Scene Grammar in Human and Machine Recognition of Objects and Scenes

Akram Bayat
University of Massachusetts Boston
Boston, MA, USA
akram@cs.umb.edu

Do Hyong Koh
University of Massachusetts Boston
Boston, MA, USA
dohyong.koh001@umb.edu

Anubhaw Kumar Nand
University of Massachusetts Boston
Boston, MA, USA
Anubhaw.Nand001@umb.edu

Marta Pereira
University of Massachusetts Boston
Boston, MA, USA
pereira.m@usp.br

Marc Pomplun
University of Massachusetts Boston
Boston, MA, USA
marc@cs.umb.edu

Abstract

In this paper, we study the effects of violating the high level scene syntactic and semantic rules on human eye-movement behavior and deep neural scene and object recognition networks. An eye-movement experimental study was conducted with twenty human subjects to view scenes from the SCEGRAM image database and determine whether there is an inconsistent object or not. We examine the contribution of multiple types of features that influence eye movements while searching for an inconsistent object in a scene (e.g., size and location of an object) by evaluating the consistency prediction power of the trained classifiers on fixation features. The results of the eye movement analysis and inconsistency prediction reveal that: 1) inconsistent objects are fixated significantly more than consistent objects in a scene, 2) the distribution of fixations is the main factor that is influenced by the inconsistency condition of a scene which is reflected in the ground truth fixation maps. It is also observed that the performance of deep object and scene recognition networks drops due to the violations of scene grammar. The class-specific visual saliency maps are created from the high-level representation of the convolutional layers of a deep network during the scene and object recognition process. We discuss whether the scene inconsistencies are represented in those saliency maps by evaluating their prediction powers using multiple well-known metrics including AUC, SIM, and KL. The results suggest that an inconsistent object in a scene causes significant variations in the prediction power of saliency maps.

1. Introduction

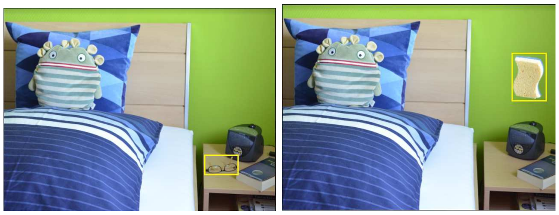
Natural scenes follow a set of semantic and syntactic rules that are initiated from the regulations that are recognized and generally accepted throughout our life. For example, an object can fit to some specific temporal or special semantic contexts (e.g., bed is usually found in the bedroom and not in the bathroom). Furthermore, syntactic rules in any natural scene are stablished based on physics. For instance, objects do not hover in air because of the gravity since they need a surface to rest or a place to be hung. These semantic and syntactic rules construct a scene grammar that define the relations between objects and scenes [8]. Studying the effects of violations from these semantic and syntactic rules in natural scenes on the human visual system is useful for getting insight into the underlying high-level cognitive mechanisms that can help to improve the performance of the visual systems in detecting inconsistency in natural scenes. Inconsistency detection is important due to its wide applications in surveillance systems, driving assistance, and virtual reality scene design.

From the perspective of human perception and behavior, scene grammar facilitates identification of objects within a scene which reduces the computational load of perceptual processes [1, 2]. It has been shown in a broad range of eye movement research that objects inconsistent with the global identity of a scene are processed differently than other objects (e.g., different fixation durations) [7]. In addition, studies on scene perception have shown that semantic relations between objects of a scene which includes information about detailed interaction among them is processed concurrently with object identification in humans [5], which is evidence for the importance of scene grammar in the processing of a scene. However, in deep convolutional neural networks for object and scene recognition, the process of learning meaningful contextual information in terms of consistent relation between objects of a scene and their arrangements has not collected considerable attention. Taking advantage of the deep Convolutional Neural Networks (CNNs) with multiple processing layers and training on huge databases with millions of samples and instances have considerably improved the state-of-the-art in visual scene and object recognition systems [13, 4]. In this work, we aim to test whether deeper layers of a deep convolutional neural network can learn high-level representations that describe the typical relation between objects or object and scene representative of a scene grammar. Furthermore, what are the effects of violations of scene grammar in prediction performance of deep object and scene recognition networks? These investigations can assist in better understanding of the high-level representations of deep neural networks for establishing syntactic and semantic rules in deep CNNs during the learning process.

The present work is the first work to the best of our knowledge that addresses the effects of inconsistency in a scene in human eye-movement behavior and in biologically inspired visual systems such as deep CNNs. We comprehensively analyze human eye movements on the SCEGRAM database [14] which is a well-controlled image data in multiple states of violating the scene grammar. The extracted fixation patterns are employed to evaluate whether the last convolutional layer of a deep CNN (e.g., AlexNet) can learn high-level cognitive factors indicators of the sematic relations between objects and a scene. Furthermore, we analyze the influence of factors such as object location and size in a scene in modulating eye-movement behavior while viewing a scene with an inconsistent object, similar to influence of text characteristics (e.g., grammar, topic, and layout) on eye-movement behavior [3].

2. Eye-movement behavior and scene grammar

We analyze eye movements during viewing and perception of the SCEGRAM database images. Eye movements of 20 participants are recorded when they search to find inconsistent objects that are not embedded into the semantics of the natural scenes. The SCEGRAM database consists of 62 real-world indoor scenes in multiple consistency conditions, however, we use the 62 scenes that are consistent and extremely inconsistent (totally 124 images). Figure 1 illustrates a sample scene in consistent control condition (e.g., eyeglass on table) and in extreme semantic or syntactic violations (e.g., kitchen sponge floating in bedroom).



(a) A consistent object (b) An inconsistent object

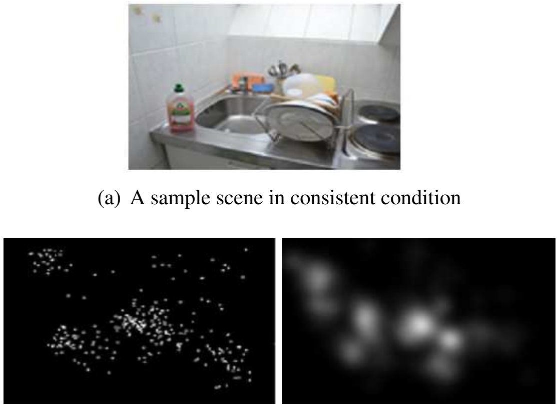
Figure 1. Examples of a scene in (a) consistent condition (b) inconsistent condition

2.1. Eye movement experimental design

Twenty participants were recruited from the Computer Science Department of the University of Massachusetts Boston to participate in the experiment with financial compensation. The average age of subjects was around 26 and half of them were wearing glasses or contacts during the experiment. No one had vision abnormality. Subjects eye movements were recorded by a video-based EyeLink 1000 system with desktop-mounted setup and 1000Hz sampling frequency. A chin-rest was used to increase the tracking stability. A total of 124 images from the SCEGRAM image data base were taken as stimuli, 62 of which were consistent with the semantic context of the scene while the remaining ones were not. Images were displayed on a 27-inch LCD monitor with 1400x1050 pixels resolution. For each subject, images were shuffled, and each image was displayed for 5 seconds. A fixation dot was displayed on the center of the screen for 3 seconds between images. During that moment, subjects were asked to enter their decision regarding the consistency of the preceding image. An auditory feedback was provided to subjects after the keyboard input. A wrong answer triggered a low pitch sound and a high pitch sound was played for the correct answer. After the calibration procedure for the eye tracker, the experimenter gave a brief tutorial about the experiment to each subject who then practiced the task on several images. The experiment lasted approximately 15 minutes for each subject and the eye movements, final score and response time were recorded.

2.2. Eye movements in object-scene semantics perception

It has been shown in the previous eye movement research that scene semantics can regulate the eye-movement behavior during free viewing or task-driven viewing in which the subjects task is to decide about the semantic congruity of a scene. For example, longer and more fixations are observed for inconsistent objects in the scene [10, 16]. These findings



(a) A sample scene in consistent condition

(b) The binary fixation map (c) The continous fixation map

Figure 2. (a) Examples of a scene and (b) the binary fixation map and (c) the continuous fixation map.

are used to show that scene-object semantics are obligatorily processed during natural scene and object recognition tasks.

In this section, our goal is to explore how the semantic and syntactic inconsistencies of scenes in the SCEGRAM database influence on subjects eye-movement behavior. Then, the fixation maps are constructed for evaluation of saliency maps. The binary and continuous (the blurred binary fixation map) human fixation maps are created (Figure 2). The Gaussian low-pass filter with cut off frequency of 8 cycles per image (approximately equivalent to 1 degree of visual angle) is used to create continuous fixation maps

Different eye movement measures are analyzed. Normality of the distribution is verified using the Kolmogorov-Smirnoff and the Shapiro-Wilk tests. Parametric test, t-test for repeated measures, is used to compare the following eye movement measures in both consistent and inconsistent scenes: number of fixations on target object (consistent or inconsistent objects), total fixation count, ratio of fixations, and average fixation duration.

Number of fixations on target object: We could observe that on average, participants fixate significantly (t(61)=88.50, $p<0.05$) more on the target object in inconsistent scenes (M=152.92, Standard Error Mean (SE)=9.03) than in consistent scenes (M= 64.42, SE=5.80). This result proves that fixation density in a scene varies according to the semantic features of the image, since the gaze tends to be fixated more on semantically inconsistent objects.

Total fixation count: This measure is defined as the total number of fixations during the entire scene viewing time. It was observed that subjects fixated more in consistent scenes (M=568.03, SE=3.76) than inconsistent scenes (M=484.74, SE=5.25), and this difference was statistically significant (t(61)=83.29, $p<0.05$).

Average fixation duration: This measure is computed by dividing the total fixation time by the total fixation count. The average fixation duration is higher in inconsistent scenes (M=303.71, SE=3.46) than in consistent scenes (M=260.84, SE=2.33). This means that, on average, subjects made significantly (t(61)=-42.87, $p<0.05$) longer fixations when looking at inconsistent scenes than when looking at consistent scenes.

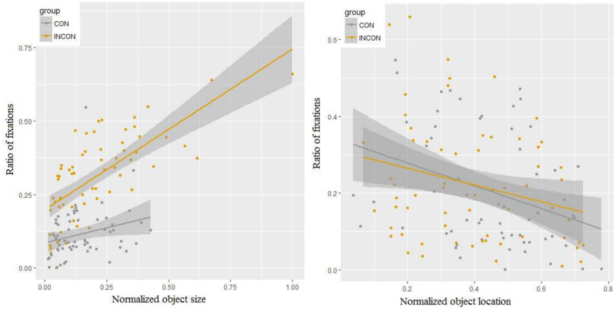
Ratio of fixations: is the number of fixations on the target object in each image over the total number of fixations on the image. The ratio of fixations in inconsistent scenes is significantly (r= 0.28, t(61)=3.91, p=3.98E-16) higher (M= 0.315 SE= 0.021) than in consistent scenes (M=0.113, SE=0.007). Moreover, the ratio of fixations on target objects (consistent or inconsistent) to the rest of the image by average varies between 11.3% in CON to 31.5% in INCON scenes.

Both total fixation count and average fixation duration seem to be an accurate measures of object processing during scene viewing, as certified by previous studies [10]. In fact, not only do subjects tend to look more at the inconsistent object but they also spend, on average, more time per fixation on that same object. It is also observed that the ratio of fixations is another eye-movement variable that is influenced by the scene semantic contiguity.

Taken together, these results reveal that eye-movement patterns during scene viewing are driven by high-level semantic content and specific changes in this consistency are accurately detected by eye-movement behavior. However, there are some other cues that seem to influence eye movements while viewing an inconsistent scene: 1) size of the inconsistent object, 2) spatial location of the inconsistent object. We investigate this by evaluating the variation of the ratio of fixations by each of these cues.

Object size: 120 out of 124 images (96.8%) in the SCEGRAM dataset take up less than 10% of the image size. We compute the normalized object size by normalizing the ratio of the object area to the image area. However, there is a moderate correlation (r=0.582) between the ratio of fixations and the normalized object size that shows they are dependent variables. In addition, since the average ratio of the object size to the image size in the INCON scenes (M=4%, SE=0.0012) is significantly (r= 0.54, t(61)=3.91, p=0.013) larger than for consistent objects (M=2.7%, SE=0.0004) in the CON scenes, then, we consider the object size as an influencing factor in the ratio of fixations that we use for classifying consistent and inconsistent objects (Figure 3(a)).

Object location: The normalized spatial distance between the center of an object (in a scene) and the center of that scene is computed as an indicator of object location. There is no significant (r= 0.18, t(61)=3.91, p=0.55) difference in normalized object location between inconsistent scenes (M= 0.425 SE= 0.029) than consistent scenes (M=0.406, SE=0.037). However, as is illustrated in Fig-



(a) (b)

Figure 3. Variations of the ratio of fixations in each scene versus (a) the normalized object size and (b) the normalized object location. CON and INCON scenes have been marked in different colors.

ure 3(b), there is a weak negative correlation between the normalized object location and the ratio of fixations on that image.

We refine the SCEGRAM image database in order to reduce the influence of object size and location on the ratio of fixations as an important feature that is modulated due to semantic content and specific changes in consistency condition of a scene.

The image database is refined based on the object size in which objects whose sizes is greater than 5% of the image size are filtered. Ninety-five (43 INCON images) out of 124 images satisfy this condition which are considered as Database I.

Similar to Object size filtering, the image database is refined based on the normalized object location in which objects whose normalized location is less than 0.2 are filtered out. We call this feature as normalized filtered object size. That is, object close to the center of the image is removed. One hundred and four (55 INCON images) out of 124 images meet this constraint. We call the remaining images as Database II.

Database III with 78 images (38 INCON) contains images that are remained after the original database is filtered based on size (Database I) and location (Database II) simultaneously. We construct a binary classification model based on the ratio of fixations feature to distinguish inconsistent scenes from consistent scenes in all created subsets of the SCEGRAM database. In fact, knowledge of the prediction powers of consistency versus inconsistency of a classifier on various created databases reveals the role of various factors (e.g., the ratio of fixations, object size, and object location) in directing the visual attention in an inconsistent scene.

2.3. Classification of CON and INCON scenes

Several classifiers were tested in order to evaluate the prediction power of the individual ratio of fixations on inconsistency and the results of the best classifier for each database are summarized in Table1. We train a classifier based on the ratio of fixations as a single feature on the SCEGRAM, Database I, Database II, and Database III.

Table 1. Classification of inconsistent versus consistent scenes for selecting the determinant factors on eye movemet in multiple combinations of features

Database	Number of images	Classifier	Accuracy	Precision	Recall
SCEGRAM	124 (62 INCON)	Logistic	86.29%	0.865	0.863
Database I	95 (43 INCON)	Logistic	83.15%	0.839	0.832
Database II	104 (55 INCON)	Logistic	84.61%	0.849	0.846
Database III	78 (38 INCON)	Logistic	**80.76%**	0.809	0.808

In terms of classification accuracy, the results show that the Logistic classifier outperforms the other classification algorithms in classifying the scenes into CON and INCON scenes. In previous section, we observed that the ratio of fixations is affected by the object size and location. However, the consistency condition is not associated with size and location of an object (e.g., an object can be consistent in one scene and inconsistent in another scene) but object size and location and consistency condition vary in parallel with a third hidden factor, namely direction of attention. This association has been revealed in classification results. By removing the outliers in Database I, Database II and specifically in Database III, we observe that classification results does not drop significantly that shows the ratio of fixations is not biased considerably with the size and locations of the object in SCEGRAM database. Furthermore, it indicates that the ratio of fixations that are explicitly embedded in a fixation map, properly describes the human eye movements during viewing an inconsistent object in a scene.

3. Inconsistency in deep object and scene recognition networks

The availability of large scale labeled datasets like Places [18] or ImageNet [12] have significantly enhanced the performance of deep CNNs for object recognition and scene classification . However, some high-level contextual information like relationships between objects has not been included in the learning process which lead to classification performance becomes lower than human performance. In this section, we evaluate the deep object and scene recognition networks in presence of inconsistencies in a scene.

Two CNN models, AlexNet-ImagNet [11] for object classification and AlexNet-Places205 [17] for scene classification, are employed. AlexNet-ImagNet is the pretrained AlexNet CNN network that was trained on ImageNet dataset with over 10 million images over 1000 object categories. AlexNet-Places205 is the pre-trained AlexNet CNN for scene classification which was trained on the Places data with more than 10 million images on 205 unique scene categories. We evaluatet the results of AlexNet-ImageNet and AlexNet-Places205 on the SCEGRAM Data for object and scene classification, respectively. Then, we disscuss the influence of inconsistency in variations of predictions from consistent scenes to inconsistent scenes.

Scene Classification: Table 2outlines the classification performance of the AlexNet-Places205 network for both consistent and inconsistent scenes. The significant performance drop of about 9% in top-1 error rate is occured due to an inconsistent objects that do not fit the semantics of the scenes. We also observe that the average probability of the predicted classes is decreased by about 22% for majority of the images and increased for the rest of the scenes by about 20%. However, this increase in probability led to only a small performance increase of 3-4%.

Table 2. The influence of inconsistency on scene classification network (AlexNet-Places205), where FR, TR, PI, PD stand for False Recognition,True Recognition, Probability Increase and Probability Decrease, respectively

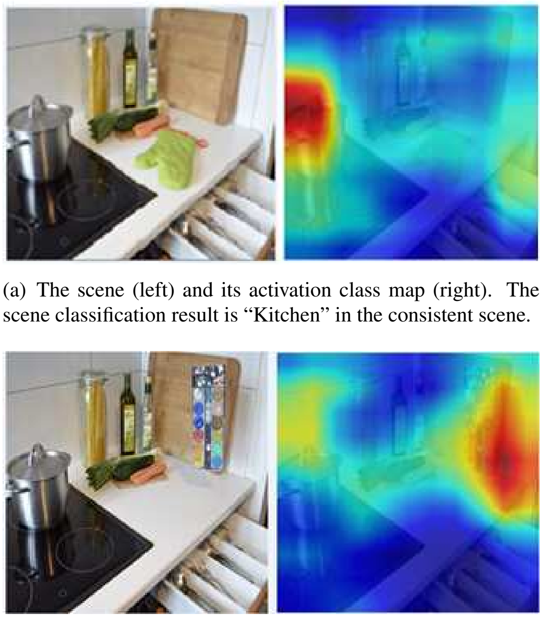
	Consistent scenes	**Inconsistent scenes**
Top-1 error	32.66%	41.53%
Top-1 Decision Changing	16.12% FR	3.2% TR
Top-1 Probability Changes	41.9% PI	58.1% PD

Object Classification: Table 3 lists the object recognition performance of the AlexNet-ImageNet network in both consistent and inconsistent scenes. We observe that there is a significant performance drop of about 13% in top-1 error rate due to inconsistent objects in the scenes. We also find that the probability of the predicted classes is increased by an average of 23% for more than half of the images and is decreased for the rest of the scenes on average by about 25%. However, this increase in probability led to a small performance increase of 4-5%.

Table 3. The influence of inconsistency on object classification network (AlexNet-ImageNet), where FR, TR, PI, PD stand for False Recognition,True Recognition, Probability Increase and Probability Decrease, respectively

	Consistent Scenes	**Inconsistent Scenes**
Top-1 error	54.84%	67.74%
Top-1 Decision Changing	17.74% FR	4.8% TR
Top-1 Probability Changes	53.22% PI	48.38% PD

An examples of scene classification results in consistent and their correspondence inconsistent scenes is illustrated in Figure 4. The heatmaps show the image regions that highly contribute to the identification of the scenes categories. The heatmaps are generated from the activation maps of the fifth convolutional layer (CONV5) of the AlexNet network based on this fact that object detectors to emerge inside the inner layers of the CNN network trained for scene classification [17].



(a) The scene (left) and its activation class map (right). The scene classification result is "Kitchen" in the consistent scene.

(b) The same scene as in (a) in inconsistent condition, the inconsistent object changes the predicted class to "shower" (the prediction score is increased) which leads to an incorrect result.

Figure 4. Examples of a sample scene in two consistency conditions. The corresponding class activation heatmaps show the discriminative regions of each scene that determine the predicted class.

4. High-level saliency and semantic consistency

In this section, we seek to explore the other effects of violation of scene grammar in deep CNN recognition systems other than the reduced performance. First, we investigate the multiple levels of representation in various layers of a deep network to find out whether those representations can detect inconsistency in object-scene relation, however, no supervision is provided for scene-inconsistency learning. Subsequently, we consider the convolutional units of inner layers of a deep CNN network that can generate high-level visual concept (e.g., object) detectors. The Class Activation Maps (CAM) technique [17] is used to generate the activation maps during a particular recognition task. These activation maps can identify important regions of a scene that contribute to the specific class prediction. The weights of the output layer is mapped back onto the last convolutional layer. Using this technique we can localize the class-specific discriminative regions most relevant to the particular category used for scene or object classification. For this purpose, we first train the Alexnet-CAM network on the ImageNet database for the object recognition and on the Places205 database for the scene classification, respectively. The Alexnet-CAM network has the AlexNet architecture in which its fully connected layers are removed and the average global pooling layer is added to the last convolutional layer to preserve the localization ability. The fifth convolutional layer of these networks (Alexnet-CAM-ImageNet and Alexnet-CAM-Place205) are employed for generating the activation maps. Each generated class activation map represents the probability of each corresponding pixel in the image to discriminate regions of images for that specific class category. The generated class-specific activation maps can serve as saliency maps. In the next section, we evaluate to what extent the generated saliency maps are comparable with the human fixation map which are driven by high-level semantic content and specific changes in consistency conditions of a scene.

4.1. Evaluation of the class-specific saliency maps

The influence of scene inconsistency on the class-specific saliency maps based on the high-level representations learned during object or scene classification on the SCEGRAM images can be investigated by evaluating the similarity of those saliency maps with the ground truth fixation maps obtained from eye-movement analysis in two consistency conditions. During eye-movement analysis, we observed that inconsistent objects with semantics of the scenes attract human attention more than consistent objects because they are fixated more than and longer than other regions. Therefore, the similarity level between class-specific saliency map and the fixation map as ground truth indicates the influence of the high-level concept of inconsistency on the deepest convolutional layer of the deep object and scene recognition networks whose are represented in a saliency map. We compute similarity between saliency maps and fixation maps using different evaluation metrics and across two consistency conditions. Table 4 summarizes the results of evaluating class-specific saliency models during object and scene classification on the SCEGRAM dataset. We use the evaluation metrics that are commonly used on the MIT Saliency Benchmark [6] like AUC Judd (area under the ROC curve, Judd version), SIM (similarity, also referred to as histogram intersection), and dissimilarity metrics like KL (Kullback-Leibler divergence). We employ these evaluation metrics since each of them can evaluate the similarity between saliency and fixation maps from different perspective.

We also select a deep convolutional neural network for visual saliency prediction model from MIT saliency benchmarks that hold promising results in various evaluation metrics: SalGAN [15]. The architecture of the SalGAN is based on generative adversarial networks (GANs) [9] fit a deterministic function to generate saliency values from images. The saliency map of any given scene in the SalGAN model provides the probability of each corresponding pixel in that scene to capture human attention.

We aim to compare the evaluation metrics across two scene consistency conditions first and then compare the metrics across three saliency models: class-specific saliency maps extracted through scene classification, class-specific saliency maps extracted through object classification, and SalGAN. The values of ACU-Judd, SIM, NSS, and KL in the inconsistent condition are discussed in the following section.

The AUC-Judd evaluation metric is the most widely used metric for evaluating saliency maps. AUC-Judd values indicate whether the class-specific saliency maps generated during the scene classification and object recognition processes in two consistency conditions show a significant difference in predicting the number of ground truth fixations they capture in successive threshold values. The results of AUC-Judd evaluation for class-specific saliency maps show that the similarity values between the regions of a scene that significantly contribute to the scene or object recognition and the fixation locations in that scenein two consistent and inconsistent scenes do no. AUC-Judd values for two consistency conditions. AUC-Judd values for two consistency conditions are illustrated in Figure 5.

However, the AUC-Judd evaluation on the SalGAN saliency prediction model (one of the best deep saliency models in MIT saliency benchmarks) results in better score due to a higher correspondence with ground truth fixation map. In addition, there is a significant variation between consistent and inconsistent scenes.That is, the saliency maps that are generated from the SalGAN model (Figure 6) can better capture scene inconsistency than class-specific saliency maps for the scene and object recognition system (e.g., AlexNet).

The SIM evaluation metric measures the similarity between saliency maps and continuous fixation distribution. A SIM of one indicates identical distributions and a SIM of zero indicates no overlap. SIM is very sensitive to false negatives (missing values) and penalize them more significantly than false positives. It is observed that values of SIM significantly drop for inconsistent scenes because class-specific saliency maps seem to fail to detect all inconsistent object locations that have higher fixation density than other regions. These missing values lead to lower similarity values.

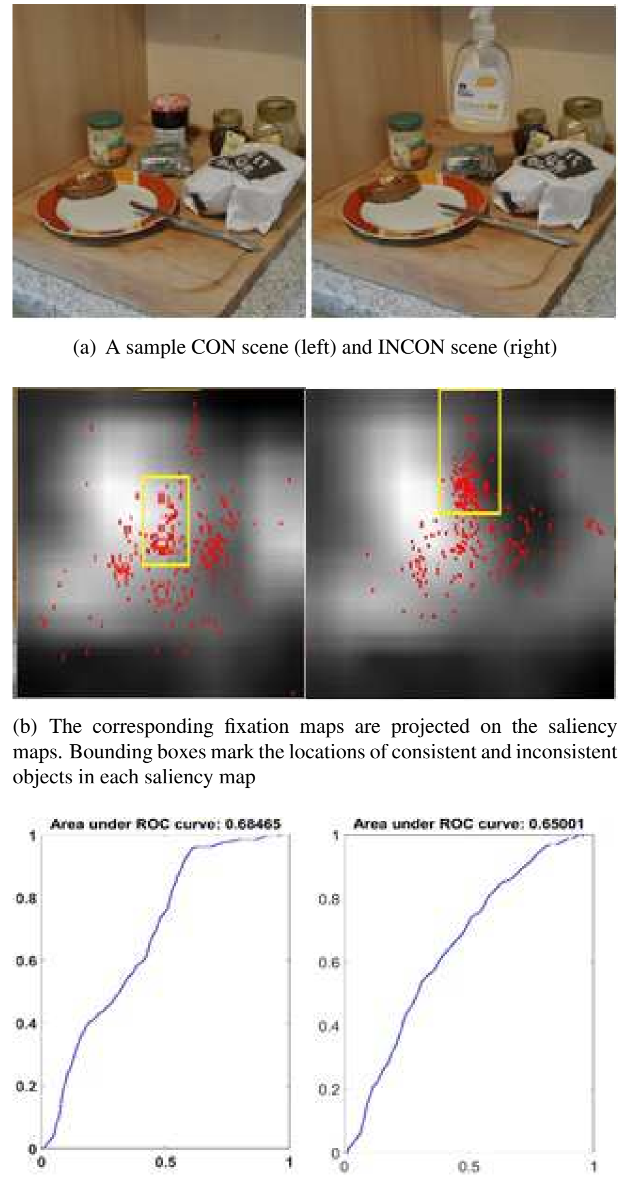
The KL evaluation metric measures the divergence between fixation and saliency maps distributions. A lower score of KL shows the better prediction power of the saliency map. KL is more sensitive to false negatives than the SIM metric. Similar to the SIM metric, inconsistency in a scene affects KL values in all saliency models and causes significant differences.

5. Conclusions

In this work, we studied the influence of the scene inconsistency on human eye-movement behavior and in deep CNNs for object and scene recognition systems. The results of eye-movement analysis on an experiment with twenty

Table 4. The evaluation of saliency maps in presence of inconsistency on the SCEGRAM database

Recognition Process	**Scene-Classification**		**Object-Classification**		**SalGAN**	
Scene consistency condition	Consistent	Inconsistent	Consistent	Inconsistent	Consistent	Inconsistent
Auc-Judd	0.546	0.518	0.665	0.650	0.799	0.822
SIM	0.421	0.351	0.475	0.397	0.628	0.552
KL	1.123	1.462	0.902	1.246	0.669	0.813



(a) A sample CON scene (left) and INCON scene (right)

(b) The corresponding fixation maps are projected on the saliency maps. Bounding boxes mark the locations of consistent and inconsistent objects in each saliency map

(c) The corresponding ROC curves and AUC scores

Figure 5. Evaluation of the various saliency maps in predicting human eye movements in the consistent and inconsistent conditions

participants while task-driven viewing of the SCEGRAM database revealed that human fixations during scene exploration differ significantly in terms of frequency and duration as a result of semantic consistency manipulations that moves their attention towards the inconsistent object. That was, subjects tended to look more often and longer at the inconsistent object than consistent object. In addition, subjects eye movements accurately detected high-level semantic and syntactic scene grammar violations that were reflected in the fixation maps. This was proved this by evaluating the influence of the object size and location on eye-movement behavior and concluding that these factors do not significantly modulate the fixation ratio. We then evalu-

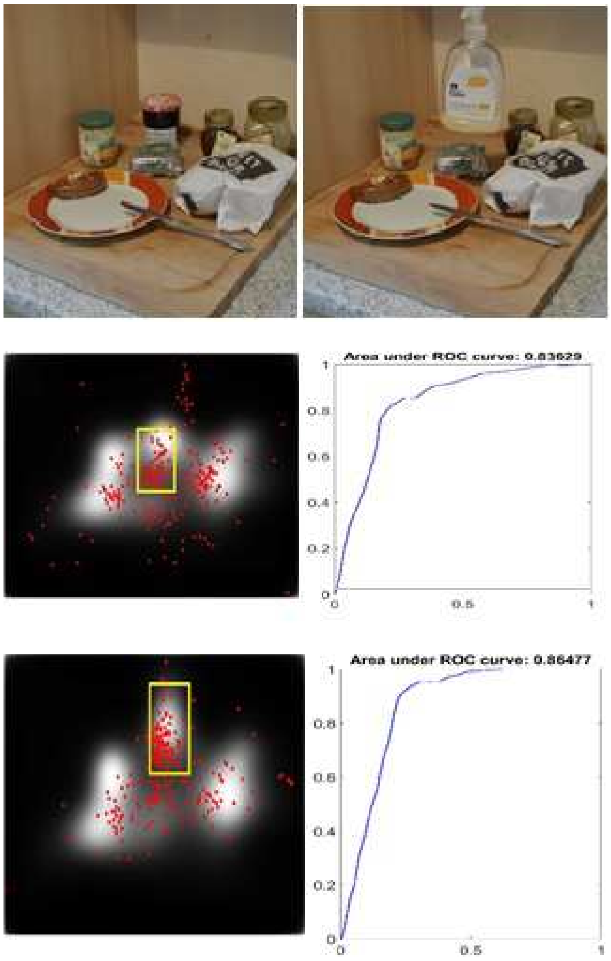


Figure 6. Example of saliency maps obtained by SalGAN models and thier prediction performance by AUC evaluation metric . The SalGAN outperforms the class specific saliency maps in capturing the inconsistent object in the scene.

ated the performance of the deep convolutional neural networks for classifying objects and scenes with inconsistencies in their scenes. The results showed that deep recognition networks perform poorly in detecting the category of objects or scenes when there is an inconsistency in their scene. However, the performance of the object recognition network (AlexNet-ImageNet) was affected more than a scene recognition network (AlexNet-Places205). Subsequently, the class-specific saliency maps were derived from high-level representations of deep object and scene recognition networks (e.g., AlexNet-CAM) in order to analyze the influence of the scene inconsistency in those saliency maps. Taken together, it can be concluded that class-specific saliency maps perform poorly in predicting the human fixation locations in inconsistent scenes compared to consistent scenes.

References

[1] M. Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617, 2004.

[2] M. Bar. The proactive brain: memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1235–1243, 2009.

[3] A. Bayat and M. Pomplun. The influence of text difficulty level and topic on eye-movement behavior and pupil size during reading. In *Signal Processing and Intelligent Systems (ICSPIS), International Conference of*, pages 1–5. IEEE, 2016.

[4] A. Bayat and M. Pomplun. Deriving high-level scene descriptions from deep scene cnn features. In *Image Processing Theory, Tools and Applications (IPTA), 2017 Seventh International Conference on*, pages 1–6. IEEE, 2017.

[5] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology*, 14(2):143–177, 1982.

[6] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark, 2015.

[7] T. H. Cornelissen and M. L.-H. Võ. Stuck on semantics: Processing of irrelevant object-scene inconsistencies modulates ongoing gaze behavior. *Attention, Perception, & Psychophysics*, 79(1):154–168, 2017.

[8] D. Draschkow and M. L.-H. Võ. Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific reports*, 7(1):16471, 2017.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[10] J. M. Henderson, P. A. Weeks Jr, and A. Hollingworth. The effects of semantic consistency on eye movements during complex scene viewing. *Journal of experimental psychology: Human perception and performance*, 25(1):210, 1999.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[13] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[14] S. Öhlschläger and M. L.-H. Võ. Scegram: An image database for semantic and syntactic inconsistencies in scenes. *Behavior research methods*, 49(5):1780–1791, 2017.

[15] J. Pan, C. Canton, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017.

[16] M. L.-H. Võ and J. M. Henderson. Does gravity matter? effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision*, 9(3):24–24, 2009.

[17] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2921–2929. IEEE, 2016.

[18] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.