational visual recognition toward more human-centric outputs, we need a better understanding of how people perceive and judge the importance of content in images. In this paper, we explore how a number of factors relate to human perception of importance. Proposed factors fall into 3 broad types: 1) factors related to composition, e.g. *size, location, 2) factors related to semantics,* e.g. *category of object or scene, and 3) contextual factors related to the likelihood of attribute-object, or object-scene pairs. We explore these factors using what people describe as a proxy for importance. Finally, we build models to predict what will be described about an image given either known image content, or image content estimated automatically by recognition systems.*

Figure 1. An image from the UIUC Pascal sentence dataset [20] with 3 descriptions written by people.

1. Introduction

Consider Figure 1. Despite the relatively small image space occupied by the people in the boat, when humans describe the scene they mention both the people ("3 adults and two children", "Four people", "Several people"), and the boat ("raft", "canoe", "canoe"). The giant wooden structure in the foreground is all but ignored, and the cliffs in the background are only mentioned by one person. This suggests a significant and interesting bias in perceived content importance by human viewers!

Now that visual recognition algorithms are starting to work – we can reliably recognize tens or hundreds of object categories [11, 6, 24, 10], and are even beginning to consider recognition at a human scale [2, 17] – we need to start looking closely at other questions related to image understanding. Current systems would treat recognition of all objects in pictures like Fig. 1 as equally important, despite indications that humans do not do so. Because people are often the end consumers of imagery, we need to be able to adopt human-centric views of recognition, especially in user applications such as image or video search. For example, in response to an image search for "tree", returning an image with a tree that no person would ever mention is not desirable.

In this paper we consider the problem of understanding and predicting perceived importance of image content. A central question we pose is: what factors do people inherently use to determine importance? We address this question using descriptions written by people as indicators of importance. For example, consider Fig. 2. Despite containing upwards of 20 different objects, people asked to describe the image tend to mention quite similar content aspects: the man, the baby, the beard, and the sling (and sometimes the kitchen). This suggests there are some underlying consistent factors influencing people's perception of importance in pictures.

We study a number of possible factors related to per-

Figure 2. Not all content is created equal – as indicated by the descriptions people write (**right**). Some objects (*e.g.* man, baby, sling) seem to be more important than others (*e.g.* ladder, table, chair). Some attributes seem to be more important (*e.g.* beard) than others (*e.g.* shirt, or glasses). Sometimes scene words are used (*e.g.* kitchen), and sometimes they aren't. We examine a number of compositional and semantic factors influencing importance.

ceived importance, including: a) factors related to image composition such as size and location, b) factors related to content semantics such as category of object (*e.g.* perhaps people are more important to viewers than rocks), and category of scene, and c) factors related to context, including object-scene or attribute-object context. The influence of each factor is first explored independently in a number of experiments on large datasets with labeled image content and associated descriptions written by people. Next, we build models for predicting importance using combinations of our proposed factors. Two scenarios are explored, models given known image content, and models given estimated image content predicted by state of the art visual recognition methods.

Our paper makes several novel contributions:

- We propose a variety of factors for human perceived importance, including factors related to composition, semantics, and context.
- We evaluate the individual influence of each of our proposed factors on importance using (existing and gathered) image content labels and descriptions written by people.
- We build models to predict what humans describe given known image content (to understand factor importance), or automatically estimated content (to understand what is possible using current computer vision techniques).

A relatively small amount of previous work has investigated content importance [4, 14, 22, 23], but these all take an object-centric stance, predicting the importance of objects in the image. We expand the problem to also include prediction of importance for scenes and attributes. For example, an image might portray a particularly iconic example of a kitchen (*e.g.* Fig. 3, 4th picture from left), resulting in users describing the scene as a "kitchen". Sometimes an attribute of an object might be relatively unusual, *e.g.* a "pink elephant" or "bearded man", resulting in all viewers describing these salient characteristics (*e.g.* Fig. 2). We also expand the types of factors considered, moving from purely compositional [22, 23, 4] toward factors related to semantics or context, which we find to be *much stronger indicators* of importance than factors related purely to composition. Finally, we complete the prediction loop to take an input image, estimate content using visual recognition, and make predictions for description.

1.1. Related work

Predicting Importance: Elazary and Itti [4] consider object naming order in the LabelMe dataset [21] as a measure of the interestingness of an object and compare that to salient locations predicted by computational models of bottom-up attention. In the ESP game [25], two players type words related to an image and receive points as soon as a matching word is obtained. Intuitively, important content elements should occur earlier in this process, but noise, game scheming, and results spread across multiple short games for each image make translating between game results and importance somewhat difficult. Hwang et al [14] use Kernel Canonical Correlation Analysis to discover a "semantic space" that captures the relationship between ordered tag cues and image content to improve image retrieval.

Perhaps most similar to our goals, elegant work by Spain and Perona [22, 23] tackles the task of examining factors to predict the order in which objects will be mentioned given an image. Our work is distinct in a number of ways: 1) we study whether image contents are included in natural language descriptions as opposed to studying ordered lists of objects, 2) we consider how category of object influences importance, something Spain and Perona [22, 23] do not consider, 3) we examine importance factors on a larger scale with datasets of 1000 and 20,000 images, as opposed to 97 images, 4) we explore importance for scenes and attributes in addition to objects, and 5) we use the output of computer vision algorithms to predict what is described.

Human-Centric Recognition: Computational recognition is beginning to move toward examining content from a hu-

man perspective, including the shift away from purely object based outputs, toward including attributes [26, 1, 7], scenes [18, 21], or spatial relationships [15, 13]. Other related work includes attempts to compose natural language descriptions for images [15, 16, 19, 8]. Especially relevant – in fact almost the "dual" to this paper – is recent work in natural language processing predicting whether pieces of text refer to visual content in an image [3]. However, none of these approaches focus on predicting perceived importance and how it could influence what to recognize in (or describe about) an image.

1.2. Overview of the Approach

We start by describing the data we use for investigating importance (Sec 2.1), how we gather labels of image content (Sec 2.2), and how we map from content to descriptions (Sec 2.3). Next we examine the influence of each of our proposed importance factors, including compositional (Sec 3.1), semantic (Sec 3.2), and contextual (Sec 3.3). Finally, we train and evaluate models to predict importance given known image content (Sec 4.1) or given image content estimated by computer vision (Sec 4.2).

2. Data, Content Labels, & Descriptions

To study importance in images we need three things: large datasets consisting of *images*, *ground truth content* in those images, and indicators of *perceptually important content*. For the first requirement, we use two large existing image collections (described in Sec 2.1). For the second requirement, we use existing content labels or collect additional labels using Amazon's Mechanical Turk service (Sec 2.2). For the last requirement, we make use of existing descriptions of the images written by humans. As illustrated in Fig. 2, what people describe when viewing an image can be a useful proxy for perceived importance. Finally, we map between what the humans judge to be important (things mentioned in descriptions) to the labeled image content by hand or through simple semantic mapping techniques (Sec 2.3).

2.1. Data

We use two data collections to evaluate our proposed factors for importance: the ImageCLEF dataset [12], and the UIUC Pascal Sentence dataset [20]. ImageCLEF is a collection of 20K images covering various aspects of contemporary life, such as sports, cities, animals, people, and landscapes. The original IAPR TC-12 Benchmark [12] includes a free-text description for each image. Crucially, in its expansion (SAIAPR TC-12), each image is also segmented into constituent objects and labeled according to a set of (275) labels [5]. From here on we will refer to this dataset, and its descriptions as ImageCLEF. This dataset is quite large, and allows us to explore some of our proposed factors at scale. However, some factors are still difficult to measure well – *e.g.* scene, or contextual factors – because they require collecting additional content labels not present in the dataset, somewhat difficult for the large scale ImageCLEF data.

Therefore, we use the UIUC Pascal Sentence data set (UIUC), which consists of 1K images subsampled from the Pascal Challenge [6] with 5 descriptions written by humans for each image. As with all Pascal images, they are also annotated with bounding box localizations for 20 object categories. This dataset is smaller than the ImageCLEF dataset, and is a reasonable size to allow collecting additional labels with Mechanical Turk (Sec 2.2). Crucially this dataset has also been explored by the vision community, resulting in object detectors [10], attribute classifiers [15], and scene classifiers [28] for effectively estimating image content.

2.2. Collecting Content Labels

We use three kinds of content labels: object labels, scene labels, and attribute labels. Object labels are already present in each of our data collections. ImageCLEF images are segmented and each segment is associated with an object category label. In the UIUC dataset there are bounding boxes around all instances of the 20 PASCAL VOC objects. To gather additional content labels – for scenes and attributes in the UIUC dataset – we use Mechanical Turk (MTurk).

To procure scene labels for images, we design a MTurk task that presents an image to a user and asks them to select the scene which best describes the image from a list of 12 scenes that cover the UIUC images well. Users can also provide a new scene label through a text box denoted with "other," or can select "no scene observed" if they cannot determine the scene type. In addition, we ask the users to rate their categorization, where a rating of 1 suggests that the scene is only barely interpretable from the image, while 5 indicates that the scene category is obviously depiction. Each image is viewed and annotated by 5 users.

Our MTurk task for labeling attributes of objects presents the user with a cropped image around each object. Each of 3430 objects from the UIUC data are presented separately to labelers along with a set of possible attributes. Three users select the attributes for each object.

2.3. Mapping Content Labels to Descriptions

We also require a mapping between labeled image content and text descriptions for each image. For example, if an image contains 2 people as in Fig. 2, we need to know that the "man" mentioned in description 1 refers to one of those people, and the "child" refers to the other person. We also need mappings between scene category labels and specific words in the description, as well as between attribute categories and words (e.g. the "bearded man" refers to a person with a "has beard" attribute). In the case of a small

Figure 3. We hypothesize 3 different kinds of factors for perceived importance: **Compositional factors** related to object size or placement, **Semantic factors** such as the type of object (people might be inherently more important to a human observer), or obviousness of a scene category (extremely iconic or strong depictions of a scene might influence whether the scene name will be mentioned), **Context factors** such as unusual objects in a scene (trees in the water), or relatively unusual attributes of objects (bearded men).

dataset like the UIUC collection, this mapping can be done by hand. For the larger ImageCLEF dataset, we devise an automatic method for mapping. Note that for ImageCLEF, only objects are explored due to cost of labeling other content (Sec 2.2).

For ImageCLEF data, we need to map between objects labeled in the image, and mentions of those objects in associated descriptions. Determining if a label is referred to is not as simple as matching the nouns in the description with the label type because the writer might use a synonymous term (*e.g.* boy for person) or a more specific term (*e.g.* "Chihuahua" for dog). We use a simple WordNet based measure of semantic distance [27] and find that this works well (F1 score of 0.94), especially for sibling terms (*e.g.* "boat" and "ship"). The hierarchical nature of WordNet also provides valuable matching capability because image labels tend to be general (*e.g.* "person") while terms used in descriptions are often more specific (*e.g.* "cyclist").

3. Exploring Importance Factors

We propose a number of potential factors for perceived importance and evaluate how each impacts importance by studying its relationship with user descriptions. The factors we examine come in 3 flavors: (a) compositional factors related to object placement and size (Sec 3.1), (b) semantic factors related object or scene type (and depiction strength) (Sec 3.2), and (c) context factors related to common vs unusual (attribute-object or object-scene) content pairs (Sec 3.3). Figure 3 illustrates each kind of factor.

3.1. Compositional Factors

We consider two types of compositional factors, object size and location (left two images in Fig 3 demonstrate their effects). Since we have images with labeled content – labeled segmentations (ImageCLEF) or bounding boxes (UIUC) – we can automatically measure the impact of size and location on whether an object is described by a human viewer. Size is measured as the object size (number of pixels), normalized by image size. Location is measured as the distance from the image center to the object center of mass, normalized by image size.

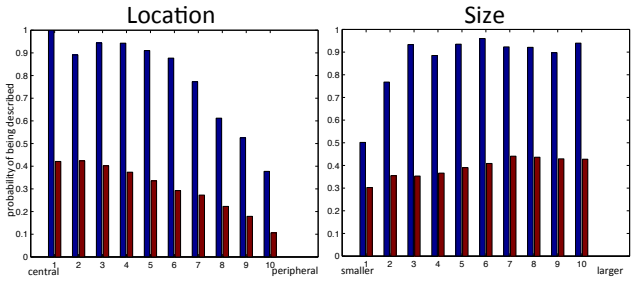

Figure 4. Objects further away from the image center are less likely to be mentioned (left). The bigger an object is, the more likely it is mentioned (right), unless it is very large. In general objects from ImageCLEF (red) are less likely to be described than objects from the UIUC (blue) dataset.

Fig. 4 displays the effect of location (left) and size (right) on description probability for the two datasets (where larger bin number indicates further from the center, and larger respectively). The results support the intuition that objects that are larger and closer to the center are more likely to be described (perceived as important). Note the slight downhill slope in the size plot in Fig 4, 3 rightmost bins; this is presumably because sometimes very large objects are inherently background content and are not mentioned *e.g.*, "sky" (similar effects were observed in previous work [22, 23]).

3.2. Semantic Factors

In addition to the compositional factors explored in past work [22, 23], we also examine semantic factors related to the categories of content present within an image. We study two kinds of semantic information – how the category of an object influences the probability that the object will be described (Fig. 3, 3rd image), and how the scene category of an image and strength of its depiction influences the probability that the scene type will be described (Fig. 3).

Object Type: Intuitively, we hypothesize that some object

Top10	Prob	Last10	Prob
firework	1.00	hand	0.15
turtle	0.97	cloth	0.15
horse	0.97	paper	0.13
pool	0.94	umbrella	0.13
airplane	0.94	grass	0.13
bed	0.92	sidewalk	0.11
person	0.92	tire	0.11
whale	0.91	smoke	0.09
fountain	0.89	instrument	0.07
flag	0.88	fabric	0.07

Table 1. Probability of being mentioned when present for various object categories (ImageCLEF).

	Prob-ImageCLEF	Prob-Pascal
Animate	0.91	0.84
Inanimate	0.53	0.55

Table 2. Probability of being mentioned when present for Animate versus Inanimate objects.

Object	Prob	Object	Prob
horse	0.99	bus	0.80
sheep	0.99	motorbike	0.75
train	0.99	bicycle	0.69
cat	0.98	sofa	0.59
dog	0.96	dining table	0.56
aeroplane	0.97	tv/monitor	0.54
cow	0.95	car	0.43
bird	0.93	potted plant	0.26
boat	0.90	bottle	0.26
person	0.81	chair	0.26

Table 3. Probability of being mentioned when present for various object categories (UIUC).

categories are more important to human observers than others. For example, being human we expect that "people" in an image will attract the viewers attention and thus will be more likely to be described. For example, in Fig. 3, 3rd picture, the caption reads "Girl in the street" despite the large and occluding bicycle in front of her.

Table 1 and Table 3 show the 10 most probable object types and 10 least probable object types from the ImageCLEF and UIUC datasets respectively, sorted according to the probability of being described when present in an image. A few observations can be drawn from these statistics. Very unusual objects tend to be mentioned; we don't see fireworks very often, so when an image contains one it is likely to be described. The person category also ranks highly in both lists because of our inherent human tendency to pay attention to people. In contrast, the objects deemed non-salient tend to be very common ones (sidewalk, tire), too generic (smoke, cloth), or part of something more salient (hand). In the UIUC data (Table 3) we observe that (Table 3) less semantically salient objects (*e.g.* chair, potted plant, bottle) are described with lower probability than more interesting ones (cow, cat, person, boat).

Rating	1	2	3	4	5
Prob	0.15	0.21	0.21	0.22	0.26

Table 5. Probability of Scene term mentioned given Scene depiction strength (1 implies the scene type is very uncertain, and 5 implies a very obvious example of the scene type, as rated by human evaluators). Scenes are somewhat more likely to be described when users provide higher ratings.

It is also interesting to note that *animate objects* are much more likely to be mentioned when present than inanimate ones in both datasets (Table 2). From these results one could hypothesize that observers usually perceive (or capture) the animate objects as the subject of a photograph and the more common objects (*e.g.* sofa, grass, sidewalk) as background content elements.

Scene Type & Depiction Strength: In Sec 2.1 we described our Mechanical Turk experiments to gather scene labels and depiction strength ratings for the UIUC images. In addition, we also annotate whether the scene category is mentioned in each image's associated descriptions. The relationship between scene type and description is shown in Table 4. Some scene categories are much more likely to be described when depicted (*e.g.* office, kitchen, restaurant) than others (*e.g.* river, living room, forest, mountain). In general we find that the scene category is much more likely to be mentioned for indoor scenes (ave 0.25) than for outdoor scenes (ave 0.12).

Finally, we also look at the relationship between scene depiction strength – whether the scene is strongly obvious to a viewer as rated by human observers – and description. We expect that the more obvious the scene type (*e.g.* the iconic kitchen depiction in Fig. 3), the more likely it is to be described by a person (*e.g.* "kitchen in house"). Using a scene word can be more succinct and informative than naming all of the objects in the scene, but the scene type needs to be relatively obvious for an observer to name it. We observe (Table 5) that scene depiction strength has some correlation with description probability.

3.3. Context Factors

We examine two kinds of contextual factors and their influence on description. The first is object-scene context, hypothesizing that the setting in which an object is depicted will have an impact on whether or not the object will be described. We visualize the probability of an object being described given that it occurs in a particular scene in Fig. 5. Some interesting observations can be made from this plot. For example, bicycles in the dining room are more likely to be described than those on the street (perhaps because they are in an unusual setting). Similarly, a TV in a restaurant is

office	airport	kitchen	dining room	field	living room	street	river	restaurant	sky	forest	mountain
0.29	0.13	0.36	0.21	0.16	0.13	0.18	0.1	0.28	0.18	0.0	0.07

Table 4. Probability of description for each Scene Type.

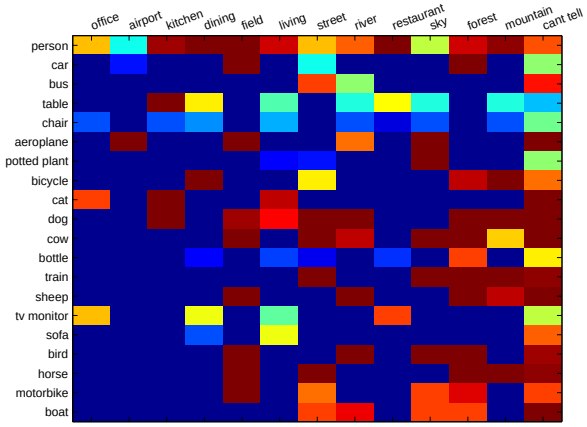

Figure 5. The impact of Object-Scene context on description. Colors indicate the probability of an object being mentioned given that it occurs in a particular scene category (red - high, blue - low). Objects in relatively unusual settings (*e.g.* bicycles in the dining room) are more often described than those in ordinary settings (*e.g.* bicycles in the street).

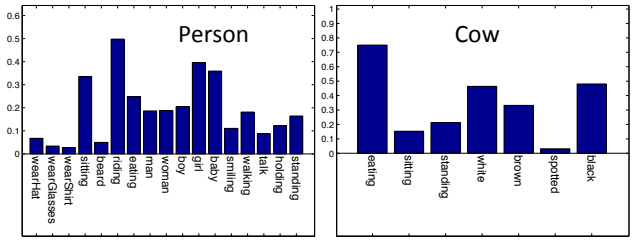

Figure 6. The impact of Attribute-Object context, showing the probability of an attribute being mentioned given that it occurs with a particular object. More unusual attributes (*e.g.riding* person) tend to be mentioned more often than relatively common attributes (*e.g.standing*).

more likely to be described than a TV in a living rooms. In images where the scene is unclear (perhaps because they are object focused images), the objects present are very likely to be described.

The second contextual factor we explore is attribute-object context. Here we compute the probability of an attribute being described given that it occurs as a modifier for a particular object category. Example results for the "person" and "cow" categories are shown in Fig. 6. Notice that for person, riding is more likely to be described than other actions (*e.g.* sitting, smiling), perhaps because it is a more unusual activity. Wearing a hat is also more likely to be described, followed by beard, wearing glasses, and wearing shirts (a similar ordering to how usual or unusual these attributes of people tend to be). For cows, eating is the most likely attribute, over actions like standing or sitting (lying down). Color attributes also seem to be described frequently. Similar observations were made for the other object categories.

Model	Features	Accuracy% (std)
Baseline (ImageCLEF)		57.5 (0.2)
Log Reg (ImageCLEF)	$K_o^s + K_o^l$	60.0 (0.1)
Log Reg (ImageCLEF)	K_o^c	68.0 (0.1)
Log Reg (ImageCLEF)	$K_o^c + K_o^s + K_o^l$	69.2 (1.4)
Baseline (UIUC-Kn)		69.7 (1.3)
Log Reg (UIUC-Kn)	$K_o^s + K_o^l$	69.9 (0.6)
Log Reg (UIUC-Kn)	K_o^c	79.8 (1.4)
Log Reg (UIUC-Kn)	$K_o^c + K_o^s + K_o^l$	82.0 (0.9)
Baseline (UIUC-Est)		76.5 (1.0)
Log Reg (UIUC-Est)	$E_o^s + E_o^l$	76.9 (1.1)
Log Reg (UIUC-Est)	E_o^c	78.9 (1.4)
Log Reg (UIUC-Est)	$E_o^c + E_o^s + E_o^l$	79.52 (1.2)

Table 6. Accuracy of models for predicting whether a given object will be mentioned in an image description on the ImageCLEF and UIUC datasets given known image content (UIUC-Kn) and visual recognition estimated image content (UIUC-Est). Features – K_o^c for known object category, K_o^s for known object size, K_o^l from known object location, E_o^c estimated object category, E_o^s estimated object size, E_o^l estimated object location.

4. Predicting Importance

We train discriminative models to predict importance – using presence or absence in a description as proxy for an importance label. These models predict importance given as input: known image content (Sec 4.1), or estimates of image content from visual recognition systems (Sec 4.2). For each input scenario, we train 3 kinds of models: a) models to predict whether an object will be mentioned, b) models to predict whether a scene type will be mentioned, and c) models for predicting whether an attribute of an object will be mentioned. We use four fold cross validation to train and estimate the accuracy of our learned models. This is repeated 10 time with different random splits of the data in order to estimate the mean and standard deviation of the accuracy estimates. Logistic Regression is used for prediction with regularization trade-off, C, selected by cross-validation on subsets of training data.

Model	Features	Accuracy% (std)
Baseline (UIUC-Kn)		86.0 (0.2)
Log Reg (UIUC-Kn)	K_s^c+K_s^r	96.6 (0.2)
Log Reg (UIUC-Est)	E_s^d	87.4 (1.3)

Table 7. Accuracy of models for predicting whether a particular scene will be mentioned in the UIUC dataset given known (UIUC-Kn) and visual recognition estimated image content (UIUC-Est). Features – K_s^c indicates known scene category, K_s^r user provided scene depiction rating, and E_s^d estimated scene descriptor (classification scores for 26 common scene categories).

Model	Features	Accuracy% (std)
Baseline (UIUC-Kn)		96.3 (.01)
Log Reg (UIUC-Kn)	K_a^c+ K_o^c	97.0 (.01)
Log Reg (UIUC-Est)	E_a^d+E_o^c	96.7 (.01)

Table 8. Accuracy of models for predicting whether a specific attribute type will be mentioned in the UIUC dataset given known (UIUC-Kn) and visual recognition estimated image content (UIUC-Est). Features – K_a^c known attribute category, K_o^c known object category, E_o^c estimated object detection category, E_a^d estimated attribute descriptor (vector of predictions from 21 attribute classifiers on the object detection window).

4.1. Predicting Importance from Known Content

Object Prediction: We train a model to predict: given an object and its localization (bounding box), whether it will be mentioned in descriptions written by human observers. We first look at a simple baseline, predicting "Yes" for every object instance. This provides reasonably good accuracy (Table 6), 57.5% for ImageCLEF and 69.7% for UIUC. Next we train several models using features based on object size, location, and type – where size and location are each encoded as a number $\in [0, 1]$, and type is encoded in a binary vector with a 1 in the kth index indicating the kth object category (out of 143 categories for ImageCLEF and 20 for UIUC). Results are shown in Table 6, with full model accuracies (using all features) of 69.2% and 82.0% respectively on the two datasets. Interestingly, we observe that for both datasets, the semantic object category feature – not included in some previous studies predicting importance – is the strongest feature for predicting whether an object will be described, while compositional features are less helpful.

Scene Prediction: We train one model for each common scene category (categories described at least 50 times). For example, the kitchen model will predict: given an image and its scene type, whether the term "kitchen" will appear in the description associated with that image. Positive samples for a scene model are those images where at least one human description contains that scene category, negative samples are the rest. Our descriptor is again a binary vector, this time encoding image scene type, plus an additional index corresponding to user provided rating of scene strength. Results are shown in Table 7. The baseline – always predicting "No" scene mentioned – obtains an accuracy of *86.0%*. Prediction using image scene type and rating improves this significantly to *96.6%*.

Attribute Prediction: We train one model for each common attribute category (categories described at least 100 times in the UIUC dataset). For example, the "pink" model will predict: given an object and its appearance attributes, whether the attribute term "pink" will appear in the description associated with the image containing the detection. Positive samples for an attribute model are those detections where at least one human image description contains the attribute term, negative samples the rest. Our input descriptor for the model is a binary vector encoding both object category and attribute category (to account for attribute semantics and attribute-object context). Results are shown in Table 8. The baseline – always predicting "No" attribute mentioned for all detections – obtains an accuracy of 96.3% due to apparent human reluctance to utilize attribute terms. Our model improves prediction accuracy to 97.0%.

4.2. Predicting Importance from Estimated Content

Next, we complete the process so that importance can be predicted from images using computer vision based estimates of image content. Specifically we start with a) an input image, then b) estimates of image content are made using state of the art visual recognition methods, and finally c) models predict what objects, scenes, and attributes will be described by a human viewer. Recognition algorithms estimate 3 kinds of image content on the UIUC dataset: objects, attributes, and scenes. Objects for the 20 Pascal categories are detected using Felzenszwalb *et al.*'s mixtures of deformable part models [9]. 21 visual attribute classifiers, including color, texture, material, or general appearance characteristics [15] are used to compute a 21-dimensional appearance descriptor for detected objects. We obtain scene descriptors for each image by computing scores for 26 common scene categories using the SUN dataset classifiers [28].

Object Prediction: For objects, we train models similar to Sec 4.1, except using automatically detected object type, size, and location as input features. Results are shown in the last 4 rows of Table 6. Note that though the object detectors are somewhat noisy, we get *comparable* results to using known ground truth content (and better results than the baseline of classifying all detections as positive). This may be because the detectors most often miss detections of small, background, or occluded objects – those that are also less likely to be described. Performance of our complete model is *79.5%* compared to the baseline of *76.5%*.

Scene Prediction: As in Sec 4.1, for each scene category we train a model to predict: given an image whether that scene type will be mentioned in the associated description. However, here we use our 26 dimensional scene descriptor

as the feature. Results are shown in Table 7 bottom row. Although the set of scenes used to create our descriptor are not exactly the same as the set of described scene categories in the dataset, we are still able to obtain a classification accuracy of *87.4%* compared to the baseline of *86.0%*.

Attribute Prediction: As in Sec 4.1, for each attribute we train a model to predict: given an object detection, what attribute terms will be used in an associated image description. We use our 21d attribute descriptor as a feature vector. Results are shown in Table 8 bottom. Attribute prediction based on estimated attribute descriptor and object category yields *96.7%* compared to the baseline of *96.3%*.

5. Discussion and Conclusion

We have proposed several factors related to human perceived importance, including factors related to image composition, semantics, and context. We first explore the impact of these factors individually on two large labeled datasets. Finally, we demonstrate discriminative methods to predict object, scene and attribute terms in descriptions given either known image content, or content estimated by state of the art visual recognition methods. Classification given known image content demonstrates significant performance improvements over baseline predictions. Classification given noisy computer vision estimates also produces smaller, but intriguing, improvements over the baseline.

Acknowledgments: Support of the 2011 JHU-CLSP Summer Workshop Program. T.L.B and K.Y. were supported in part by NSF CAREER #1054133; A.C.B. was partially supported by the Stony Brook University Office of the Vice President for Research; H.D.III and A.G. were partially supported by NSF Award IIS-1139909.

References

[1] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010. 3

[2] J. Deng, A. C. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *ECCV*, 2010. 1

[3] J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, K. Stratos, K. Yamaguchi, Y. Choi, H. D. III, A. C. Berg, and T. L. Berg. Detecting visual text. In *NAACL*, 2012. 3

[4] L. Elazary and L. Itti. Interesting objects are visually salient. *Journal of Vision*, 8:1–15, March 2008. 2

[5] H. J. Escalante, C. Hernandez, J. Gonzalez, A. Lopez, M. Montes, E. Morales, L. E. Sucar, and M. Grubinger. The segmented and annotated iapr tc-12 benchmark. In *CVIU*, 2009. 3

[6] M. Everingham, L. V. Gool, C. Williams, and A. Zisserman. Pascal visual object classes challenge results. Technical report, 2005. unpublished manuscript circulated on the web, URL is http://www.pascal-network.org/challenges/VOC/voc/index.html. 1, 3

[7] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 3

[8] A. Farhadi, M. Hejrati, A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences for images. In *ECCV*, 2010. 3

[9] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4. http://people.cs.uchicago.edu/ pff/latent-release4/. 7

[10] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9), Sep 2010. 1, 3

[11] G. Griffin, H. AD, and P. P. The caltech-256. In *Caltech Technical Report*, 2006. 1

[12] M. Grubinger, P. D. Clough, H. Mller, and T. Deselaers. The iapr benchmark: A new evaluation resource for visual information systems. In *LRE*, 2006. 3

[13] A. Gupta and L. S. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*, 2008. 3

[14] S. J. Hwang and K. Grauman. Accounting for the relative importance of objects in image retrieval. In *BMVC*, 2010. 2

[15] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. In *CVPR*, 2011. 3, 7

[16] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale n-grams. In *CoNLL*, 2011. 3

[17] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang. Large-scale image classification: Fast feature extraction and svm training. In *CVPR*, 2011. 1

[18] A. Oliva and A. Torralba. Building the gist of a scene: the role of global image features in recognition. *Visual Perception, Progress in Brain Research*, pages 23–36, 2006. 3

[19] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 3

[20] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *NAACL HLT Workshop*, 2010. 1, 3

[21] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 77:157–173, may 2008. 2, 3

[22] M. Spain and P. Perona. Some objects are more equal than others: measuring and predicting importance. In *ECCV*, 2008. 2, 4

[23] M. Spain and P. Perona. Measuring and predicting object importance. *IJCV*, 2010. 2, 4

[24] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *ICCV*, 2007. 1

[25] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *CHI*, pages 319–326, 2004. 2

[26] G. Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *ICCV*, 2009. 3

[27] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *ACL*, 1994. 4

[28] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 3, 7