Toward Building Computational Models for Detecting and Reasoning about Abnormalities in Images

Babak Saleh BABAKS@CS.RUTGERS.EDU
Department of Computer Science, Rutgers University, NJ, USA

Ahmed Elgammal ELGAMMAL@CS.RUTGERS.EDU
Department of Computer Science, Rutgers University, NJ, USA

Jacob Feldman JACOB@RUCCS.RUTGERS.EDU
Department of Psychology, Rutgers University, NJ, USA

Ali Farhadi ALI@CS.UW.EDU
Department of Computer Science, University of Washington, WA, USA

Spotting abnormal images and reasoning about what makes them look strange is a great capability of human visual system. However, this challenging problem has been understudied in the field of computer vision and machine learning. We study various types of atypicalities[1] in images by proposing a new dataset of abnormal images, extracting a list of reasons of atypicality, enumerating distinct modes or types of abnormal images, and deriving computational models motivated by human-level reasoning. We derive probabilistic models to learn typicality of objects and images, and infer abnormality as meaningful deviations from this typicality models.

1. Introduction

Humans begin to form categories and abstractions at an early age. The mechanisms underlying human category formation are the subject of many competing accounts, including those based on prototypes, exemplars, density estimation, and Bayesian inference. But all modern models agree that human category representations involve subjective variations in the typicality or probability of objects within categories. For example, bird category includes both highly typical examples such as robins, as well as extremely atypical examples like penguins and ostriches, which while belonging to the category seem like subjectively "abnormal" examples. Visual images can seem abnormal, where they exhibit features that depart in some way from what is typical.

There are several issues and concerns in abnormality detection. *First*, researchers are not in an agreement about what is a typical sample of a category and what makes humans distinguish typical instances from atypical ones. The definition of abnormality in the visual space is even more complex. For example, there is no general rule as what is a typical car. Even if there were such a rule, it might vary across people and categories.

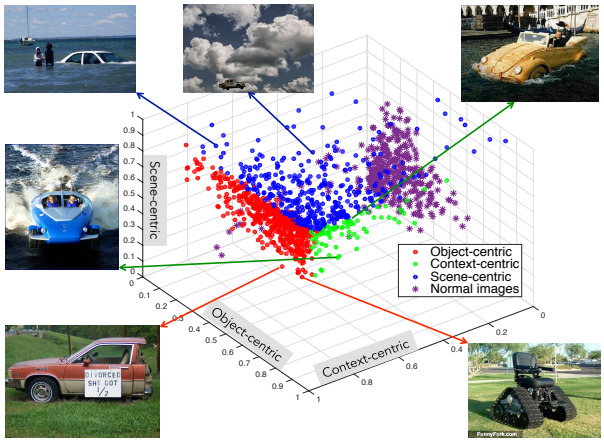

Figure 1. Projection of images based on how abnormal/atypical they look like, where they are plotted based on scores form our computational models of abnormality. While normal images (purple) fall close to the center of coordinate, we can find abnormal images by moving along three axes of abnormality (scene, object, and context). Points (images) are colored-coded based on the most important reason of abnormality that they present.

Second, abnormality in images is a complex notion that happens because of a diverse set of reasons that can be related to shape, texture, color, context, pose, location or even a combination of them (Figure 1)

Third, there is a gradual transition from typical to atypical instances, so simple discriminative boundary learning between typical and atypical instances does not seem ap-

[1] We will use typicality/atypicality when referring to objects, scenes and context, while we will use normality/abnormality when referring to images. However, at some points we use these words interchangeably.

Category name	"Expanded abnormality list" of reasons from Amazon Mechanical Turk Experiment
1 Scene centric	Strange event happening in the scene(21); Strange scene(20)
2 Context centric	Atypical object size(19); Strange location of the object(18); Atypical object pose(17); Weird combination of objects and scene(16)
3 Object centric	Unexpected part(7); Weird shaped part(6); Misplaced part(5); Missing part(4); Body posture(14); Mixture of object classes(13); Unnameable shape(12); Object is not complete(3); Unknown object(15); Object in the shape of another object(2); Atypical pattern(10), Weird color(9), Strange material(11), Weird texture(8); Strange object contour(1)

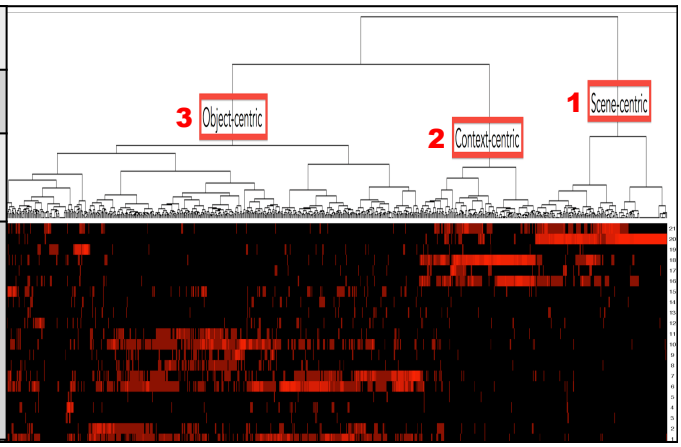

Figure 2. Right: Agglomerative clustering of abnormal images (columns) based on the human subject responses for each abnormality reason (rows). The dendrogram on top shows how the abnormal images in our dataset can be grouped to make three clusters, reflecting three major latent categories of abnormality. Each cluster corresponds to a specific list of abnormality reasons. Details of these three categories of abnormality, and fine-grained reasons of abnormality can be found in the table on the left.

propriate. Figure 1 elaborates this concept where images are plotted based on how abnormal they look.

Fourth, with the limited number of abnormal images it is hard to be comprehensive with all aspects of abnormality. This suggests that computational models for identifying abnormalities should not use samples of abnormal images. This is also aligned with how humans are capable of recognizing abnormal images while only observing typical samples.

2. Computational Models of Abnormality

We made the dataset of "1001 Abnormal Images" by downloading images from the web, where we looked for strange, weird, atypical or abnormal images. We conducted a two-phase human-subject experiment to determine a typology of images judged abnormal by human observers and collect data that facilitates discovery of a taxonomy of atypicality. In the first phase, subjects took on-site tests to describe abnormal images in their own words. We compiled a fine-grained list of abnormality reasons (Listed in the left side of Figure 2). In the second phase we conducted a large-scale human subject experiment on Amazon Mechanical Turk, where 60 annotators rated abnormal images based on 21 fine-grained reasons of abnormality. Analysis of the data lead us to a coarse taxonomy of three reasons for abnormality: object-centric, scene-centric, and contextual (Saleh et al., 2016).

We propose a Bayesian generative model for typical scenes and objects, depicted in Figure 3. This model formulates the relation between objects, context and other information in the scene that is not captured by objects or the context (e.g. scene characteristics such as Sunny or Crowded). This is a model of typicality, and atypicality/abnormality is detected as a deviation from typicality. Hence, this model is trained using only typical images and relies on visual attributes and categories of both objects and scenes.

Visual attributes have been studied extensively in recognition (Lad & Parikh, 2014; Parikh & Grauman, 2011). In contrast to low-level visual features (e.g. HOG, SIFT), attributes represent a valuable intermediate semantic representation of images that are human understandable (nameable). Example attributes can be "Open area", "Sunny weather" for scenes and "wooden" or "spotty" for objects. Attributes are powerful tools for judging about abnormality. For example, the object-centric model of (Saleh et al., 2013) mainly used attribute classifiers to reason about abnormality. However, the response of an attribute classifier is noisy and uncertain. As a result, we categorize the object based on low-level visual features apart from its attributes scores. Later, out model at the level of the object focuses on deviations between categories of the objects and its meaningful visual characteristics (attributes). In short, if low-level features predict an object to be a car, while attribute responses do not provide evidence for a car, that is an indication of abnormality.

As a similar argument stands at the level of scenes, we model the typicality of low-level visual features (F) and attributes (A) for both objects (O) and scenes (S). Figure 3 shows that assuming we observe a normal image I, any distribution over scene category S imposes a distribution over the categories of objects O that are present. This procedure holds for all K objects in the image (left plate is repeated K times). Each object category imposes a distribution over object's low-level features F^o and attributes A^o. Similarly, scene categories impose a distribution over

scene's low-level features F^s and attributes A^s. However, extracted visual features for scenes are different from ones extracted for objects. We define two disjoint sets of attributes for objects ($A^o = \{A_i^o\}_1^n$) and attributes for scenes ($A^s = \{A_i^s\}_1^m$).

Learning the model involves learning the conditional distribution of object-attribute, given object categories ($\{P(A_i^o|O_k), i = 1\cdots n, k = 1\cdots V\}$), and scene-attribute conditional probability distribution given scene categories ($\{P(A_i^s|S_j), i = 1\cdots m, j = 1\cdots J\}$), where each of these distributions is modeled as a Gaussian. We also learn probabilities of object categories given scene categories ($\{P(O_k|S_j), k = 1\cdots V, j = 1\cdots J\}$) , where V and J are number of object and scene categories.

2.1. Measuring Abnormality of Images

Scene-centric Abnormality Score: For any scene category, some visual attributes are more relevant (expected). This is what we call relevance of i^{th} scene attribute for the j^{th} scene category, denoted by $\Omega(A_i^s, S_j)$ [2]. We compute this term by calculating the reciprocal of the entropy of the scene-level attributes for a given scene category $\Omega(A_i^s, S_j) = 1/H(A_i^s|S_j)$ over normal images. This relevance term does not depend on the test image.

For a given image, applying scene classifiers produce a distribution over scene categories. Assuming a scene category, we compute the information content in each scene-attribute classifier response ($I(A_i^s|S_j) = -logP(A_i^s|S_j)$). This information content is a measure of the surprise by observing an attribute for a given scene class. Since attribute classifiers are noisy, depending on the concept that they are modeling, we need to attenuate the surprise score of a given attribute by how accurate is the attribute classifier. We denote this term by $\Upsilon(A_i^s)$, which measures the accuracy of the i^{th} scene attribute classifier on normal images. Therefore the scene surprise score (*Surprise*$_S$) is computed by taking the expectation given $P(S_j)$ as following:

$$\sum_j P(S_j)[\sum_i I(A_i^s|S_j)\Upsilon(A_i^s)\Omega(A_i^s, S_j)] \qquad (1)$$

Context-centric Abnormality Score: An image looks abnormal due to its atypical context if one of the following happens: first, an unexpected occurrence of object(s) in a given scene. (e.g. elephant in the room); second, strange locations of objects in the scene (e.g. a car on top of the house); or inappropriate relative size of the object. We propose Eq. 2 to measure the context-centric surprise (Surprise$_C$) of an image based on aforementioned reasons:

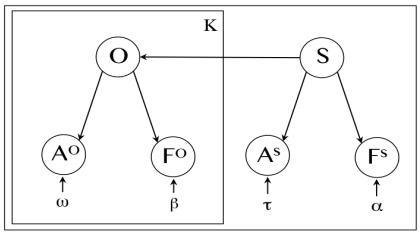

Figure 3. Graphical Model of Normal Images

$$\sum_k \sum_j \Lambda(O_k)[\hat{I}(O_k|S_j) + I(L_k|O_k)]. \qquad (2)$$

The term $\hat{I}(O, S)$ measures the amount of surprise stemming form the co-occurrence of the objects in the scene (Eq. 3). We measure the surprise associated with each object classes appearing in the scene S_j by computing the information content of each combination of scene categories and object classes, $I(O_k|S_j)$, modulated by the probability of the object and scene categories.

$$\hat{I}(O_k|S_j) = P(S_j)P(O_k)I(O_k|S_j). \qquad (3)$$

On the grounds that we use a distribution as the output of classifiers rather than a single class confidence, we do not need to involve the accuracy of neither the object classifier nor the scene classifier to tackle the uncertainty output.

The term $I(L_k|O_k)$ measures how much surprising is the location of the object k in the image. Assuming we know category of the object (O_k), we expect to see it in certain locations in the image. By considering one object category at a time, we learn a distribution of possible locations for the object in normal images and use it to compute the information content of the object location in a test image.

Finally we aggregate the co-occurrence and location term and modulate the score by multiplying it with $\Lambda(O_k)$, which stands for the importance of the size of the object relative to the whole image in judging the context atypicality. If the object of interest is tiny or huge in the image, the contextual surprise should be modulated down. To model $\Lambda(O)$ for each object category(O) we learn the distribution of its relative size by considering the normal images with typical context and for the test image compute its probability based on this distribution.

Object-centric Abnormality: For *Surprise*$_O$ we check if the objects in the image look typical or not independently. We assume that we take the object out of the scene and measure how abnormal it is based on its predicted object class and visual attributes. This term is in part similar to work of Saleh et al (Saleh et al., 2013). However, we are different from their work as we classify the objects

[2]For simplicity, we slightly abuse the notation and use A_i^s to denote both the i^{th} attribute, and the i^{th} attribute classifier response for scene attributes. The same holds for object attributes as well.

Experiment Number	Method	Accuracy	Training images		Testing images	
			Normal	Abnormal	Normal	Abnormal
I	Object-centric baseline (Saleh et al., 2013)	0.9125	Pascal	Not Used	Pascal	Dataset of (Saleh et al., 2013)
	Our Model - Object-centric	**0.9311**	Pascal	Not Used	Pascal	Dataset of (Saleh et al., 2013)
II	Context-centric baseline (Park et al., 2012)	0.8518	SUN	Not Used	SUN	Subset of (Park et al., 2012)-without human
	Our Model - Context-centric	**0.8943**	Pascal	Not Used	SUN	Subset of (Park et al., 2012)-without human
III	One Class SVM - based on Attributes	0.5361	Pascal	Not Used	Pascal	Our dataset
	Two Class SVM - based on Attributes	0.7855	Pascal	Our dataset	Pascal	Our dataset
	One class SVM - based on Deep features (fc6)	0.5969	Pascal	Not Used	Pascal	Our dataset
	Two class SVM - based on Deep features (fc6)	0.8524	Pascal	Our dataset	Pascal	Our dataset
IV	Our Model - No Object-centric score	0.8004	Pascal	Not Used	Pascal	Our dataset
	Our Model - No Context-centric score	0.8863	Pascal	Not Used	Pascal	Our dataset
	Our Model - No Scene-centric score	0.8635	Pascal	Not Used	Pascal	Our dataset
	Our Model - All three reasons	**0.8914**	Pascal	Not Used	Pascal	Our dataset

Table 1. Evaluating the performance (AUC) of different methods for classifying normal images vs. abnormal images.

based on low-level visual features F^o rather than visual attributes A^o. We formulate the object-centric surprise score (Surprise$_O$) as:

$$\sum_k P(O_k) * (\sum_i I(A_i^o|O_k) * \Upsilon(A_i^o) * \Omega(A_i^o, O_k)) \quad (4)$$

Where $P(O_k)$ is the distribution over object categories obtained from low-level visual features. $I(A_i^o|O_k) = -log(P(A_i^o|O_k))$ denotes the amount of the surprise by observing the response of the i-th attribute classifier, given class O_k. Similar to scene-centric surprise score, $\Upsilon(A_i^o)$ adjusts the weights of visual attributes based on how reliable one attribute performs on normal images. $\Omega(A_i^o, O_k)$ models the relevance of attribute A_i^o to object k, however this is computed based on ground truth annotation rather than the conditional entropy of attributes.

2.2. Parametric Model for Typicality

For the final decision about abnormality of an image we should compare the three surprise scores and pick the maximum as the the most important reason of abnormality. However, there are two issues that prevent us from using the maximum of raw surprise scores. These described surprise scores are based on quantifying the information content, therefore these measures are unbounded (as the probability approaches zero, the surprise approaches infinity). The other issue is that these surprise scores are not comparable since the information content in each of them are modulated differently. As a result it is hard to compare the values of *Surprise*$_O$, *Surprise*$_S$, and *Surprise*$_C$ to determine which of these reasons gives rise to the abnormality in the image, if any. To tackle these issues, we propose to model the distribution of the surprise scores for normal images.

Toward this goal, we compare fitting different parametric models to the empirical distributions of three surprise scores, computed over normal images. For model selection we consider simplicity of the distribution, as well as how well it fits the empirical data based on Akaike Information Criterion (AIC) (Akaike, 1974). We are interested in simpler distributions, because of their better generalization and the ability to derive simpler (sometime closed from) CDFs. Our experiments show that independent of the reason of abnormality, surprise scores follow exponential family of distributions. We pick *"Inverse Gaussian"* distribution as the underlying distribution. Due to limited space, we put more analysis in the supplementary material. Given these probabilistic models, we can compute the probability of observing a given surprise score instead of the raw surprise scores. Then we can classify the reason of abnormality in an image by comparing the CDFs of the parametric models, i.e.,

$$\underset{o,s,c}{\text{argmax}}(\phi_o(Surprise_O), \phi_s(Surprise_S), \phi_c(Surprise_C)) \quad (5)$$

Where $\phi_o(\cdot)$,$\phi_s(\cdot)$,$\phi_c(\cdot)$ are the inverse Gaussian CDFs for the object, scene, and context -centric parametric surprise models respectively. Parameters of each model are estimated only from the normal training data.

Experimental Results Experimental Results Table. 1 represents the performance of our proposed models for the task of finding ab- normal images. First two boxes (first four rows) shows that our proposed object-centric and context-centric models outperform state-of-the-art. In the third box (Box III) we conducted experiments with vanilla features of AlexNet (Krizhevsky et al., 2012) for the task of abnormality classification. We also evaluated the performance of two-way SVM classifiers for this challenging task. Interestingly, our final model without training on abnormal images can beat these two-way classifiers. In the last box (Box IV), we conducted an ablation experiment to investigate the importance of three components of our final model: Object, context, and scene centric. Results show that our model benefit from aggregating all these scores. However, object-centric abnormality plays the most important part of our final model.

Acknowledgment: This research was supported by NSF award IIS-1218872.

References

Akaike, Hirotugu. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems*, 2012.

Lad, Shrenik and Parikh, Devi. Interactively guiding semi-supervised clustering via attribute-based explanations. In *European Conference on Computer Vision (ECCV)*, 2014.

Parikh, Devi and Grauman, Kristen. Relative attributes. In *International Conference on Computer Vision*, 2011.

Park, Sangdon, Kim, Wonsik, and Lee, Kyoung Mu. Abnormal object detection by canonical scene-based contextual model. In *European Conference on Computer Vision (ECCV)*, 2012.

Saleh, Babak, Farhadi, Ali, and Elgammal, Ahmed. Object-centric anomaly detection by attribute-based reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013.

Saleh, Babak, Elgammal, Ahmed, Feldman, Jacob, and Farhadi, Ali. Toward a taxonomy and computational models of abnormalities in images. In *The Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, 2016.