CONTEXT-BASED IMAGE CONCEPT DETECTION AND ANNOTATION

by

Esfandiar Zolghadr

A Dissertation Submitted to the Faculty of

The College of Engineering and Computer Science

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

Florida Atlantic University

Boca Raton, FL

December 2016

Copyright by Esfandiar Zolghadr 2016

CONTEXT-BASED IMAGE CONCEPT DETECTION AND ANNOTATION

by

Esfandiar Zolghadr

This dissertation was prepared under the direction of the candidate's dissertation advisor, Dr. Borko Furht, Department of Computer and Electrical Engineering and Computer Science, and has been approved by the members of his supervisory committee. It was submitted to the faculty of the College of Engineering and Computer Science and was accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

SUPERVISORY COMMITTEE:

Borivoje Furht, Ph.D.
Dissertation Advisor

Robert Cooper, Ph.D.

Hari Kalva, Ph.D.

Ankur Agarwal, Ph.D.

Nurgun Erdol, Ph.D.
Chair, Department of Computer and
Electrical Engineering and Computer Science

Mohammad Ilyas, Ph.D.
Dean, College of Engineering and
Computer Science

Deborah L. Floyd, Ed.D.
Dean, Graduate College

November 28, 2016
Date

ACKNOWLEDGEMENTS

I wish to express sincere gratitude to my committee members for all of their guidance and support, and special thanks to my advisor, Dr. Borko Furht for his persistence, patience, and encouragement during the writing of this dissertation. I feel extremely privileged for having Dr. Robert Cooper's guidance and instrumental commentaries which were essential in preparation of this manuscript.

ABSTRACT

Author: Esfandiar Zolghadr

Title: Context-based Image Concept Detection and Annotation

Institution: Florida Atlantic University

Thesis Advisor: Dr. Borivoje Furht

Degree: Doctor of Philosophy

Year: 2016

Scene understanding attempts to produce a textual description of visible and latent concepts in an image to describe the real meaning of the scene. Concepts are either objects, events or relations depicted in an image. To recognize concepts, the decision of object detection algorithm must be further enhanced from visual similarity to semantical compatibility. Semantically relevant concepts convey the most consistent meaning of the scene.

Object detectors analyze visual properties (e.g., pixel intensities, texture, color gradient) of sub-regions of an image to identify objects. The initially assigned objects names must be further examined to ensure they are compatible with each other and the scene. By enforcing inter-object dependencies (e.g., co-occurrence, spatial and semantical priors) and object to scene constraints as background information, a concept classifier predicts the most semantically consistent set of names for discovered objects. The additional background information that describes concepts is called context.

In this dissertation, a framework for building context-based concept detection is presented that uses a combination of multiple contextual relationships to refine the result of underlying feature-based object detectors to produce most semantically

compatible concepts.

In addition to the lack of ability to capture semantical dependencies, object detectors suffer from high dimensionality of feature space that impairs them. Variances in the image (i.e., quality, pose, articulation, illumination, and occlusion) can also result in low-quality visual features that impact the accuracy of detected concepts.

The object detectors used to build context-based framework experiments in this study are based on the state-of-the-art generative and discriminative graphical models. The relationships between model variables can be easily described using graphical models and the dependencies and precisely characterized using these representations. The generative context-based implementations are extensions of Latent Dirichlet Allocation, a leading topic modeling approach that is very effective in reduction of the dimensionality of the data. The discriminative context-based approach extends Conditional Random Fields which allows efficient and precise construction of model by specifying and including only cases that are related and influence it.

The dataset used for training and evaluation is MIT SUN397. The result of the experiments shows overall 15% increase in accuracy in annotation and 31% improvement in semantical saliency of the annotated concepts.

CONTEXT-BASED IMAGE CONCEPT DETECTION AND ANNOTATION

TABLES

FIGURES

CHAPTER 1: INTRODUCTION

1.1 Introduction to Scene Understanding

Scene understanding means describing visual properties pictured in a scene to convey the essence of depicted concepts and the real meaning of the scene. The result of such system will be a set of related words that together provide a relatively clear mental picture of the content.

By looking at the image shown in Figure 1 for instance, a human observer immediately gathers the information necessary to categorize the scene. Based on the presence of a car on the foreground and the mountains and the sky in the background, one can describe this scene as "car on the road." More detail study of the scene reveals that it is showing a side view of a brown Honda SUV and there is no passengers or driver, so it is in the parked position. Adding the newly discovered concepts to initial observation results in the description such as "a brown Honda CR-V, parked alongside an open road shown from the driver side." If the location of the image was known, additional valuable information such as the time of the year that this picture was taken could have been inferred based on the color of plants. Other information such as estimation of the age of the car may also be beneficial to someone who is looking for 2016 Honda CR-V pictures. In applications such as video or image retrieval, more detail information translates into better results for queries. The image in Figure 1 can be the result of query about location, time of the year, or specifics about make and model of the car.

Scene understanding also attempts to describe the actions, events, emotions, and more details about the situations and timelines, particularly in videos. For

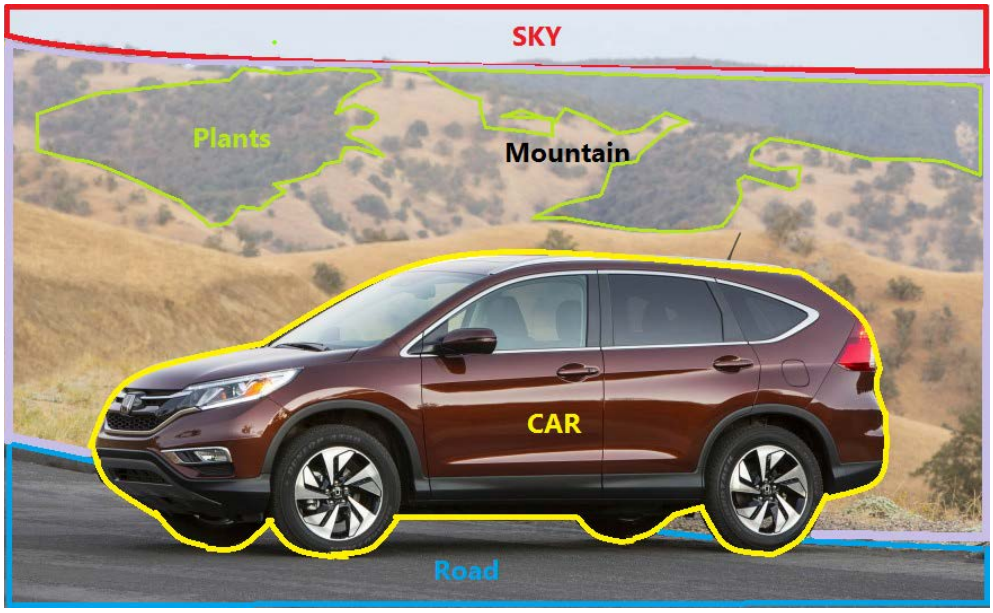


Figure 1. A typical picture of a car on the road.

instance “A man entered the room” or “player number 19 scored a goal in second half” could be expected annotations results of a video clip.

Most concepts involve objects, their conditions, and interactions, so the focus of scene understanding is on the development of high-performance object recognition techniques that meet sensitive requirements of applications like robotics, self-driven car, aviation, and military control system among others.

Concept recognition relies on object detectors to analyze visual properties (e.g., pixel intensities, texture, color gradient) of sub-regions (or window) in an image to provide a list most likely object class labels for every window. Some of more popular approaches widely used in designing classifiers are:

- Pixel-based Detectors

In computer vision, pixel-based methods model pixel intensities. A pixel may have exclusive membership to one class or mixtures of classes with a proportional probability of membership to each class. The latter is often called sub-pixel based

a) Original image b) Some of the strongest SURF features

Figure 2. Key-points found by SURF

approach. When pixels are analyzed at isolation valuable information about the co-occurrence of neighboring pixels inside object boundaries are not considered. One way to mitigate this limitation is by using adjacent pixel similarities to aggregate pixels into groups called superpixels. Using super-pixels allows integration of vital relations such as scale and location of relative spatial alignment for the whole-part type of classification.

In spectral imaging, the electromagnetic spectrum is used to classify each pixel. Hyperspectral image unmixing is one of the branches of spectral imaging widely employed in applications such as remote sensing, signal and image processing (Bioucas-Dias et al., 2012). Models in this field are crafted based on the assumption that the spectral measurement of each pixel which is a function of wavelength is a non-negative linear combination of the spectral signature of some pure materials called *endmembers* and fractional abundance map of the endmembers within a simplex constraint.

- Shape-based Detectors

Shape-based approaches characterize objects based on their shape geometry, contour or texture. These methods are often successful in applications such as

automated object tracking in the video with a limited number of areas of interest (Chiverton, Xie, & Mirmehdi, 2012; Lu & Little, 2006). Shape-based methods are complex particularly in applications that deal with clutter and many occluded objects.

- Color-based Detectors

Color-based descriptors are easier to acquire from image color histogram and can be used to determine a quick set of candidate objects in a picture (Han, Ye, & Jiao, 2008; Khan, Gu, & Backhouse, 2011).

- Feature-based Detectors

Feature-based approaches try to solve recognition problem by transforming images to features and then perform classification on the set of features. State-of-the-art classifiers are formulated either as a probabilistic (e.g., Gaussian Mixed Model), generative (Latent Dirichlet Allocation) or a discriminative (Conditional Random Field) problems. These methods often use parameter estimation and maximization to achieve optimum prediction.

1.2 Image Representation and Feature Space

An object detection system contains three components: feature extraction, object classifier on vast multimedia datasets, and fusion engine (Smeaton, Over, & Kraaij, 2006). Choice of visual feature defines image representation. Visual features are highly distinctive attributes of a region used to identify a class of object or a scene with high probability. Features are invariant and often more compressed representation extracted around salient patches of the image that gather rich local information about that location. Features are often called feature-points or key-points, and during the training phase, a set of all features is constructed which called feature space. Application dictates the choice of the feature which can increase the complexity of the classifier with direct impact on the accuracy of the results.

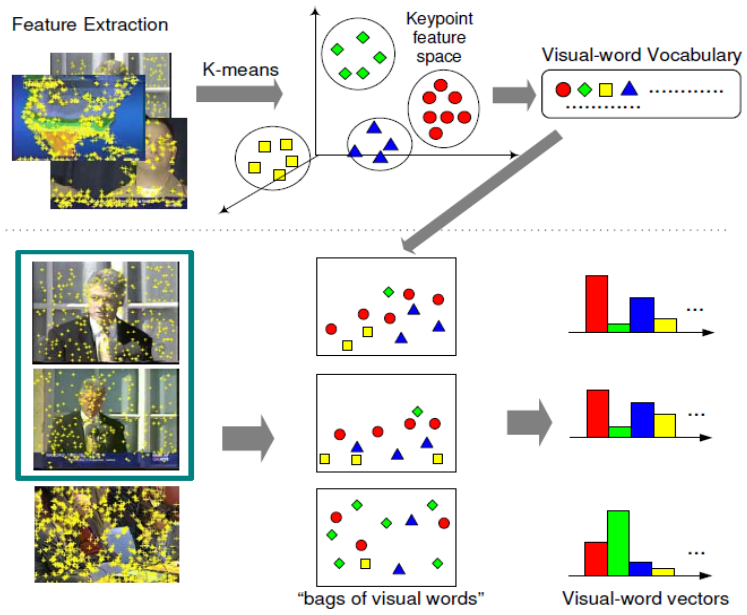


Figure 3. Image representation and encoding using Bag-of-Words

One of the chief problems with using visual features to detect objects is the presence of variations in real life images. Adverse effects of variations in point of view, scale and morphological transforms can be mitigated at feature level by applying a robust algorithm. Some of the most common algorithms to craft the feature descriptors are Scale-Invariant Feature Transform (SIFT) (Lowe, 2004), Speed-Up Robust Features (SURF)(Bay, Ess, Tuytelaars, & Van Gool, 2008), Binary Robust Invariant Scalable Keypoints (BRISK)(Leutenegger, Chli, & Siegwart, 2011) and Oriented Fast and Rotated Brief (ORB)(Rublee, Rabaud, Konolige, & Bradski, 2011). Gradient-based algorithms such as SIFT sample best key-points in a scale pyramid type and extract orientation using the directed gradients or moments. These features are invariant to image scaling and rotation, and partially invariant to change in illumination and 3D camera viewpoint(Lowe, 2004). Feature-points are calculated during the training phase. To classify an

object, feature-points of that object (or a variation of it) is compared to set of key-points for that class.

Feature-points are used to construct feature descriptors, a building block of the visual word in the vocabulary of the image representation in *Bag-of-Words* (BoW) or *Bag-of-Feature* (BoF) (see Figure 3 from (J. Yang, Jiang, Hauptmann, & Ngo, 2007)) methodology. In BoF feature descriptors are clustered into homogeneous groups to build a visual vocabulary regarding the cluster centers as visual words. Encoding an image in terms of visual word frequencies is performed by mapping their feature-points into the ones in vocabulary. A feature vector containing these frequencies is built as input to classifiers such as SVM (Van De Sande, Gevers, & Smeulders, 2009).

1.3 Concept Recognition

Concept recognition systems focus on labeling detected objects with the most semantically compatible labels that best describe the scene. As one of the most important applications of computer vision, parsing real-life scenes suffers from an array of variations such as pattern textures, lighting conditions, and scale of the objects. There are many permutations of different poses (e.g., front, back or side of the car or body part articulations like a door opened), illumination (e.g., brown, white) and occlusions of the objects being analyzed (see Figure 4). Training such object detectors will require a significant number of sample images with enough sample images for each variation. When the number of images increases, the number of visual signatures needed to recognize an object increases exponentially. Many objects are deformable or capable of articulation and changing the geometry of their shape. The human body is a good example of such a complex composition. Limbs, legs, and head are attached to the body with certain geometrical constraints and can move in many directions. The body itself can pose in many different ways. Even objects like cars have windows, doors, trunk and hood that can open and

Figure 4. Variations of training images for car object.

close and that is a limiting factor of conventional object recognition methods. One way to approach this problem is building object detectors for every part and use a part-based approach to solving the problem of their composition (Pandey & Lazebnik, 2011).

Some real life images may contain objects that are not directly recognizable using their visual features or appearance, and other attributes may be required. For instance, some of the objects are too similar to other more familiar objects and can be easily mistaken. Many real life images acquired by portable devices suffer from poor quality, low resolution, glare, saturation of colors, noise, skewed objects and grainy pixels among other things.

In summary, the practical concept detectors will need to solve the following main challenges of the object detection to be successful:

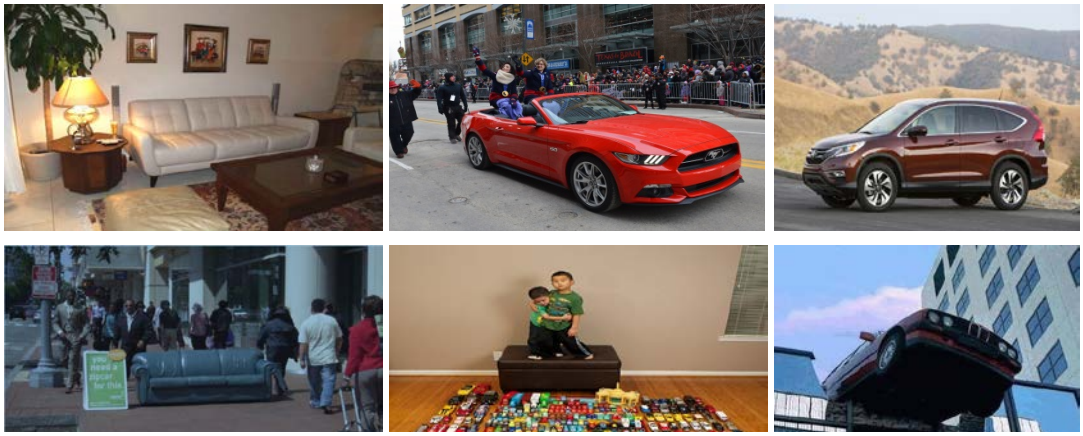
- *High dimensionality* of feature space required to learn all permutation of that an object appear in real life images,

- *Weak Features*: even most superior feature transform algorithms fail to produce high-quality features in extremely impoverished and degraded image conditions,
- *Deformity and occlusion* make object detection based on visual features very challenging,
- *Semantical Structure*: Analyzing objects at the isolation discard vital semantical structure that can be essential to an accurate description of the scene. Semantic structure within each scene category shows how physical objects interact with each other (e.g., the car always has tires or car is always on the road) and how they relate to the scene (i.e., the sky is on the top and grass is always on the bottom).

Concept classification requires the candidate object to meet appearance similarity criteria and to comply with constraints model that describe the inter-object and object-scene relationship in context to maximize the semantical consistency among the detected concepts. State-of-the-art methods and their extensions explored in this dissertation integrate the contextual constraints at classification level and perform scene-wide optimization until the most desirable set of objects, captions are determined. These captions are shown to be the most salient concept and closest to human semantics. Only possible refinements to the result of these methods are ontology based inference.

1.4 Context in Scene Understanding

Motivated by challenges that concept detectors face, a combination of visual features with other sources of information obtained from visual image properties called "*contextual information*" is used to increase objects classification accuracy (Chang et al., 2008). Studies on human neurophysiology and psychology suggest brain activity increases when viewing objects in right context. Neuroscientists attribute this increase in brain function to the analysis of semantical relationships



a) Semantical context b) Scale context c) Spatial context

Figure 5. Constancy (top) and inconstancy (bottom) of named relations.

(Gronau, Neta, & Bar, 2008). Contexts encode additional knowledge about a scene and objects, their relationships, and in generalized form include any information that can improve semantical concept recognition. We define context as "any information that can maximize semantical relevance and accuracy of the detected concepts in an image or a video" This definition implies that context is directly related to increasing the relevance of the discovered concepts and not the process of the discovery itself. Also, type of information that context conveys at each level of the concept detection may vary but as long as the objective of contextual relevance is facilitated context is involved.

Classification of weak lower-level features depends on influencing contextual properties on inference to produce more accurate meaning of scene concepts. Some applications use external source information such as camera setting or location priors to infer more high-level semantical concepts not directly obtainable from the visual attributes.

Other approaches use global descriptors of an image as contexts to classify the scene as a high-level category such as indoor or outdoor to minimize the number expected objects to reduce dimensionality. When the scene type is known (e.g.

busy street corner), the context model suggests a list of possible object classes that may appear in that type of scene with a probability of their appearance. For example cars, bicycles, pedestrian, signs, buildings, and trees are highly anticipated to appear on the picture of the street.

We assume the scene category unknown in advance in this dissertation. Therefore, semantical label consistency of the discovered objects is achieved by finding the configuration that maximizes the overall relevance of predicted label settings.

If classification results in a caption that is incompatible with other objects, contextual constraints shall enforce pairwise and global consistency constraints to enforce semantical similarity among assigned object class name. A concept recognition system may build an additional layer on top of the regular object classifiers to refines their output based on the constraints described above.

The role of context is more important in order-less feature-based approaches (i.e., BoF) because these representations don't capture information about scene layout. Adding spatial, scale and co-occurrence context combined with the flexibility of BoF approach compensate for missing information while benefiting from the efficiency of these methods.

Poor image quality, occlusion, and low definition in cluttered images also underscore the importance of contextual information in correcting initially misclassified instances. Real life scene configuration is often very complex, and multiple types of contexts may be required to ensure the detected objects are labeled correctly. In addition to the number of context types, the configuration of object relationships is also very importance. The study shows pairwise relationship may not be able to identify inconsistencies among labels assigned to objects but higher order (e.g., ternary) can.

Table 1. Taxonomies of context

2D context	models global scene statistics (e.g., gist)
3D context	corresponds to the 3D geometric structure of a scene and surface layout
Cultural context	photographer bias, dataset selection bias, visual clichés (Divvala et al., 2009)
Folk Context	heritage and traditions and customs related to a country, a religion or any groups of people
Illumination context	color invariant descriptors used to obtain similarity invariant descriptors
Photogrammetric context	information extracted from photos by measuring points of interest occasionally used in satellite imaging and remote sensing
Scale/Size context	information pertaining relative size of objects
Semantic context	probability appearance of objects together
Sensor context	any information obtained from acquisition device such as geographic location time of capture or depth of field
Spatial/Position context	the relative alignment of objects
Temporal context	similar content in spatiotemporal domain
Weather context	describes the meteorological characteristics of a region (e.g., Caribbean is tropical) at a point in time (Divvala et al., 2009). For instance temperature in Alaska falls below zero in winter. Winters of Florida are sunny and warm
Web Context	classification of a scene, event or pattern based on social media response and web bias

Figure 5 shows some classification challenge of images with consistent and inconsistent object arrangements. Knowledge of co-occurrence, scale or spatial context helps to modify the type of objects to more compatible ones or to identify out of context objects. For instance, the scale context discriminates cars in the bottom middle picture to be classified as toy cars.

1.5 Sources of Context

Various levels of image representation contain contextual information that encodes relations in that layer. Scene-level context encodes high-level semantics of

the image. Mid-level context captures inter-object configurations and object-scene relations. Region-level context refers to intra-object with the whole-part type of relationship among sub-regions or superpixels. Local-feature contexts relate to neighborhood information of the location where the feature belongs to. This information can be the ordering of the neighbors considered or specific details that make the feature invariant or scalable. Table 1 lists the most common categories of contexts widely studied in the literature.

The position of this dissertation is only to include sources with the strong discriminatory power which could be automatically obtained from our training datasets in the experiments. The following is the types of context explored:

- Semantical Context

  Semantical context is the strongest and the most popular sources of context (Carolina Galleguillos & Belongie, 2010; Shotton, Winn, Rother, & Criminisi, 2006). It models object relationships such as co-occurrence statistics in an image. These relationships can have positive (e.g. car always appears on the road) or negative (e.g. car never appears on a tree) association with visual entities.

- Spatial Contexts

  The spatial context contains object relative location information about other objects or their absolute location in the scene. Relative location context captures inter-objects location relationships based on various taxonomies such as horizontal or vertical relative positions (Endo & Takeda, 2005; Fink & Perona, 2004; Hock, Gordon, & Whitehurst, 1974).

- Scale Context

  As one of the most challenging context to model, scale context captures objects relative scale and size relationships and can be a useful discriminating constraint as shown in Figure 5(Murphy, Torralba, & Freeman, 2004).

1.6 Context Selection and Modeling

Selection of context source is an important design consideration and can be dependent on the scene and dataset. Some datasets offer rich labeling information that can be used to compute contexts, and some are very limited. Information obtained from different sources of context varies significantly so there is a benefit in including as many contextual relations as possible.

There is a trade-off between accuracy of the model inference as a result of types and degree of relationships and complexity cost. This trade-off has encouraged many researchers to limit types and order of context in their work. For instance, many have only explored co-location semantics between objects in a pairwise order. This dissertation investigates more generalized context modeling and context integration with state-of-the-art classifiers.

1.7 Context-based Framework

Context-based approaches required building a scalable and yet flexible framework that allows incorporation of multiple sources with various configurations into one unified model. Many questions have to be answered in creating such a versatile system such as how to represent the contextual relations to be able to combine them into one single measurable attribute. Another important question is how to interpret contextual relationships. For instance one of the challenging issues of existing models is the propagation of initialization errors and lack of recovery strategy. A sensible solution may require label consistency optimization while reinforcing constraints at each level using feed forward and backward system.

CHAPTER 2: OBJECTIVES AND CONTRIBUTIONS

2.1 Objectives

The efficiency of scene parsing techniques often suffers from analyzing objects in isolation and ignoring their relations. Relying on visual features only and discarding information such as scene layout information is a source of poor results and incoherent semantical interpretation of the content. As discussed in the previous chapter (section 1.4), contexts can be used to improve classification accuracy and increase the semantical relevance of assigned labels. Objectives of this study are as following:

- To explore how to model complex hybrid context structure that captures important visual attributes and combines them with external sources of knowledge,
- To demonstrate the role of context in concept classification by applying a hybrid context model to improve the result of state-of-the-art classifiers to recognize better and categorize rich media content,
- To use context model in salient objects selection that better describe the semantical meaning of the image closer to the understood human semantics.

2.2 Challenges

There are many problems in modeling contexts and building, optimizing and running a robust context-based framework including:

- Creating efficient context models requires significantly large labeled dataset,
- Dataset of real life images usually contain lots of noise,
- Images in the dataset must contain sufficient number of object types to show all typical relationships and semantical structures,

- Building a context for large datasets is an ambitious task and complexity of the task increases when multiple contexts are combined (e.g., co-location, spatial or scale). To fully facilitate contexts, the underlying context-based framework must be able to build such hybrid context models automatically and to interpret them,
- Contextual relations must capture local and global relations which require modeling long-range and high-order configurations and dependencies,
- Concept detectors must resolve issues such as high dimensionality, variance in training images, and poor image quality similar to conventional object detectors,

2.3 Contributions

The framework presented in this dissertation is a general purpose concept detection framework based on the state-of-the-art generative and discriminative graphical models which incorporate multi-source contextual information to discover scene semantical structure.

- **Framework**

A framework presented in this study that lays a sound foundation for general purpose concept detection and annotation systems. This scalable, versatile and flexible framework enables state-of-the-art classification algorithms to integrate context in scene analysis.

This framework is designed using graphical models that are very efficient methods for modeling complex structured information such as objects, their composing parts and composition constraints. It further demonstrates the role of context in salient concept detection by integrating the contextual model into three graphical approaches: 1) Generative, 2) Discriminative and 3) Gaussian Mixture Model.

- **Classifiers**

Two context-based based methods implemented to demonstrate how context can improve the performance of generative solutions for concept classification and scene categorization. These methods are extensions of Supervised Latent Dirichlet Allocation (LDA) approaches.

As a discriminative context-based approach, a conditional random field (CRF) was designed with unary, pairwise and high-ordered potentials to show superior performance over state-of-the-art baseline annotation methods and improve over-fitting of LDA.

The context-based extension of GMM was also implemented to demonstrate the influence of semantical relevance in object saliency.

Our framework facilitates seamless integration of an arbitrary number of new sources of any order.

- Context Model

We introduced a non-parametric high performance and robust contextual model for image scenes understanding that allows integration of any other sources of knowledge into an easy to interpret paradigm. The generalized model presented here comprises a systematical approach to building and incorporation of multiple sources of context into a single model.

- Context Representation

Presented context model transforms contextual relations into the quantitative measurement of semantical similarity called *Contextual Relevance Score* (CRS). CRS signifies how well an object describes the scene given the interactions with all other objects. We calculate this score for all objects and build a graph that maximizes the overall score based on the contribution of all objects in a high-ordered configuration. The score captures local and global inter-dependencies between objects in each scene.

- Context and Object Saliency

We conclude our work by studying the relationship between our annotated objects' saliency and the scene category using *Gaussian Mixture Models* (GMM). We provide a metrics to measure and quantify the similarity and importance of objects to the cognitive semantical meaning of the scene. We introduce *Object Saliency Score* (OSS) and use it to rank most important objects for scene categories. Comparing our results with data gathered from human subjects demonstrate the effectiveness of the context-based methods presented in this dissertation and its similarities human annotations.

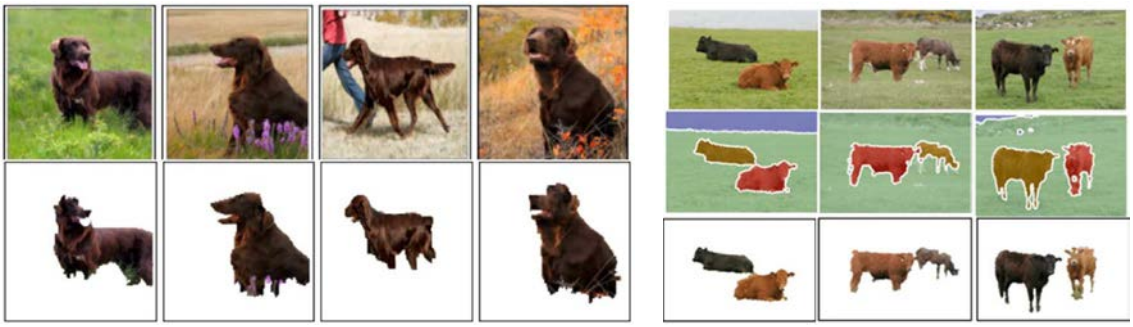
CHAPTER 3: SCENE UNDERSTANDING LITERATURE REVIEW

3.1 Introduction

Scene understanding is aiming at accurately describing details of scene concepts. For video content, the description may also include spatiotemporal concepts such as start and end of events, type of actions, human poses and their interactions. Creation of high volume of contents requires an automated annotation framework to generate the content description based on audio-visual features in an image and inferring overall scene description.

Real images can have an overwhelming number of low-level descriptors that must be analyzed during concept recognition. To address high dimensionality issue, some researchers have proposed a top-down paradigm that exploits high-level course features to infer global knowledge about the scene category such as outdoor or indoor (Murphy et al., 2004; Torralba, 2003). The objects are usually organized in tree-like hierarchical structure with each node corresponding to an object, and each edge is a score showing the level of association of that object to that scene category selected in the node. Dimensionality reduction is achieved by first transforming the global features into a descriptor that can be used in the classification of scene category. There are a limited number of objects with a strong association with a scene category which can reduce the number of possible detectors needed to analyze the image. For instance, a typical indoor scene of a family room can contain a sofa, a TV and coffee table and an outdoor image of the street may include a picture of cars and buildings.

In addition to the high dimensionality of features space being analyzed, running classifiers without any constraints and dependencies can produce potentially



a) Common foreground b) Multi-class segmentation

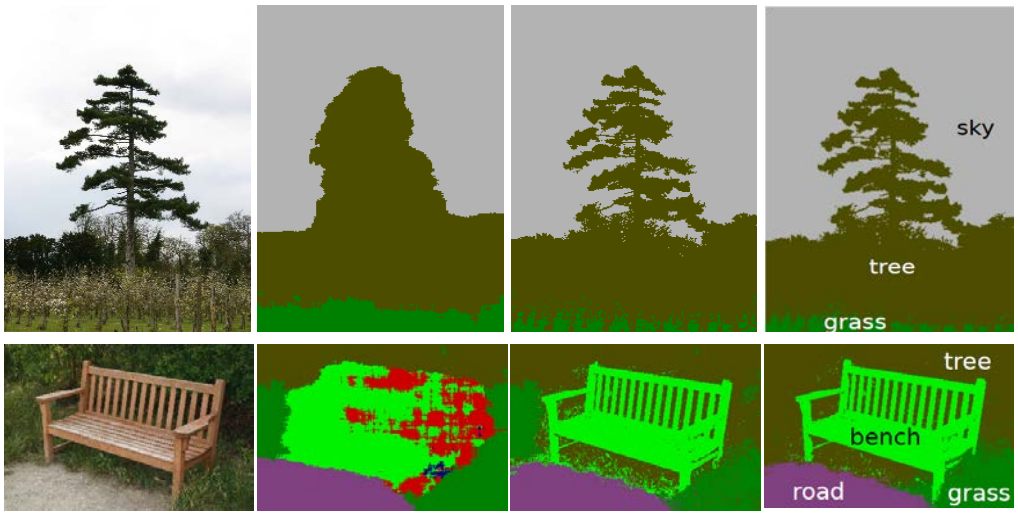
Figure 6. Image segmentation (H. Zhu et al., 2016).

incorrect prediction due to variation in visual descriptors. Contextual information can be exploited in disambiguation of visually similar and recognition of hard to detect objects in a scene. With proven role of context in improvement of recognition systems performance the significant dependencies among the scene components can be exploited by classifiers (Choi, Torralba, & Willsky, 2012b; Carolina Galleguillos & Belongie, 2010; J. Wang, Chen, & Wu, 2011; Zhang, Kalashnikov, Mehrotra, & Vaisenberg, 2014; Y. Zhu, Nayak, & Roy-Chowdhury, 2013). Contextual scene understanding frameworks have been studied in many of the previous work and two conventional approaches widely used to tackle this task are generative and discriminative modeling methods (Jones & Shao, 2014; Tang, Shao, & Zhen, 2014).

We evaluated four models based on generative, discriminative and mixed membership classification using our context-based framework. Mixed-membership models achieve high accuracy in clustering and also generate mixed-membership vectors which reduce high dimensional feature space into easy to interpret topic space. In generative classification each class has a model of the type of observation it generates. To find class of a given observation (x), probability of generating x is calculated for each class (y) using their probability distribution $p(x|y)$ to solve the

query “which class x belongs to?” This approach requires availability of distribution models of each class in (y). Using Bayesian inversion, Discriminative approaches attempt to directly map the observation to class label by calculating likelihood of class given observation $p(y|x)$. Discriminative approach is beneficial when distribution models for each class are known.

Latent Dirichlet Allocation is a generative algorithm for topic modeling which associates each document in the corpus with a probability distribution over "topics." Each topic itself is a distribution over words which are learned during parameter estimation. Dirichlet Process starts from a normal mixture model, which is a single global mixture of several distributions. Each document has its mixture distribution over the globally shared mixture components selected based on a latent variable drawn from a global mixture. Each word in each document also has its parameter drawn from a document-wide mixture. The idea is that a probabilistic mixture of some models is used to explain some observed data points that come from one of the models in the mix. A latent parameter specifies which model each data point comes from. *Conditional Random Field* (CRF) a hallmark of discriminative methods, has been widely used to identify inter-conceptual relationships in an image(W. Jiang, Chang, & Loui, 2007). These methods classify concept presented in an image by computing updated marginal probability for each concept detected by individual detectors. CRF model was first applied to natural language and text processing (Lafferty, 2001) as an alternative to Hidden Markov Models. In their work, Shotton et al. (Shotton et al., 2006) applied CRF to image classification problem and using efficient energy minimization algorithm and inference predicted the labels of observed data. Parameters of the models are estimated using the marginal distribution of subsets of training data to predict the labels of new input. CRF approximation and optimization problems require the use



a) Original Image

b) Unary classifiers result

c) Fully connected CRF, MCMC inference result

d) Fully connected CRF, (Krähenbühl & Koltun, 2012).

Figure 7. Pixel-level classification improvement with a fully connected CRF.

of inference. Figure 7 illustrates the result of CRF on pixel-wise classification after applying each potential. Many inference algorithms have been developed and successfully applied to CRF including graph cut(Boykov, Veksler, & Zabih, 2001; Kolmogorov & Zabin, 2004), mean field approximation (Krahenbuhl & Koltun, 2012) and belief propagation (Morik, Katharina & Piatkowski, Nico, 2012).

One of the limitations of CRF is the complexity of learning and inference models that capture global and local relation which is motivation for the development of more efficient methods (Adams, Gelfand, Dolson, & Levoy, 2009; Ladicky, Russell, Kohli, & Torr, 2010). For instance cross, bilateral Gaussian filter-based methods have shown to increase inference efficiency and improve the accuracy of the image segmentation (H. Zhu, Meng, Cai, & Lu, 2016).

3.2 Related work on Context

One of the pioneering works on the context in computer vision is Yakimovsky and Feldman's (Yakimovsky & Feldman, 1973) image segmentation using Bayesian

a) Input b) Superpixels c) Multiple Hypothesis d) Geometric Labels

Figure 8. Spatial context of geometric classes (From Hoiem et. al)

decision theory. They incorporated image domain as context alongside the image measurements, in their segmentation approach. Contextual information was included in many other following works. Fischler, Stra, (Fischler & Strat, 1989; Strat, 1993) incorporated contextual information obtained from expert-based perception system and applied the manually defined set of rules to infer contextual information which was suitable only for specifically synthesized datasets and incapable of the modeling the real world concepts. To be effective, scene analysis requires exceptional accuracy in the parsing of the scene topology and uncovering its semantical structure.

Biederman(Biederman, Rabinowitz, Glass, & Stacy, 1974) showed that contextual information such as biases in object arrangements, objects relative size, and location are important cues for humans and presented impossible spatial relationships as violations and suggested as number violations increase, the performance of recognition decreases. Torralbo et al. (Torralbo et al., 2013) studied humans ability to categorize scenes and showed some images were a better representative of scene category than the others. They referred to easy to classify sample images as “Good” exemplar and the difficult ones as “Bad” exemplars. Analysis of the image statistics of the good and bad exemplars showed that variability in low-level features and image structure is higher among bad than good exemplars(Torralbo et al., 2013).

Spatial context models relative object locations with respect to each other or their absolute position in the scene. In study by Freeman et al. (Freeman, Murphy, & Torralba, n.d.), spatial relationships were quantized to four prototypical relationships; $\{above, below, inside, around\}$ whereas in (Tu, 2008), a non-parametric map of spatial priors was learned for each pair of objects. Desai et al. (Desai, Ramanan, & Fowlkes, 2009) combined individual classifiers by using spatial interactions between object in a discriminative manner. Heitz and Koller (Heitz & Koller, 2008) combined a sliding window method and unsupervised image region clustering to leverage *stuff* such as the sea, the sky, or a road to improve object detection. Pandey et al., (Pandey & Lazebnik, 2011) used a deformable part-based model to learn the spatial contextual structure of regions of interest using latent SVM. Zhu et al. (S. Zhu & Yung, 2014) augmented spatial context into BOW representation. They incorporate sub-scene attributes within global descriptions by encoding sub-scenes with layout prototypes that capture the geometric structure of scenes to improve categorization performance. Spatial context aims at modelling location configuration of objects and scene topology which is discarded in order-less approaches such as BoW. An early work of Lazebnik et al. (Lazebnik, Schmid, & Ponce, 2006) on spatial pyramid incorporates spatial context by partitioning the image into increasingly smaller sub-regions and encoding each sub-region into visual word frequencies. The approach extracts spatial context of the image in the pyramid of partitions in which together represent the image. The result demonstrated significant improvement over baseline BOW. Qin and Yung used localized maximum-margin learning to incorporate different types of features into BoW which allows selection of best contextual visual word for the local region and set of candidate contextual visual words. Spatial contextual of each region was paired with neighboring regions to enhance the discriminative ability of the scene categorization with success. Zhu et al. (J. Zhu, Wu, Zhu, Yang, & Zhang, 2012)

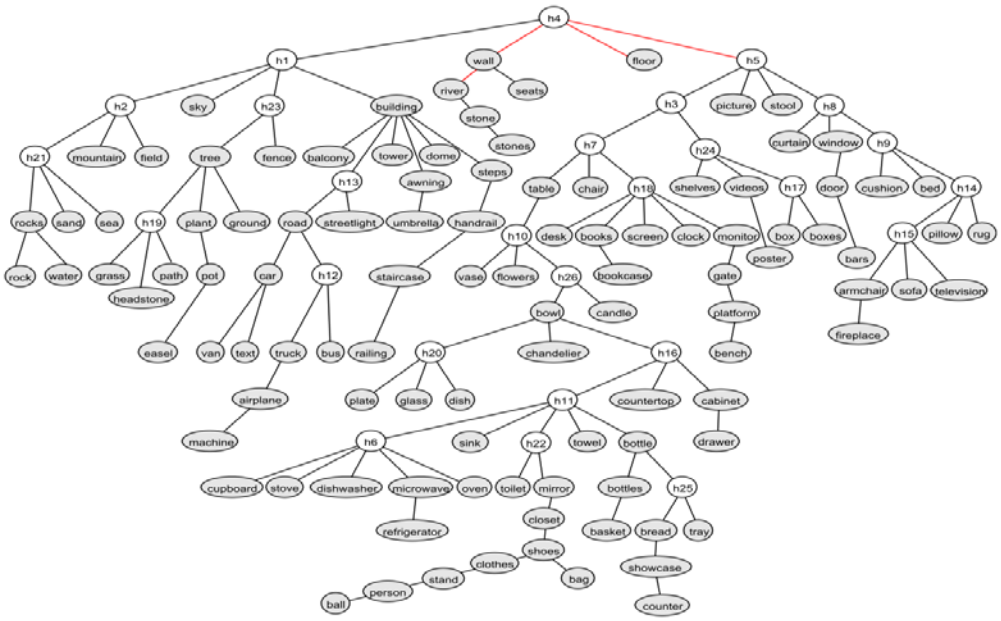


Figure 9. Tree-based representation of context model(Choi et al., 2012a).

presented a method that learned re-configurable and sparse scene representation in the joint space of spatial and appearance features using a dictionary called tangram instead of the fixed spatial pyramid. In this approach, the layout of a mid-level descriptor is approximated by a combination of basic triangles. The shape elements are combined from bottom-up to form a parse tree that contains all possible layout shapes producing a score for a given image at the each node. The spatial layout of an image can be characterized from coarser to fine while scoring the tree top-down. The follow-up works proposed by Wang et al. (S. Wang, Joo, Wang, & Zhu, 2013)replaced the basic shape element by rectangle to reduce the model complexity which expands exponentially and is very time-consuming.

As one of the most studied relation, co-occurrence has shown to be the most useful type of context in both object classification and image segmentation. One of the more influential works in studying the co-occurrence context was presented by Rabinovich *et al.* (Rabinovich, Vedaldi, Galleguillos, Wiewiora, & Belongie, 2007).

They used the output of local detectors first to assign an object label to each image segment and then adjusted these names using co-location context in CRF(Rabinovich et al., 2007). The pairwise potential of the CRF was used to model co-occurrence of objects in the fully connected graph. Galleguillos *et al.* (C Galleguillos, Rabinovich, & Belongie, 2008) and (Gould, Rodgers, Cohen, Elidan, & Koller, 2008) extended the co-occurrence context to incorporate the spatial relationships of objects by modeling pairwise relative location of objects.

Qi et al. (Qi et al., 2007) have distinguished between two types of concept detection approaches. The first one uses binary classification to detect each concept, ignoring inherent correlations between the concepts. The second one considers semantic context. These context-based approaches built on top of the independent binary detectors combining the results in a context vector. In this setting, detection errors of unreliable concept detectors of the first step propagate to the second fusion step. Qi et al. (Qi et al., 2007) have proposed a third approach that simultaneously classifies concepts and models correlations between them in a single step by using a new correlative multi-label framework.

Wang *et al.* (Y. Wang & Mori, 2010) presented a method to incorporate attributes with co-occurrence using a latent SVM (Felzenszwalb, McAllester, & Ramanan, 2008) for object recognition with attributes. In their approach, the attribute relations were pre-learned through a network of attribute nodes and then interpolated into latent SVM. The captured co-occurrence or mutually exclusive attribute relationships were used to enhance object recognition performance further.

Jiang *et al.* (Y. Jiang, Lim, & Saxena, 2012) proposed the use of Dirichlet Process topic model for capturing the distribution of objects to model human pose states in the scene, which is then was used to generate placements for objects. They extended the Dirichlet process mixture model to discover two types of topics,

one for the human configuration topics and another for the human-object relation topics. A multiple variate classifier may be used to fuse contextual descriptors, and appearance features descriptors for event or activity recognition. To incorporate mid-level descriptions, Fei-Fei and Perona (Fei-Fei Li & Perona, 2005) introduced an unsupervised method that represented images as regions called themes. A theme is a group of representative textons which are generalized visual words used in a topic model. Their themes could infer features at the mid-level description to some extent but not very efficient in addressing ambiguity due to the absence of spatial layout. In their later work, they presented a holistic model that combined scene with objects and layout information which resulted in increased accuracy(Li & Fei-Fei, 2007).

A graphical model is a probabilistic framework where a graph is used to represent the dependency structures among different variables. Combining the use of probability theory and graph theory, graphical models become an effective approach for modeling context. They allow us to represent a distribution over a collection of random variables, using the product of potential functions that are defined on small subsets of random variables. Choi et al. (Choi, Torralba, & Willsky, 2012a) presented a tree-structure graphical model to capture dependencies among object categories using objects co-occurrence statistics and spatial relationships for object recognition and out-of-context object detection applications (see Figure 9). Chen *et al.* (H. Chen, Gallagher, & Girod, 2012) built a BN-based context model by incorporating the attribute relations in a fully connected graph to discriminate between object-dependent and object-independent relationships. Torralba et al. (Freeman et al., n.d.) combined boosting, and CRFs to first detect simple objects (e.g., a monitor) and passed the contextual information to identify other more complicated objects (e.g., a keyboard). Tu(Tu, 2008) used both image patches and their probability maps estimated from classifiers to learn a contextual

model and iteratively refined the classification results by propagating the contextual information. Chang et al. (Chang et al., 2005) have described another system using a parts-based statistical approach to representing an entire key-frame as an attributed relational graph. The parts-based concept classifiers trained on the weakly labeled data are combined with conventional concept classifiers in a late-fusion scheme.

Vogel et al., (Vogel & Schiele, 2007) showed less popular images are more prone to ambiguity and global-based scene categorization approaches do not perform well. Scene composition and scene semantics are essential for these images. Shortcomings of global-based methods suggested need of mid-level description such as a region of interest (ROI), part of an object, an object, or object groups (Juneja, Vedaldi, Jawahar, & Zisserman, 2013).

Pan and Kanade (J. Pan & Kanade, 2013) presented a 3D geometric context in which unary potential contained the geometry compatibility score of the corresponding an object in 3D space and the pairwise potential captured the co-occurrence with other objects. This context used object orientation on the ground plane to identify globally and locally incompatible objects during outlier suppression and noise reduction process. A cascaded classification model in (Geremy Heitz, Gould, Saxena, & Koller, 2008) linked scene categorization, multi-class image segmentation, object detection, and 3D reconstruction.

The appearance of object classes in the image is predicted given shape and context priors. For object recognition, researchers have investigated various sources of context, including context from the scene (Babaguchi, Kawai, & Kitahashi, 2002), objects (Bay et al., 2008) and actions (Fleischman & Roy, 2008). Scene based context harnesses scene classification such as urban, landscape, or higher level indoor/outdoor designation to constrain the objects that can occur in the scene more frequently (i.e. car, trees, road, buildings, street signs appear in

outdoor/street scenes often together). In addition to co-appearance of related objects categories, additional spatial or temporal conditions add more constraints to recognition decisions (e.g. car is expected to be on a road whereas boat is sought to reside on the water).

Torralba et at showed similarity amongst images of the same category from features to spatial layout perspective allows generation of the consistent prototype which can be characterized as the *spatial envelope* (Torralba, 2003).

For the task of object detection, Hoeim et al. (D. Hoiem, Efros, & Hebert, 2005) use geometric context to build a spatial context that divided outdoor images into three hierarchical semantical classes. On the top is the sky, vertical structures such as buildings and trees are in the middle and on the bottom is ground. They use superpixels to extract the spatial context of a single image. Each superpixel is classified into a semantic class based on descriptors obtained from their low-level features such as color, texture, shape and geometry feature vector. They use AdaBoost combined with weak decision tree classifiers to label each superpixel and composing pixels with three semantical classes.

Hoiem et al. (Derek Hoiem, Efros, & Hebert, 2008) defined a framework to model object size and object scale contexts using the coarse scene geometry and used inference to select object hypothesis that is consistent with contextual constraints. Viewpoint prior and location of the horizon in the image is calculated using scene geometry defined as horizon line as the intersection of sky and ground plane and the location of objects on the ground plane. They construct a graphical model of conditional independence for viewpoint, object identities and 3D geometry of surfaces surrounding the objects. Using Pearl's (Pearl, 1988) belief propagation inference they try to determine a cluster of object hypothesis that is consistent with the size and location of an object given the geometry and horizon of the scene. The main limitation of this approach is using the position of the horizon to locate

object categories that are placed on the ground plane and are almost the same size (i.e., this method cannot be used to detect windows on facades or trees).

Wolf and Bileschi (Wolf & Bileschi, 2006) introduced *semantic layers* which are constructed by extracting and combining various features such as color, texture, geometric feature maps and saliency maps at pixel location during the learning stage. Each semantic layer represents an object category and location of objects indicates the presence of a particular object in the image at a semantic layer.

In his other work Bileschi (Bileschi, 2006) presented a new set of scale and position-tolerant feature detectors inspired by the role of ventral stream of visual cortex. Scene images are portioned into four semantical classes of buildings, roads, skies, and trees. These classes are learned from different sets of texture properties features known as HMAX (Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007).

Galleguillos *et al.* (Carolina Galleguillos, McFee, Belongie, & Lanckriet, 2010) explored pairwise interactions between pixels, regions, and objects to extract and learn three sources of context semantic, boundary support and contextual neighborhood.

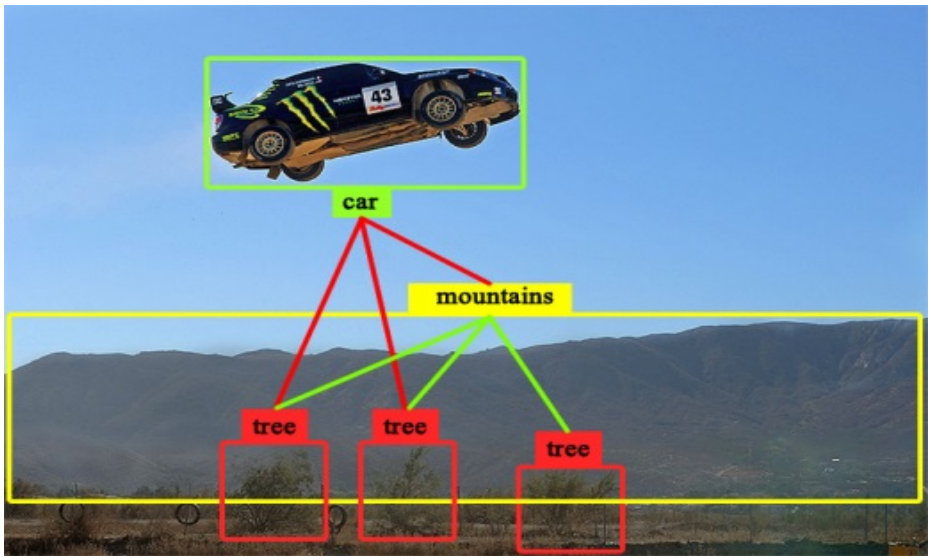


Figure 10. Inconsistency in pairwise versus high-order relationships.

Torralba *et al.* (Choi et al., 2012b) introduced a simple framework for modeling the relationship between context and object properties. Scale context was used to provide a reliable cue for size selection in the detection of high-level structures as objects. Contextual features were obtained from a set of training images and object properties were based on the correlation between the statistics of low-level features across the entire scene.

Jones and Shao (Jones & Shao, 2014) studied pairwise contextual interactions of events and scene elements in a clustering application. They demonstrated performance improvement over state-of-the-art clustering methods.

High-order relationships are examined in (G. Chen, Ding, Xiao, & Han, 2013; Kohli, Ladickýy, & Torr, 2009; Myeong & Lee, 2013) on single source of context such as co-occurrence combined with Robust P^n optimization model. The shortcoming of these methods is when discriminative contextual cues may appear in other contextual modalities such as scale or spatial context.

On the other side of the spectrum, generative models such as (Fergus, Perona, & Zisserman, 2003) are widely used to model multi-context relations. The

limitation of these frameworks and the generative process is the independence assumption on observed data to make the inference tractable which is very restrictive.

3.3 Our Approach to Scene Understating

Previous work shows the success of context-based methods in improving the performance of object localization and recognition. We extend previous work to exploit high-order multi-modal contextual relationships instead of pair-wise approach. We propose a high-order context framework that learns consistent appearance, structure and semantical constraints of the scene and infers its parameters based on the importance of contextual relationship among object types. Objects co-occurrence statistics is defined in high-order to capture scene level semantics. For example objects in "car, motorcycle, road, sky" tend to appear in outdoor street images and "car, truck, rubber duck, Mickey Mouse" represents a set of children toys and most likely is taken from indoors.

Spatial and scale contexts are critical in learning layout topology. Location and size information is obtained from bounding box information in training dataset and transformed into the set of contextual spatial attributes during the learning process. Bounding box information is acquired from the image annotations provided in Sun397[1] dataset (Xiao, Ehinger, Hays, Torralba, & Oliva, 2016).

Context model represents a scene with a graph with fully connected cliques consisting objects at each node. These nodes are linked to undirected edges, and

[1] http://vision.princeton.edu/projects/2010/SUN/

each edge is assigned with contextual relevance score (CRS) that quantifies the strength of the relations between two objects given the dominated context for that clique. CRS is defined to maximize semantical consistencies including scale and location in a scene.

Some contextual inconsistencies may not manifest in pairwise relations wherein ternary relation a clear violation is evident. Figure 10 shows such condition that would have been interpreted differently in the pair-wise association. The image of car illustrates the concept of a flying car and only using relative location constraint enforced by spatial context this event could be differentiated with ordinary "driving car." The object-scene score is scalable and extendible to other datasets since it is not dependent on visual primitives. Contextually related objects form semantically coherent cliques in our graph representation and are labeled accordingly.

CHAPTER 4: CONTEXT MODEL

4.1 Introduction

Given an annotated dataset, our objective is to learn a class-specific model which predicts the existence of a relation between objects in an image using global features of the scene and local features of regions. We also want to learn the importance of those relationships and weight them. To adequately capture the essential structure of mid-level semantics, a full range of relations must be examined in high-ordered configuration to build the contexts. An intuitive representation of such complex relationships is achieved using graphs. Graph-based methods are easier to infer and optimize with many sophisticated techniques already developed.

Our graph-based framework builds on co-relation metrics called the Context Relevance Score (CRS) which is calculated using the high-ordered interaction of the objects. The objective of CRS is to maximize the inter-object semantical coherency and improve their relevance to the scene. The value of CRS for each object pairs determines the relevancy of each object to the clique they belong to and strength of their relation. So, each edge in this graph signifies this relation with a strength determined by CRS. The model introduced here will be used throughout this dissertation wherever referred to context.

4.2 Context Model

Our motivation is that using pairwise relations among image objects is not always informative and there are exceptional cases where information obtained from pairwise relations is not improving to the overall meaning of the scene. We want to identify important relations in the training images to include in our model

and predict the strength of those relations. We then use an iterative approach to maximize the strength of these relationships and form our contextual graphs. To formally describe our model, let's assume there are n object-classes $y = \{y_1, \dots, y_n\}$ in our dataset. Our goal is to learn a function; $\omega(l, y_*)$ which represents strength of relationship l between object-classes given in the vector $y_* \subset y$. Scores for relationship between two or more objects are learned using the following objective function, which maximizes the labeling relevance in the training annotation. If our training dataset contains n images, and image contains arbitrary number of regions $r = \{r_1, \dots, r_k\}$ then the cost function can be written as:

$$\mathcal{L} = \sum_{t=1}^{m} \left(\sum_{i} a_i^t f_a(r_i) + \sum_{y_* \in \Delta} F_c(y_*) a_*^t \right) \tag{1}$$

In this function, a_i^t is n-dimensional ground truth annotation vector for region r_i , the function $f_a(r_i)$ represents the appearance features similarity score of that region in the image with modeled features of all trained objects in the dataset. The result is a n-dimensional vector comprising similarity scores. The matrix $F_c(y_*)$ is of dimension $n \times n$ representing contextual compatibility of object-classes in y_*. Finally, a^t is a $n \times n$ matrix which contains ground-truth annotations of the object-classes and Δ existing relations in the image t such that $\{y_* : F_c(y_*) > 0\}$. The function $F_c(y_*)$ can be written in the term of high order relationships of all context types L:

$$F_c(y_*) = \sum_{l=1}^{L} \omega^l(y_*)$$

The function in Equation (1) can be maximized if contextual constraints have the most compatibility. Also, the learned feature scores should be such that the consistent relationships should have a higher score as compared to the other ones. Maximizing equation (1) allow us to obtain a subset of all relationships in all images to form our contextual graphs.

4.3 High-order Relations Formulation

High-order relationships is formulated based on n-order pure dependence rule to quantify relationship of M binary variables (Hou, He, Zhao, & Song, 2011):

$$\omega^{s}(y_*) = \log \prod_{m=0}^{M} \prod_{\mathbb{x} \in A_{y_*}^{(m)}} P_{\mathbb{x}}^{(-1)^{(M-m)}} \tag{2}$$

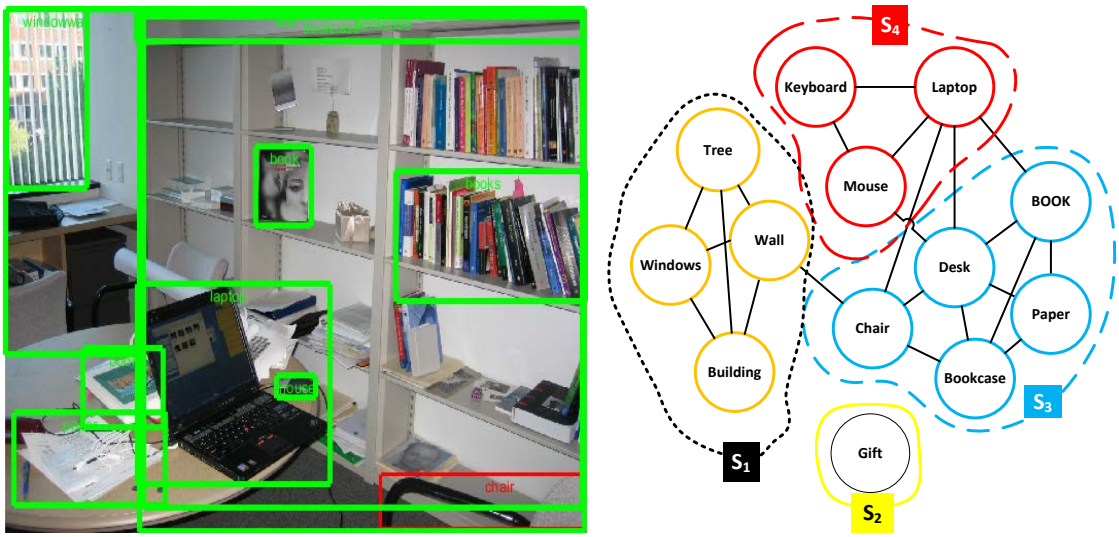
where M is number of object classes and $A_{y_*}^{(m)}$ is the set of all configurations of m-order relations (more detail below) and $P_{\mathbb{x}}$ is the probability of participation of the objects in the relation s within the configuration $\mathbb{x}$.

Encoding contextual possibilities of M-objects into one simple to interpret distribution can be achieved by using Equation (2). To formulate the high-order relations, let's assume $y_* = \{y_1, \dots, y_M\}$ is the set of variables representing object classes in relation, and $A_{y_*}^{(m)}$ represents the set of all assignments of M object-classes participating in m-order relations. For example considering set of four object classes $y_* = \{y_1, \dots, y_4\}$, the set of object configurations terms for all relations will be as follows:

$$A_{y_*}^{(0)} = \binom{4}{0} = \{0000\}$$

$$A_{y_*}^{(1)} = \binom{4}{1} = \{1000,0100,0010,0001\}$$

$$A_{y_*}^{(2)} = \binom{4}{2} = \{0011, 0101, 1001,0110, 1010, 1100\}$$



a) Image of an office scene

b) Inter-object relationships forming cliques of strongly related objects based on their SRS calculations.

Figure 11. Scene contextual representation of object relations.

$$A_{y_*}^{(3)} = \binom{4}{3} = \{0111, 1011, 1101, 1110\}$$

$$A_{y_*}^{(4)} = \binom{4}{4} = \{1111\}$$

At simplest form co-occurrence or semantical context can be directly obtained from Equation (2). For example the third-order co-appearance relation object assignments for $y_* = \{y_1, y_2, y_3\}$ can be calculated using the following formula:

$$\omega^s(y_{ijk}) = log\left(\frac{P_{001}P_{010}P_{100}P_{111}}{P_{000}P_{110}P_{011}P_{101}}\right)$$

Each probability term ($P_{000}\ to\ P_{111}$) encodes possibility of co-appearance of the objects for the given configuration (e.g., $P_{101} = P(y_1 = 1, y_2 = 0, y_3 = 1)$) .

Interpretation of high-order relationship based on values of ω is straightforward. Positive values imply existence of a correlation between objects represented by y_* and negative values represent negative relation and zero means no relation in other words variables are independent.

4.4 Contextual Relevance Score

Contextual relevance score is more generalized form of Equation(2) which incorporates several types of contexts:

$$\omega(y_*) = \omega^s(y_*) + \omega^l(y_*) + \omega^x(y_*) = log \prod_{m=0}^{M} \prod_{\mathbb{x} \in A_{y_*}^{(m)}} \left(P_{\mathbb{x}} P_{L_{1\dots n}^v|\mathbb{x}} P_{X_{1\dots n}|\mathbb{x}}\right)^{(-1)^{(M-m)}} \quad (3)$$

The term $\omega^s(y_*)$ is the probability which encodes semantical relationship among the objects and can be directly calculated using Equation (2). The term $\omega^l(y_*)$ is location context and is defined as conditional probability of relative vertical location of an object in respect to others in high-order relation specified in $A_{y_*}^{(m)}$ configurations given their co-occurrence. For three objects example the location context can be obtained from $\omega^l\left(y_{ijk}\right) = log\left(\frac{P_{L_{123}^v|001} P_{L_{123}^v|010} P_{L_{123}^v|100} P_{L_{123}^v|111}}{P_{L_{123}^v|000} P_{L_{123}^v|110} P_{L_{123}^v|011} P_{L_{123}^v|101}}\right)$. Relative vertical location configuration is determined by comparing centroids of each object's bounding box in a vertical alignment. For example expected relative probability distribution of "*sky*" is "*above*" the object "*grass*" and "*building*". Figure 12 shows histograms of object-pairs spatial relations fitting results obtained by applying Equation (3) on Sun397 dataset ground truth. Strong inter-object constraints show clear influence of the spatial and scale probabilities as spike in the histogram. The fitting results shown in Figure 12 demonstrate gamma distribution is the best fit to model this context.

$\omega^x(y_*)$ is high-order scale context and is defined as joint probability distribution of $X_1, X_2, \dots, X_{\mathrm{n}}$ where X_{i} is the expected relative scale relation obtained by transforming the image plane into 3D coordinates for relatives scale measurements based on labeled training sets. Information gathered from a relative horizontal location does not offer discriminative information and is not modeled.

Equation (3) can be generalized to incorporate other sources of contextual information required to describe the mid-level semantical structure better. The following equation shows high-order CRS calculation for four contexts:

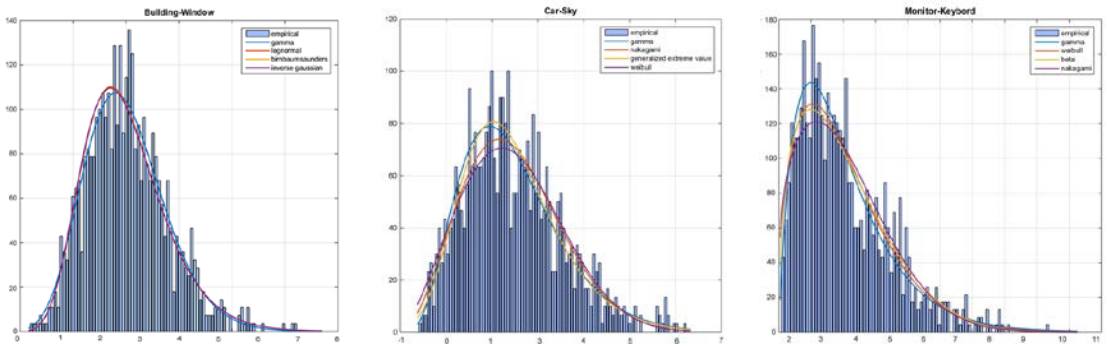
$$\omega(y_*) = log\left[\prod_{m=0}^{M}\prod_{\mathbb{x}\in A_{y_*}^{(m)}}\left(P_{\mathbb{x}}P_{L_{1\dots n}^{v}|\mathbb{x}}P_{X_{1\dots n}|\mathbb{x}}P_{R_{1\dots n}|\mathbb{x}}\right)^{(-1)^{(M-m)}}\right] \tag{4}$$

$P_{R_{1\dots n}|\mathbb{x}}$ is a generalized high order relationship among objects defined over a set of contextual constraints. For instance R can be ordering of objects in a scene like car is *on* the road, or tree *in front of* building or clouds are *over* the sky. Then $P_{R_{1\dots n}|\mathbb{x}}$ encodes the relationship $\{on, under, over, behind, exclusive\}$ among those objects.

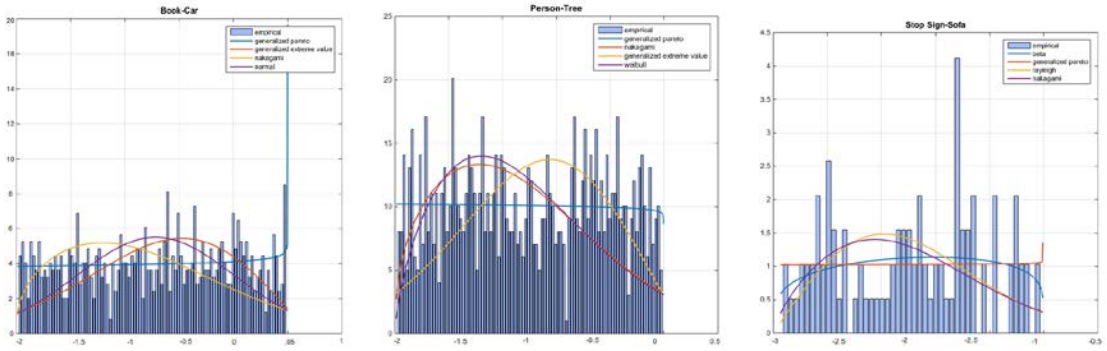
To construct the graph in Figure 11, given an image t with m objects, first we evaluate all pairwise relationship by computing ω for all objects pairs and only connect the edges with positive score. We query ground-truth annotation data to obtain the probabilities needed to evaluate ω. We repeat these steps for object combinations of higher-order (i.e., 3, 4, 5) and compute ω for those relationships. Based on outcome of high-order ω, an existing edge could be dropped or a new edge maybe added to the nodes corresponding to related objects.

Contextual score for each object is stored in S, as a vector of size M^2. These scores are computed for each node of contextual graph $G^c = (v, e)$ using softmax over their total scores as following:

$$s_i = \frac{(\sum_{\varepsilon\in v^i}\varepsilon)}{\sum_{i=1}^{M}(\sum_{\varepsilon\in v^i}\varepsilon)} \tag{5}$$



a) Strict contextual interactions show strong relationships and form a peak on positive side. These distributions are fitted to Gamma distribution.



b) Histograms of weak relationships are scattered around negative numbers.

Figure 12. Histograms of the object pairs CRS calculated using Equation (3).

In this equation, ε is set of all edges connected to node v^i which is the i^{th} object in the context graph.

In practice computing probabilities of relationships of orders higher than four become very tedious while they become less informative in forming semantically consistent cliques. We have only explored pairwise and ternary relationships in our experiments for this dissertation.

CHAPTER 5: GENERATIVE APPROACH TO SCENE UNDERSTANDING

5.1 Introduction

Many unsupervised clustering approaches have been studied to discover the latent structure of observed data such pixels among them probabilistic mixture models or generative models such as Naïve Bayes. Mixture models in general however suffer from the limitation which restricts a data point to have mixed memberships to various classes with varying degrees. This rigidity is not desirable in many applications particularly in scene understating and inspired extensions to mixed membership including the state-of-the-art topic modeling approach *Latent Dirichlet Allocation* (LDA) (Blei, Ng, & Jordan, 2003).

LDA is a probabilistic generative model that represents a collection of discrete data such as a document or image features as a finite mixture over a set of topics. LDA uses latent variables to capture document or image semantics by discovering the intrinsic structure of data and decomposing distribution of the population into groups called topics. A document is represented as topic mixtures and topics viewed as a multinomial allocation of data.

Standard LDA is an unsupervised method and is unable to capture semantic structure of the document. Therefore, many semi-supervised and supervised variations have been presented in the discovery of a set of topics with most of the semantical correlation. In computer vision supervised extensions of LDA has been applied to the discovery of possible relationship of visual image patterns for clustering and classification applications. These methods use a class label variable in the *variational inference* of generative model.

There are many studies on image annotation such as correspondence-LDA (corrLDA) (Blei & Jordan, 2003) and multivariate regression LDA (LDA-bin) (Putthividhya, Attias, & Nagarajan, 2010) and multi-dimensional binary supervised LDA (xLDA-bin). LDA has been applied to a variety of other tasks for instance labeled LDA (Ramage, Hall, Nallapati, & Manning, 2009) (labLDA) for credit attribution in multi-labeled corpora, semiLDA (Y. Wang, Sabzmeydani, & Mori, 2007) for human action recognition in videos among others.

Two of the pioneering and most influential work in the area of classification and image annotation are classLDA (cLDA) (Fei-Fei Li & Perona, 2005), and supervised LDA (sLDA)(C. Wang, Blei, & Li, 2009) which employ mean-field variational inference in approximation of conditional distribution of the latent structure. These models outperformed competing models in regards to classifications and annotations, but later studies show their inability to influence class information effectively in semantic topics discovery for large samples (Rasiwasia & Vasconcelos, 2013). These experiments show that the classification accuracies of cLDA and sLDA are not superior to those of unsupervised topic discovery either.

In this study, a context-aware framework is presented that builds on several previous works in topic modeling. Contextual relations influence topic formation based on various contextual relations that are disregarded by Dirichlet process. Global relations of the low-level and mid-level descriptions are accounted in high-range and high-order relationship to uncover scale and location configurations. In contrast to pairwise relationships which only act on local and neighboring regions, high-order dependencies cover overall scene and provide a higher level contextual consistency. The new supervised models learn topic distributions conditioned on contextual inter-object relations for image classification and annotation tasks. The first model presented uses multivariate Bernoulli distribution with logistic link

function and the second model utilize non-central hypergeometric distribution called *Wallenius* (Chesson, 1976; Fog, 2008) to infer object classes.

5.2 Feature Representation

The most appropriate feature representation for LDA is *Bag of Features* (BoF)which we will use to present our observed data. As in original BoW from text processing, ordering is ignored and as result need for incorporation of contextual structure alongside a robust scale and rotation invariant feature descriptor is paramount. Consequently, SURF(Bay et al., 2008) feature points are extracted to build feature descriptors. We build code-book vocabulary by applying K-means clustering algorithm to a subset of the strongest feature descriptors. This quantizes the feature space into a more manageable mid-level representation of visual words. Hence, visual words are centroids of each cluster and encoding is done by mapping feature space to closest of these visual words. The number of visual words in vocabulary has a direct impact on the model performance, and optimum dictionary size can be set empirically. To encode an image, visual words frequency vector of the image features is computed using the trained vocabulary.

5.3 Formulation

Let Ψ be set of all images in D, and each image is observations from random variable X defined on feature space $\mathcal{X}$ of visual feature descriptors. Image is divided into patches each represented as feature vectors. Consequently an image is represented by bag of independently sampled feature vectors $\mathcal{I} = \{x_1, x_2, \dots, x_N\}$, $x_n \in \mathcal{X}$ and $|\mathcal{I}| = N$. Feature space $\mathcal{X}$ is quantized into $|\mathcal{V}| = M$ bins and the centroid of each bin is considered a visual word in the vocabulary $\mathcal{V}$. Each feature of image is mapped to nearest feature x_n in vocabulary. Feature vectors of an image $R = \{r_1, r_2, \dots, r_N\}$ are mapped to visual words in $\mathcal{V}$, where r_{n} is the bin containing x_{n} and $|r_n| = M$ is unit-basis vector that represents the correspondence

to only 1 visual word in our vocabulary V. Presence of each object-class caption in an image is stored in binary vector c. A non-zero entry c_i denotes that caption i is present in that image. Each image class is associated with a random variable Y that denotes scene class category $y = \{1, \dots, C\}$ making $D = \{(\mathcal{I}_1, y_1), \dots, (\mathcal{I}_D, y_D)\}$. Contextual relationships are represented by graph $G^c = (v, e)$ where each node v corresponds to a caption and the edges represent relationships in the term of contextual relevance score (see generalized Equation (4)) between the captions. A collection of D image-label pairs is denoted as $\{R_d, C_d\},\ d \in \{1,2, \dots, D\}$ where R is the image feature vector and C is annotation words' presence vector for document d.

5.4 LDA Graphical Model

Unsupervised LDA plate graphical diagram is shown in Figure 13. This diagram illustrates a process for generating an image. This is joint distribution of observed variables (i.e. visual words (r_n)) and latent variables (i.e. topic assignments). Each plate indicates repetition.

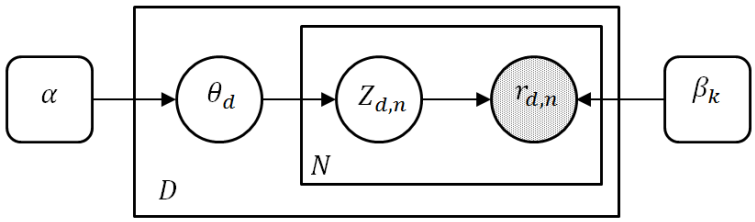


Figure 13. Unsupervised LDA plates

The Unsupervised LDA does not incorporate class labels and can follow the following generative process:

For each image in D

Draw topic proportions from Dirichlet prior $\theta \sim Dir(\alpha)$.

For each visual word $r_n, n \in \{1, 2, \ldots, N\}$*:*

Sample a topic assignment $z_n|\theta \sim Mult(\theta)$

Sample a region visual word $r_n|z_n \sim Mult(\beta_k)$

end for

end for

where $r_n|z_n \sim Mult(\beta_k)$ is a distribution of categories on $\mathcal{V}$ with parameter $\beta_{1:k}$. One of common methods to learn model parameters is expectation maximization (EM) algorithm. However E-step inference is known to be intractable and must be approximated using approximation algorithm such as Variational inference or a sampling method like Gibbs sampling.

Supervised LDA uses class labels in topic discovery for image annotation task(Mcauliffe & Blei, 2008). Many supervised approaches emerged since with objective of influencing theme structure by incorporating class label for annotation(Huang, Zhou, & Zhang, 2014; Putthividhya et al., 2010; Y. Zhu et al., 2013).

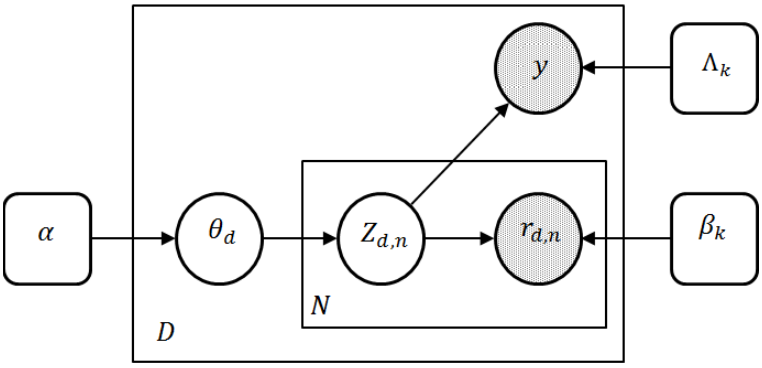


Figure 14. Supervised LDA graphical plates.

In reality, images may contain multiple objects. For example pictures of the "Kitchen" class may include patches of "stove" or "cabinets." Therefore supervised

learning of such distributions requires training set that provides a label for each patch, such that the label is a class in Y. The graphical model of fully supervised LDA is illustrated in Figure 14 and follows the following generative process:

For each image in D

Draw image class label $y \sim P_Y(y;\eta)$

Draw topic proportions from Dirichlet prior $\theta \sim Dir(\alpha)$.

For each visual word $r_n, n \in \{1, 2, \ldots, N\}$:

Sample a topic assignment $z_n|\theta \sim Mult(\theta)$

Sample a region visual word $r_n|z_n \sim Mult(\beta_k)$

Draw class label c_i *conditioned on topic class* $P_{Y|Z}(c_i|Z;\Lambda_{1:k})$

end for

end for

In this process, all classes share a topic simplex. When class label influence increases discrimination by imposing topic supervision, it reduces the model ability to discover latent structure.

5.5 Context-aware LDA Methods

The LDA approach to latent topic discovery is order-less, and words appearance in the document is the only information sampled. Adoption of this method to image processing, resulted in the loss of critical information and making assumptions that visual images and textual words are very similar. Every word in the lexicon is well defined, and its semantical contribution to the meaning of the document is consistent for each topic. However, encoded visual words of a picture can be highly variable because of the way they are constructed and the factors that have an impact on visual features such as image quality, camera movements,

illumination, and image distortion. Adding contextual information is crucial to stabilizing and smoothen this semantical gap in feature-based representation.

In next sections, two approaches are presented with different classification methods and are evaluated and compared with baseline approach. The first model is xLDA-bin and the second model is WLDA which is contextual framework built on generative process defined in xLDA-bin.

5.5.1 xLDA-bin

xLDA-bin is extended version of sLDA-bin(Putthividhya et al., 2010) which adds image-categorization to the model annotation. Figure 15 illustrates the graphical model representation of xLDA-bin with new response variable to used to predict category of image.

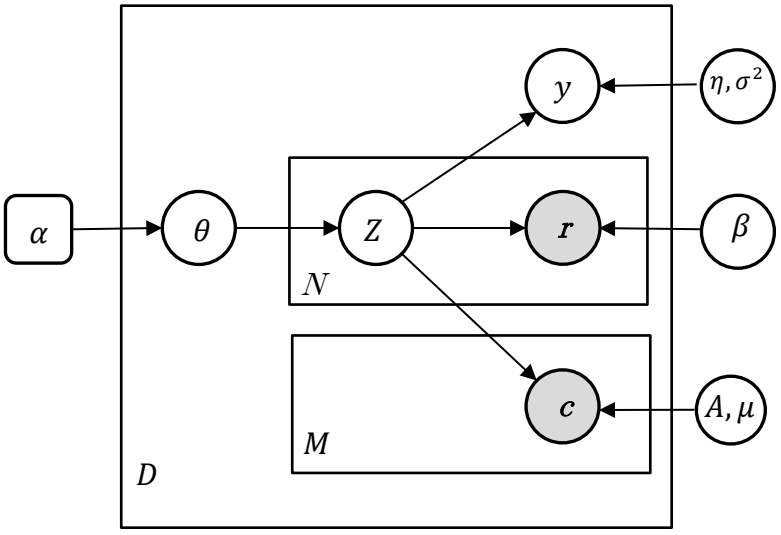


Figure 15. Graphical representation of xLDA-bin model.

In xLDA-bin shown in Figure 15, nodes represent random variables; gray nodes are observed variables and plates denote replicated graph structure. The process of generating images in which every visual word is annotated with an image category label response variables is as follows:

Draw topic proportions from Dirichlet prior $\theta \sim Dir(\alpha)$.

For each visual word $r_n, n \in \{1, 2, \dots, N\}$*:*

Draw topic assignment $z_n|\theta \sim Mult(\theta)$

Draw region visual word $r_n|z_n \sim Mult(\beta)$

Draw response variable $y|z_{1:N}, \eta, \sigma^2 \sim N(\eta^\top \bar{z}, \sigma^2)$

Given empirical topic proportion $\bar{z} = \frac{1}{N}\sum_{n=1}^{N} z_n$*., for each caption word* $i \in \{1, 2, \dots, M\}$ *draw Bernoulli* $c_i \sim p(c_i)$

Our objective is to build a model to predict caption data as the response variable by learning topic proportions that are predictive of the caption words given the contextual model. Distribution of binary response variable is modeled using multivariate Bernoulli distribution. The probability of each caption is defined as a logistic function. Each class label is influenced by topics from all image regions as well as by a selected image region depending on the corresponding regression coefficients. This association model is thus more general and accurately reflects the process of how the correct annotation is generated.

The probability of classifying the objects with class label c_i is given as follows (Putthividhya et al., 2010):

$$p(c_i|Z, A, \mu) = \sigma(a_i^T \bar{z} + \mu_i)^{c_i} \sigma(-a_i^T \bar{z} - \mu_i)^{1-c_i} \quad (6)$$

where $p(c_i|Z, A, \mu)$ is a Bernoulli distribution, A, μ are regression model parameters, $p(c_i) \sim c_i \in \{0,1\}$, $\sigma(x) = (1 + e^{-x})^{-1}$ is logistic function, and $\{a_i, \mu_i\}$ are regression coefficients.

To model categories of images in our dataset (Ψ), we add a response variable associated with each image to LDA to represent its category index. We jointly model the image and the responses, in order to find latent topics space that maximizes prediction of unlabeled images. The probability of image class categories can be obtained as:

$$p(\Psi, y|\, \alpha, \beta, \eta, \sigma^2) = \prod_{i=1}^{M_L} p(R_i, y_i\, |\alpha, \beta, \eta, \sigma^2). \prod_{j=1}^{M_U} p\big(R_j\, |\alpha, \beta\big) \tag{7}$$

Where Ψ is set of images in our dataset, M_L is number of labeled images, M_U is number of unlabeled images, R_i is i^{th} labeled image, R_j j^{th} unlabeled image, first term is likelihood of labeled documents and second term is likelihood of unlabeled documents.

The response variable comes from a normal linear model. By regressing the response on the topic frequencies, we treat the response as nonexchangeable with the object captions. The image (i.e., captions and their topic assignments) is generated first, under full word exchangeability; then, based on the image, the response variable is generated(Mcauliffe & Blei, 2008). Marginal distribution of a labeled images and its response can be written as:

$$\begin{aligned} & p(R, y|\alpha, \beta, \eta, \sigma^2) \\ & = \int p(\theta|\alpha) \prod_{i=1}^{N} \left(\sum_{z_{1:N}} p\,(z_n|\theta) p(c_i|z_n, A, \mu) \right) p(y|z_n, \eta, \sigma^2\,) d\theta \end{aligned} \tag{8}$$

The likelihood of unlabeled image can be obtained from the following LDA model:

$$p(R, y|\alpha, \beta) = \int p(\theta|\alpha) \prod_{i=1}^{N} \left(\sum_{z_{1:N}} p\,(z_n|\theta) p(c_i|z_n, A, \mu) \right) \tag{9}$$

In these equations R is an image, α, β, σ are model hyper-parameters, c_j is object caption(as described in 4.4), r_i is visual word and y_i is response variable for image class category.

5.5.2 Wallenius Context-based LDA(WLDA)

Multivariate non-central hypergeometric or Wallenius distribution is the second model used to study contextual dependencies of the objects in an image. Under this distribution, the probability of observing a sample depends on previous observations and also on the remaining terms expected. Furthermore, it allows more biased sampling by incorporation of object type dependent bias weights directly in the probability calculation.

Under the WLDA model, the generative process of annotating a candidate object with its class label response variables is as follows:

Draw topic proportions from Dirichlet prior $\theta \sim Dir(\alpha)$.

For each visual word $I_n, n \in \{1, 2, \dots, N\}$:

Draw topic assignment $z_n|\theta \sim Mult(\theta)$

Draw region visual word $I_n|z_n \sim Mult(\beta_r)$

For each object class label

Draw a label conditioned on contextual constraints given by

$p(c_i|Z, S) \sim Wall(c_i, Z, S)$

Draw response variable $y|z_{1:N}, \eta, \sigma^2 \sim N(\eta^\top \bar{z}, \sigma^2)$

Our objective is to obtain probability of most semantically consistent labeling configuration Y given topic distribution and vector of contextual scores $S = \{s_i, \dots, s_M\}$ where each s_i is obtained from Equation (5):

$$
\begin{aligned}
p(c_i|Z,S) &= \Lambda(c_i,Z)I(c_i,S) \\
\Lambda(c_i,Z) &= \prod_{i=1}^{k} \binom{c_i}{z_i} \\
I(c_i,S) &= \int_0^1 \prod_{i=1}^{M} \left(1 - t^{\frac{s_i}{s}}\right)^{c_i} dt
\end{aligned}
\tag{10}
$$

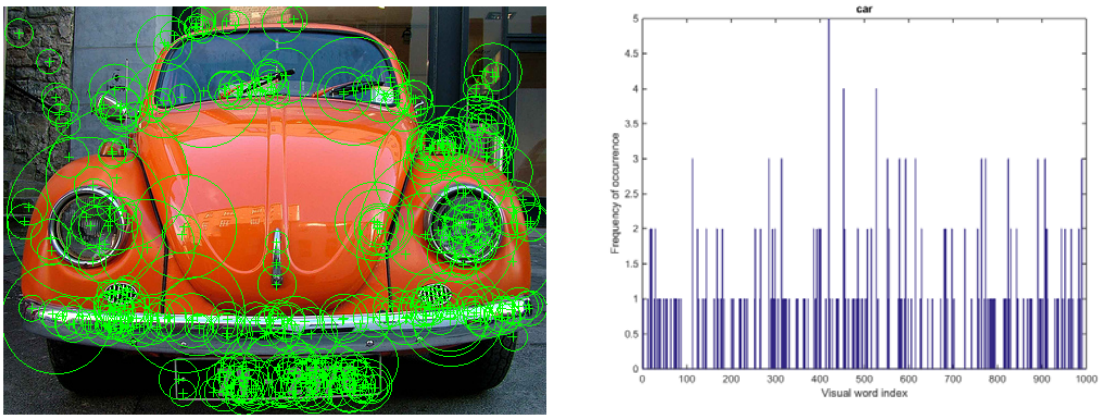
The term score $s = \sum_{i=1}^{M} s_i$ regulates the dominant context after every draw, and the integral stands for the recursive sampling from time *t=0* until all sample are drawn at *t=1*.

Integral in Equation (10) is NP-hard because of the fractional exponent and cannot be computed in polynomial time. Using binary form ($\Lambda(c_i, Z) = 1$), Equation (10) could be approximated and transformed to a polynomial once the context values are scaled to an integer by dividing to the greatest denominator.

5.6 Image Representation and Classification

The first step of training the object classifiers is extracting visual descriptors and building the vocabulary. Set of cropped images of individual objects included in training images were created using bounding box information specified in annotation meta data. Visual appearance features are obtained from extracted objects of training images using SURF descriptors of square 64x64 pixel blocks (see Figure 16-a). For each object, top m strongest features descriptors are selected and normalized across entire training set. The value of m is proportional to number of usable features obtained from weakest object images. Weakest object class usually has poor quality training images and feature descriptors have low SURF score.

To build BoF representation of the objects, feature descriptors are quantized into vocabulary sizes of V visual words. Each object is then encoded into histogram of V visual words which is used to train each individual detector. Training dataset selected contained extensive set of annotation that was also used in building of



a) SURF feature-points b) encoded visual word

Figure 16. Encoding car image into visual word frequencies

high-ordered location, scale and co-occurrence contextual relations among objects (see section 5.7.3 for details).

5.7 Experiments

For evaluation, results of both methods are compared to the state-of-the-art supervised model of sLDA.

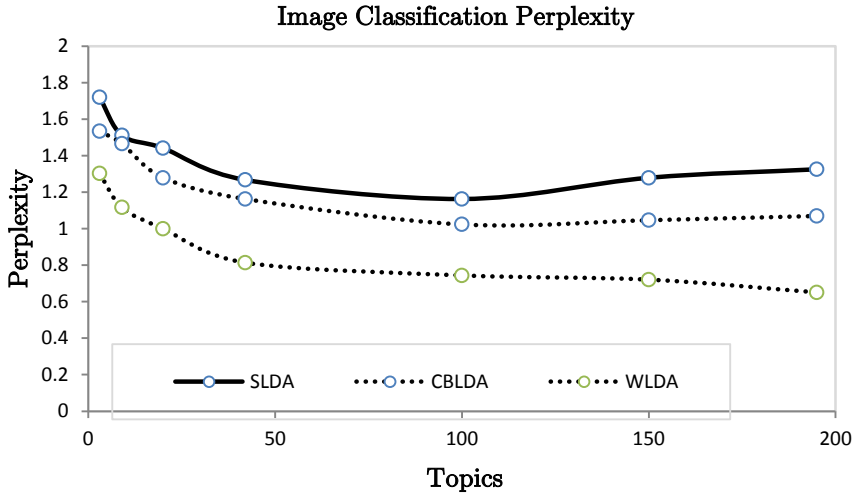
5.7.1 Datasets

Learning contextual relations requires large enough samples with higher-over relations. In our experiments, we use the SUN397 dataset of the label-me project.

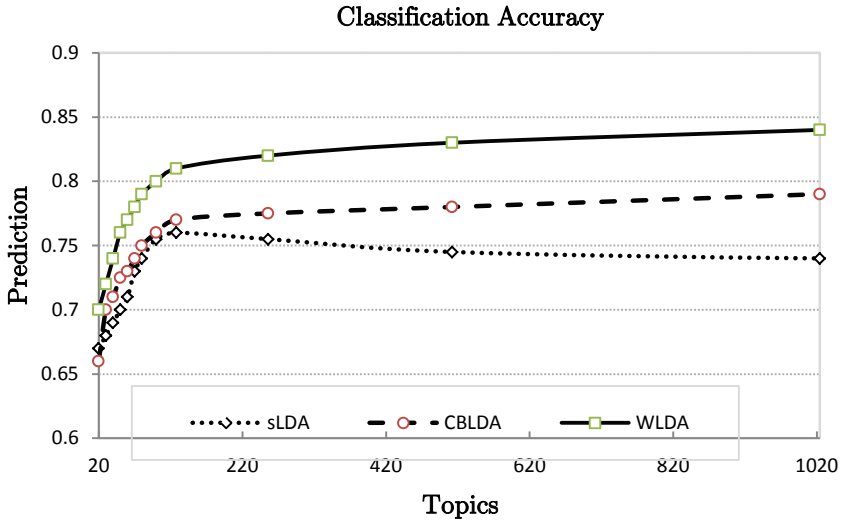
Images of this dataset are highly populated and are more suitable for learning contextual relations. Training samples were randomly selected from 8 scene categories with average nine object types and 16 object instance per image. Images were split, and 30% was allocated for training after they were normalized and preprocessed for all object categories examined. The remaining 70% of the images were used as unlabeled test images to evaluate the performance of our model.

5.7.2 Metrics and Parameters

The models were calibrated using various parameters that maximized their performance. Vocabulary size was determined empirically based on the number of



a) Perplexity of annotation (lower number is better.)



b) Prediction accuracy for 1020 topics(higher rate is better)

Figure 17. Perplexity & prediction for sLDA, xLDA-bin, and WLDA.

the objects and variation in the training samples. Vocabulary cardinality was set to 1000 visual words and selected 90% of strongest features. A sample object's visual words representation is shown in Figure 16 (image on the right) with $|V| = 1000$.

Perplexity is used as a measurement of the accuracy of object classification results and is calculated for entire dataset as shown in (11). The perplexity is a decreasing function of the log-likelihood which implies that lower number equals higher accuracy and better model performance in explaining the data.

$$Perplexity = exp\left(-\sum_{d=1}^{D}\sum_{1}^{n} log\, p(x_i) \Big/ \sum_{i=1}^{n} m_i\right) \tag{11}$$

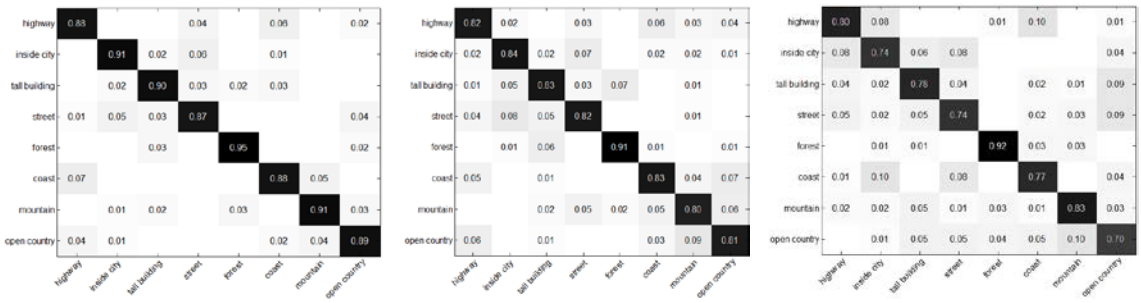
where x_i is observation with m_i features.

Perplexity results shown in Figure 17 illustrates that WLDA model provides superior predictive distribution compare to other two models. By allowing the annotation assignment to different topics, the WLDA gives the best performance in object localization and correctly labels most of the example pictures. It can assign each region to a different cluster, and the final distribution over visual words reflects the clusters which were allocated to the image areas. Over-fitting did not happen until 100 topics for sLDA and xLDA-bin and 150 in the case of WLDA. The more efficient discovery of hidden features is the result of the discriminatory role that context is playing in the topic assignments.

5.7.3 Contextual Framework and Image Classification

A meaningful contextual relevance score requires large enough scene objects with ternary or more interactions. We build our contextual model using the method described in 4.4 and based on ternary relationships. Using Equation (4) combined semantical layer relationships in combination with spatial and scale statistics.

Selected spatial context is a relative vertical alignment of the centroids of the object bounding boxes. Vertical alignment of objects is determined by comparing their individual vertical location calculated as $L_i^v = \frac{b_i^v}{b_i^h} H_i$. In this formula b_i^v is



a) sLDA b) xLDA-bin c) WLDA

Figure 18. Scene classification confusion matrix.

vertical coordination of bounding box centroid, b_i^h is bounding box height and H_i is user defined height of the i^{th} object category which is obtained from training dataset.

Scale context is determined by transformation of bounding box coordinates as $L_i^S = \frac{H_i}{b_i^h}$ to 3D co-ordinates on which a horizontal plane at intersection of ground and sky is used to calculate the relative size of the objects in this plane. Normal distribution is used to model both scale and location contexts.

5.7.4 Evaluation of Methods

To verify the effectiveness of presented algorithms, the performances of XLDA-BIN and WLDA were compared with sLDA (Wang, Blei, & Fei-Fei, 2009) annotation and classification results. All of the models were trained using samples of 8 different scene categories with average nine object classes per image and 100 images instances per object class. Image size constraint was set to minimum 250-pixel dimension, and all of the samples were scaled to minimum 250 pixels at the lowest dimension. All algorithms were executed at various iterations for E-step and M-Step with K =1024 and used the inference algorithms described in (Zolghadr & Furht, 2016a) for object label prediction. Figure 18 shows the performance improvement of the object recognition in WLDA compare to sLDA

Table 2. LDA object detection performance comparison

Object	sLDA	xLDA-bin	WLDA
Bed	0.63	0.69	0.72
Bicycle	0.74	0.79	0.88
Cabinet	0.78	0.83	0.86
Car	0.74	0.82	0.86
Keyboard	0.82	0.82	0.85
Monitor	0.77	0.83	0.88
Street sign	0.83	0.8	0.91
Table	0.7	0.81	0.89

and xLDA-BIN using Sun397 dataset. Based on the discriminatory role of context, a higher degree of disambiguation has been achieved in class label assignment of the poor quality image. This resulted in higher percentage of correct localization and classification in both WLDA and xLDA-BIN.

Figure 18 illustrates scene classification performance of the three models shown as confusion matrix. These results demonstrate that scene classification performance has increased slightly in xLDA-BIN and significantly more in WLDA. This significant increase in classification accuracy is the product of better topic selections by using context.

Topics that are semantically more consistent with the overall meaning of the scene are more predictive of scene class. Similar performance indicates that, on average, the number of hidden factors used to model a particular image is adequate to model its labels. Our object recognition results in Figure 19 shows significant improvement over sLDA and our previous experiments using one type context. The higher performance of WLDA is attributed to accurately capturing of

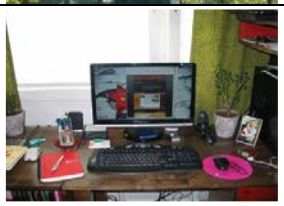
	Ground Truth	Annotation Results
	Sky, Bicycle, Water, Ship, Boat, Sand	**sLDA**: Sky, Water, Bicycle, Building **xLDA-bin**: Sky, Bicycle, Water, Boat **WLDA**: Sky, Bicycle, Water, ship, sand
	Car, Road, Sky, Grass, Cloud	**sLDA**: Car, Road, Grass, Water **xLDA-bin**: Car, Road, Sky, Grass **WLDA**: Car, Road, Sky, Grass
	Table, Chair, tree, Plates, shrubs	**sLDA**: Table, Chair, Grass **xLDA-bin**: Table, Chair, Tree **WLDA**: Table, Chair, Tree. Plates
	Street Sign, Tree, Stop Sign, Sky, Lamp Post	**sLDA**: Stop Sign, Tree **xLDA-bin** Street Sign, Stop Sign, Tree **WLDA**: Street Sign, Stop Sign, Tree
	Keyboard. Monitor, Mouse, Planter, Photo, Note pad, Window, Desk	**sLDA**: Keyboard, Monitor, Mouse, Plant **xLDA-bin**: Keyboard, Monitor, Mouse, Planter, Photo, Desk **WLDA**: Keyboard, Monitor, Mouse, Planter, Photo, Window, Desk

Figure 19. Annotation results for test images of Sun397 dataset.

contextual relations when using high-order context as score bias term during recognition.

Image annotation sample results are illustrated in Figure 19. As demonstrated in these results, both context-based models that were employed have less miss-labeled objects which are the result of contextual refinement performed using relevance score at optimization step. Both XLDA-BIN and WLDA predicted more accurate labels where there are stronger contextual cues in the images.

5.8 Conclusion

In this study, two generative models were evaluated and compared to the state-of-the-art model. The WLDA context-based model was incorporated with high-ordered and multi-source contexts to examine the influence of the context on image and performance. The results show significant improvement in object label prediction and scene categorization performance compare to the state-of-the-art methodology.

The context-based WLDA improved the classification and annotation by incorporating contextual characteristics of data which reflects more semantical coherency of the topics. Improved accuracy achieved by reduction of mislabeled and unlabeled objects due to the incorporation of semantic space in the classification process. This performance gain however incurred additional computation cost when complex high order contextual analysis performed. Among all kinds of contexts considered in this study, spatial and scale contexts played a significant discriminatory role while LDA process well-captured co-occurrence. WLDA improved the over-fitting problem associated with generative topic modeling methods but did not eliminate it.

CHAPTER 6: CONDITIONAL APPROACH TO SCENE UNDERSTANDING

6.1 Introduction

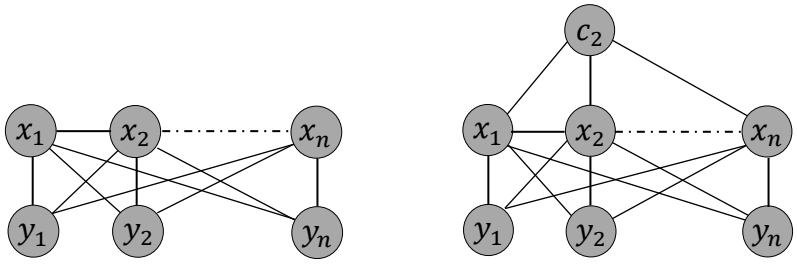
Conditional random fields (CRF) (Lafferty, 2001), a discriminative framework, is used to incorporate various features in a single model. This allows modeling intrinsic and extrinsic structure of images for a better understanding of their underlying concepts. Given observed variable X, CRFs model the conditional distribution of Y given X to encode complex dependencies of Y on X. In this dissertation we present a CRF that build using visual features and is conditioned and constrained on context. We call this model Context-based CRF (CBCRF) and combines appearance descriptors, contextual relations and layout structure of the objects likely to be present in that scene category.

This framework combines contextual consistency among the composing elements of an image introduced in section 4.4 to influence classification refinement. Using graph cut optimization contribution of an object type to the overall semantical meaning of the scene for a given context is maximized.

In this dissertation , we use context-based LDA for topic discovery. LDA maps image simplex to topic simplex in a context-aware framework to maximize inter-topic compatibility while reducing dimensionality. We show that exploiting high-order relations of topics can improve object detection, scene classification and can be beneficial to other applications such as detection of the out-of-context or black-boxed object.

6.2 Conditional Random Fields

A more advanced representation of contextual relationships can be formulated using Conditional Random Field (CRF) (Boykov et al., 2001). A conditional random field (CRF) can be viewed as a discriminative undirected probabilistic graphical model (Clifford, 1990), and is used to learn the conditional distribution over the set of class labels globally conditioned on X, the random variable representing observation sequences (shape, color or description of the domain). Formally, we define $G = (V, E)$ to be an undirected graph such that there is a node $v \in V$ corresponding to each of the output joint random variables representing an element of the set $\{y_v\}_{v \in V}$ where $Y = (y_1, y_2, \ldots, y_N)$. If each random variable y_v obeys the Markov memory-less property conditioned on X with respect to G, then (X, Y) is a conditional random field. The structure of graph G is arbitrary, provided conditional probability distribution of y_v given its adjacent label sequences is independent of the rest of nodes in the graph (see Figure 20).



a) Fully Connected CRF b) Context-based CRF

Figure 20. Graphical model for fully connected & context-based CRFs

In object recognition, X corresponds to image data sample (i.e., pixel, super-pixel, and object) or feature descriptor, Y is corresponding labels and c is

contextual feature descriptor. Conditional probability distribution of a CRF is formulated as follows (Liu, Huang, & Ma, 2016):

$$P(y|x) = \frac{1}{Z(x)} \prod_{A \in \Lambda} \Psi_A(x_A, y_A) \tag{12}$$

where $\Lambda \in G$ is set of all possible maximal cliques, $A \in \Lambda$ is maximal clique and Ψ_A is non-negative unary potential defined over A. Z is called partition function or normalization factor and has the form (Liu, Huang, & Ma, 2016):

$$Z(x) = \sum_{y} \prod_{A \in \Lambda} \Psi_A(x_A, y_A) \tag{13}$$

Ψ_A is unary potential and is defined as (Liu, Huang, & Ma, 2016):

$$\Psi_A = exp\left\{ \sum_{k} \psi_{AK} f_{AK}(x_A, y_A) \right\} \tag{14}$$

where f is set function defined on the unary potential Ψ. Taking a negative log of both side of Equation (12) gives energy function:

$$E(y|x) = -log\big(P(y|x)\big) = \sum \psi_u(y|x) \tag{15}$$

where ψ_u is the unary potential and represents relations between input local features and their labels. To maximize the conditional distribution $P(y|x)$, energy function in (15) must be minimized.

6.3 Context-based CRF Model

Figure 21 illustrates a graphical representation of the context-based CRF. To formalize definition of context-aware conditional random field, let's assume a random field Y defined over set of variables $\{y_1, \cdots, y_K\}$ to represent labels of all

objects. Domain of each variable y_i is $\mathcal{L} = \{l_1, \cdots, l_K\}$ which is set of all possible labels. Let $X = \{x_1, \ldots, x_D\}$ be the set of images in our dataset, $R_i = \{r_1, \cdots, r_n\}$ be set of visual words of i^{th} image, $C_i = \{c_1, \cdots, c_K\}$ be image sub-region category labels representing objects, and $S_i = \{s_1, \cdots, s_K\}$ be set of contextual scores (Equation (5)) under which independent measurement of semantic relevance calculated for detected objects in i^{th} image. Each image is composition of arbitrary number of object instances in the same scene category.

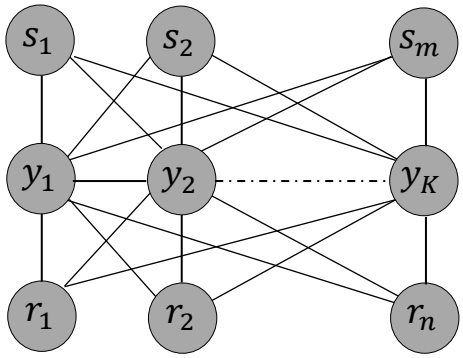


Figure 21. Graphical representation of the context-based CRF model

The use of a conditional random field allows us to incorporate appearance based descriptors, layout, and location cues in a single unified model. Our context-based CRF approach aims to find optimal configuration $Y = \{y_1, y_2, \ldots, y_n\}$ which is characterized by Gibbs distribution $P(Y|R)$:

$$P(Y|R,\theta) = P(Y|\, R,S,\theta)\; P(S|R,\theta) \tag{16}$$

where θ is model parameters, S is context and $E(Y|S,R,\theta)$ is the probability of the labeling configuration Y given visual words conditioned on the context the conditional random field defined as:

$$P(Y|\, R,S,\theta) = \frac{1}{Z(Y,S)} \exp\big(-E(Y|R,S)\big) \tag{17}$$

where z is normalization partition function. Our model is fully connected CRF with unary, pairwise and high-order potentials with following Gibbs Energy:

$$E(Y|\,R,S) = \sum_{n\in N} \psi_u(y_n) + \sum_{(i,j)\in P} \psi_p(y_i, y_j) + \sum_{i\in H} \psi_h(y_i) \tag{18}$$

where N, P, H are number of candidate objects in the image, number of pairwise and high-order cliques respectively.

6.4 Unary Potential

The model appearance, affinity, and shape are modeled using unary potential ψ_u. Unary potential is the most important potential and is sensitive to mislabeling as a result of initialization. By incorporating context the classification of objects is influenced by dominant context and hence initially mis-classified labels can be refined. Unary potential is defined as:

$$\psi_u(y_n) = p(Y|R,S) \tag{19}$$

where $S_i = \{s_1, \cdots, s_m\}$ is all contextual scores in image graphs and the term $p(Y|R,S)$ is probability that object i^{th} would be assigned label y given the set of contextual scores for the objects.

We introduced Wallenius Latent Dirichlet Allocation (WLDA), a generative process for object localization in Section 5.5.2. This process partitions an image into related groups of visual words which represent candidate objects and assigns best annotation label to them. Each label is associated with image feature data as response variable which is influenced by contextual constraints as bias weight parameter in Wallenius distribution.

6.5 Pairwise Potential

The pairwise term $\psi_p(y_i, y_j)$ reinforces contextual compatibility between label assignments of the neighboring object. It predicates on the assumptions that

objects adjacent to each other are more likely to have the compatible labels and be semantically related. Probability of label assignment follows the given context. This potential takes the form of Potts model to penalize semantically incompatible labels:

$$\psi_p(y_i, y_j) = \begin{cases} 0 & if\ y_i = y_j \\ \lambda_p exp\left(-|l_i - l_j|^2\right) & otherwise \end{cases} \quad (20)$$

where $l_n = p(y_n|S)$ and λ_p is parameter whose value is learned from training data. This potential has shrinkage bias which means it only enforces label consistency in adjacent objects.

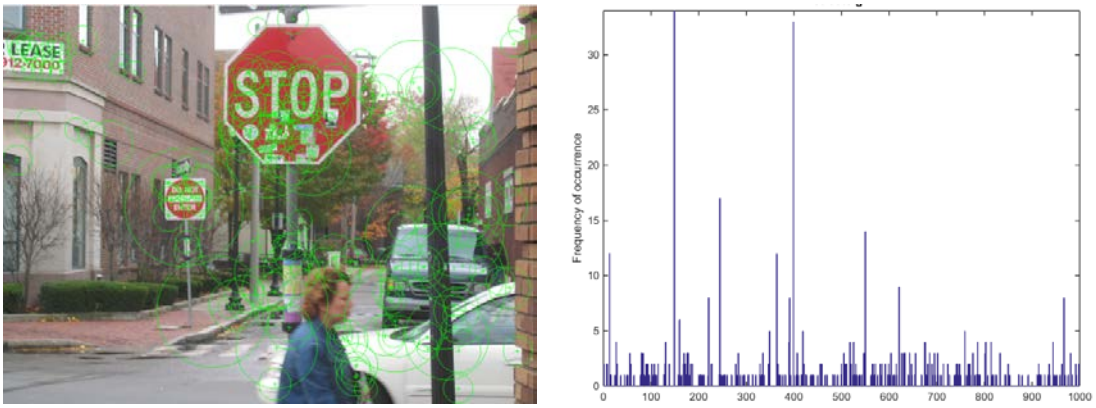
6.6 High-ordered Potential

The high-order potential $\psi_h(y_i)$ is defined to maximize contextual consistency and compatibility of the label assignment in neighborhood of an object. To achieve this, objects in an image are grouped in semantically compatible and consistent cliques. A penalty is applied to non-relevant ones to disassociate them from clique. Consistency of the clique is measured using the variance of unary feature response evaluated on all objects in that clique as follows (Kohli et al., 2009):

$$\vartheta_C = exp\left(-\frac{\|\sum_{c\in C}(I_c - \mu)^2\|}{|C_L|}\right) \quad (21)$$

where C_L is the clique, $|C_L|$ is cardinality of that clique, $I_c = p(y_n|C)$ and $\mu = \sum_{n\in C} p(y_n|C)/|C_L|$. Given the CRF model in Equation (18), high-order potential is defined as following (Kohli et al., 2009):

$$\psi_h(y_i) = \begin{cases} N\lambda_h\vartheta_C\frac{1}{Q} & if\ N \le Q \\ \lambda_h\vartheta_C & otherwise \end{cases} \quad (22)$$



a) Original image

b) Encoded image visual words histogram for the image shown on the left.

Figure 22. Encoding a sample image in visual word frequencies.

where N is number of elements in the clique y_i with label assignment that are inconsistent with dominant label in that clique and λ_h is model parameter which is obtained during the training. Consistency of this potential is controlled by threshold parameter Q which defines a cut-off point where from that point stronger penalty is imposed on very semantically consistent cliques. With the objective of finding the most probable labeling configuration that maximizes the conditional probability of Equation (17), graph-cut optimization algorithm (Boykov & Veksler, 2006) is applied to get the optimal configuration $y^* = [y_1^*, y_2^*, \dots, y_n^*]^T$.

$$y^* = arg\,max\,P(y|R) = arg\,min\,E(y) \tag{23}$$

where y_n^* is unit basis vector that represents the result of object localization for n^{th} object in the image. Contextual scores are used during the optimization to eliminate false positives and keep correct detections.

6.7 Experiments

Object recognition algorithm was evaluated on a subset of SUN397 datasets with 2152 scene images randomly selected as training set and 1010 scene images chosen as a test set from 100 object categories. The metadata of labeled images was

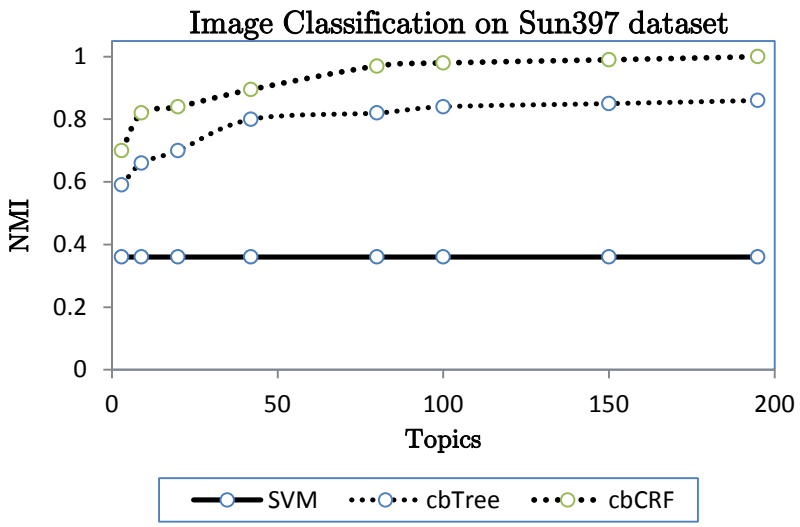


Figure 23. Object detection performance using NMI (1.0 is most accurate)

used to extract crop images of objects according to their bounding box information. In pre-processing phase, images were scaled to meet a minimum dimensional constraint.

6.8 Training

Image feature space was represented as BoF. Each codeword in the codebook is a visual appearance feature which was constructed based on SURF algorithm.

SURF feature-points were calculated from 64x64 blocks for image objects and transformed into descriptors. Top m strongest SURF descriptors were selected and normalized across entire training set. The value of m is calibrated empirically. Selected descriptors were then quantized into vocabulary sizes of V visual words using K-means clustering algorithm. Figure 22-(b) illustrates BoF representation of an image in Figure 22-(a) encoded as histograms of visual words (V =1000) which is used to train our model.

There are two sources of parameters in this study. The first one is the LDA parameter set which is learned the way is described in section 5.5. The second set

of parameters is the CRF parameter set $-\lambda_p, \lambda_h$. These parameters were all learned from the training set using the same method introduced in (Kohli et al., 2009).

6.9 Evaluation Methods

For evaluation of context-based CRF framework, multi-class support vector machine (SVM) (Desai et al., 2009) classification method was used as a baseline and compared to the state-of-the-art tree-based contextual model (Choi et al., 2012a) using the code provided at their site.

6.10 Metric

Normalized mutual information (NMI) is a metric used to evaluate the performance of clustering and to measure how well objects in test images are assigned to object categories. NMI is a number between 0 and 1 and with 1 being perfect object label assignment and is calculated as follows (Fink & Perona, 2004):

$$NMI = \frac{\sum_{h,l} |x_{h,l}| \, log\left(\frac{|X|.|x_{h,l}|}{|x_h|.c_l}\right)}{\sqrt{\left(\sum_h |x_h| \, log\left(\frac{|x_{h,l}|}{|X|}\right)\right) \sum_l log\left(\frac{c_l}{|X|}\right)}} \tag{24}$$

where X is set of images, x_h is set of images in class h, $x_{h,l}$ is number of images that are member of both classes h and l and c_l is images labeled as class l.

Figure 23 illustrates object detection NMI that was applied to the models in these experiments. The results show context-based CRF model performs better in various topic sizes of K. These experiments also demonstrate that larger number of the topics have little impact on the object detection performance but has considerable computational cost and performance degradation as the number of topics increases. When a scene contains less than K objects, the absent object categories will have very few or no members such that the impact will be small

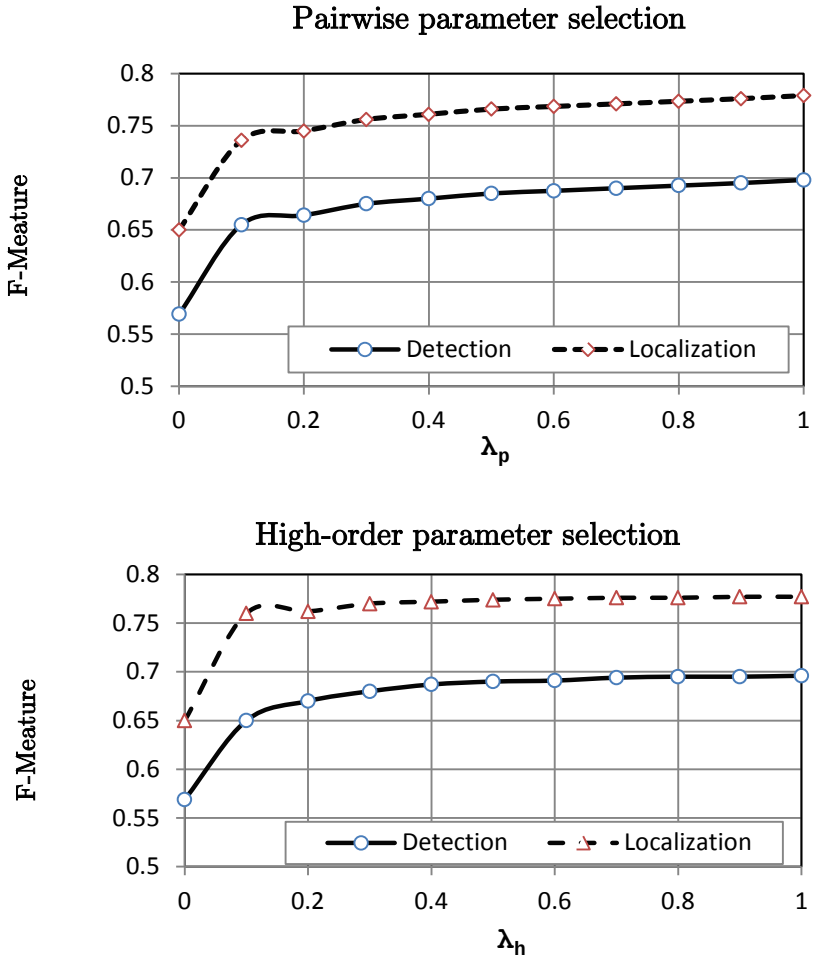


Figure 24. Parameter selection for pairwise and high-order potentials.

enough to be neglected. The optimum value of K is determined empirically and set to 135.

To evaluate performances of our framework for localization and presence F1-Measure was used which is a balanced score between precision and recall (F1) as follows:

$$F1 = \frac{2 \times precision \times Recall}{precision + Recall} \tag{25}$$

Classification performance was evaluated using objects labels in Ground-truth. Figure 24 shows how this metric was used in finding optimum parameter values for pairwise and high-order potentials.

6.11 Model Parameters

Parameters that have an influence on the distribution of topics in potentials were also investigated. Two main parameters require calibration, pairwise (λ_p) and high-order parameter (λ_h).

The tuning result on SUN397 is given in the top chart of Figure 24. Parameters λ_p and λ_h varied independently from 0 to 1 with interval 0.1 to pick the optimum value. As is illustrated in Figure 24, the performance improves as the value of the parameters increases. Slightly sharper gain in high-order potential than pair-wise demonstrates effectiveness of this potential.

6.12 Result

To build the framework, a graph was constructed to maximize contextual consistency. First scale and location context scores were calculated for all object pairs (ω_{ij}) in that image using Equation (4). Pairwise relations with $\omega_{ij} > 0$ were added to the graph and others were ignored. Next, high-order co-occurrence, location and scale contexts were computed for all cliques combinations (i.e. $\omega_{12..k}$) . Corresponding edges were added to the graph is they satisfied the positive relationship condition of Equation (4). Individual contextual scores were computed by summing up scores of each node in the graph for all relationships and applying softmax using Equation (5).

Table 3 shows the localization and presence improvement over baseline detector algorithm and the state-of-the-art context model (tree context). Both cases show significant improvements as a result of the new context discriminatory power.

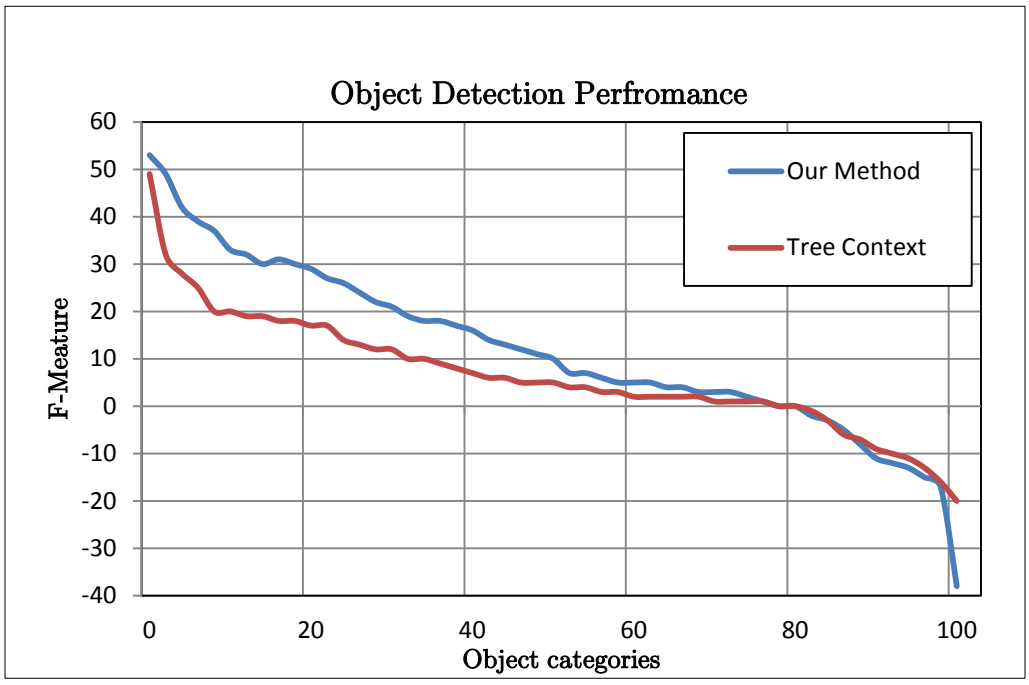


Figure 25. Object detection performance comparison.

Table 3 lists the comparisons between baseline detector SVM, tree-based context, and our framework. From the table, we see our framework produces the best performance in both object localization and presence. Table 4 shows some examples of results in which context constraints are strictly enforced to facilitate the contextually consistent detections. Results shown illustrates that context-based CRF has improved compare to the performance of the SVM and CRF in the classification of the objects.

Performances of proposed framework for object detection is illustrated in Figure 25 using F-measure. Our model shows improvement over the tree context model for most object categories by reducing number false positives as a result of class disambiguation and enforcement of contextual constraints in higher-order. In this section, we presented a discriminative model that combined the power of a generative model as unary potential and used contextual scores of section 4.4 to

encode pair-wise and high-order semantic contexts. We showed how to encode the high-order relationship among objects and build robust models to enforce location, scale, and semantical constraints. We compared our framework with another context-based model which employed similar sources of contexts in pairwise relations. Our results demonstrated that our framework outperformed the current state-of-the-art context-based object localization methods. Our generative process implemented a true context-based approach where the context was directly applied to classification problem as unary potential. We showed an inference method to solve the intractability problem of the WLDA to a solution that could be solved at the polynomial time. We then applied our framework to distinguish the contextual consistency of the candidate objects using various contextual cues. Figure 26 shows experiment results on test images of sample categories which were examined.

Table 4. CRF object detection performance comparison

	SVM	Tree Context	Context-based CRF
Bed	0.57	0.68	**0.78**
Bicycle	0.65	0.72	**0.88**
Cabinet	0.62	0.74	**0.89**
Car	0.63	0.8	**0.91**
Keyboard	0.62	0.66	**0.87**
Monitor	0.51	0.69	**0.89**
Street sign	0.67	0.71	**0.92**
Table	0.44	0.65	**0.90**

It is clear that CBCRF has improved object detection performance over LDA by forming more coherent graphs that project stronger and more robust inter-object relations by further optimizing the models using CRF. We also see that our choice of context improved the semantical structure of the interactions of the

Figure 26. Results of WLDA object detection performance.

objects in the scene. Using the discriminatory power of high-range, high ordered contextual semantics; more false positives were identified and corrected.

CHAPTER 7: OBJECT SALIENCY AND SCENE UNDERSTANDING

7.1 Introduction

"As I ate the oysters with their strong taste of the sea and their faint metallic taste that the cold white wine washed away, leaving only the sea taste and the succulent texture, and as I drank their cold liquid from each shell and washed it down with the crisp taste of the wine, I lost the empty feeling and began to be happy and to make plans." (Hemingway, 2008)

In this passage, Hemingway is describing the experience of tasting oysters and drinking wine using simple words that are easy for a person to visualize the scene and internalize his experience. Human scene analysis and synthesis facilitate extremely sophisticated interpretation and visualization capabilities. A person can perceive semantical composition of a scene in a glance and provide a precisely accurate description of the content minus the hidden details. The most impressive aspect of the human recognition system is the ability to compare similarities between scenes and detect irregularities within scenes incredibly fast (Duncan & Humphreys, 1989).

Inspired by the impressive performance of the human visual system, computer vision artificial intelligence researchers are working on the opportunities to achieve similar performance. As a result, extensive work on the salient object recognition (Elazary & Itti, 2008; Walther & Koch, 2006), context based scene understanding (Farhadi et al., 2010; Kulkarni et al., 2013; Y. Yang, Teo, Daumé, & Aloimonos, 2011) and object-based scene taxonomy methods(Berg et al., 2012; Cai & Leung, 2011; Elazary & Itti, 2008; Shafieyan, Karimi, Nasr-Esfahani, & Samavi, 2014),

image clutter (Asher, Tolhurst, Troscianko, & Gilchrist, 2013; Bravo & Farid, 2008; Henderson, Chanceaux, & Smith, 2009), object saliency (Elazary & Itti, 2008; Pan, Wang, & Jiang, 2015) and connection between internal perception and external objects (Fleming, 1977) has been performed.

An important common objective is to obtain the most semantically accurate meaning of the scene. However, is the human perception of the scene semantical structure and relations always accurate? Which features better represent the most semantically relevant structure and interactions in an image? Researchers on human brain suggest that parallel interpretation of a scene occurs simultaneously in both brain and visual cortex. Mumford (Mumford, 1992) claims brain builds an abstract and coarse image of the subject in the higher cortical area while the lower part is processing more concrete details. Processing attempts to create a complete picture by fitting the initial abstract image on the captured details until clear picture perceived (Kok & de Lange, 2014).

Inspired by visual processing in humans, a context-based framework is presented for studying the influence of various features on semantic understanding. The objective is to investigate the impact of the vital information such as the conceptual structure of preliminary knowledge of the scene. Our hypothesis is that initial perception of the scene can be refined interactively by fusion of feature-level, mid-level, and object-level semantics. In other words discovery of new detail information about an image unveils deeper explanation which refines initial understating. A multi-layered recognition process is proposed for measuring feature specific conceptual improvement.

Many of state-of-the-art scene understanding frameworks lack flexible inference to recover from initialization errors (Nematollahi & Zhang, 2014). A feed-back and feed-forward mechanism between the layers that conveys sensitive information between classifiers at each level is instrumental in boosting their labeling

confidence to higher levels of accuracy based on the synchronizing dominant context of other layers.

A successful video annotation framework, however, must go beyond detecting objects in scenes and capture attributes such as events, actions, subject's pose, expressed or implied intentions and clear description of appearance. There are many studies on locating salient or relevant visual information (Dubey, Peterson, Khosla, Yang, & Ghanem, 2015; Elazary & Itti, 2008; Isola, Xiao, Torralba, & Oliva, 2011). Traditionally study of video annotation focuses on the salient objects and point of interest, with little attention to salient semantic of a scene. Occasionally there is an object in odd relations with respect to other objects or violation with scene semantics that must also be narrated, so it is critical to determine what needs to be annotated and to provide a most accurate description.

While each of attributes mentioned above is well studied for specialized domains like news, sports, surveillance feeds, building a general purpose framework remains a challenge due to a large number of possibilities in visual appearances and variations in their relations. For instance ImageNet dataset (Jia Deng et al., 2009) consists of over 21,814 event classes (14 million image) but state-of-the-art event detectors participating in Large Scale Visual Recognition Challenge pre-train only on 10% of that dataset with 1,000 class (i.e. 1.2 million images) (L. Jiang et al., 2015; Lai, Pan, Liu, & Yan, 2015; Lai et al., 2015; Mettes, Koelma, & Snoek, 2016; Szegedy et al., 2014; Xu, Yang, & Hauptmann, 2014). The pre-training over this collection using deep convolutional neural network starts by extracting low-level features of frames and pooling for video representation (Russakovsky et al., 2015; Seddati, Dupont, & Mahmoudi, 2015; Simonyan & Zisserman, 2014).

Scene semantics are embedded in all levels with dependence on low-level inter-pixel interactions (i.e., with pair-wise and long-range spatial association) and mid-level inter-object relations (i.e., co-presence and spatial relationship) and higher

cognitive inference. The high dimensionality of low-level features makes recognition of significant numbers of the object for the state-of-the-art approaches virtually impossible. Using deep neural networks running on a vast number of nodes to label the huge set of objects and their relations is unrealistically expensive and therefore not unfeasible for most research labs.

The objective of this study is to investigate the gap between computer vision scene understanding and human perception and interpretation of the underlying scene concepts. By exploring the correlations of object interactions and high-level scene semantics, we aim at locating objects that play a pivotal role in representing scene concepts.

To measure this gap, a new training dataset was built by human subjects in a weakly-supervised and assisted environment using interactive process. For each image in pre-selected scenes category (i.e. street, dining room) users are asked to annotate their perception of the scene taxonomy and salient objects participating in those interactions ranked by the importance of their role. Users also are given system-generated annotation and other annotation from other participants and asked to revise their previous input until finalized. Our primary objective is to uncover the role of semantical object saliency in the way human perceive the meaning if a scene. We are also interested in novel high-level image semantics representation to exploit more intrinsic properties such as object relations and attributes of the events or concepts. We start at higher level semantics and work our way down to image primitives.

In summary, our contribution as follows

- Introducing semantical object saliency measurement for each scene category. We show the relationship of such metrics with the consistency of scene high-level concepts and mid-level interactions.

- Also, we present a framework that incorporates object saliency with high-level descriptors in a probabilistic model as before maximizing the scene contextual consistency and consequently objects recognition optimization.
- We apply high-level semantics learned in this model to complement the mid-level and low-level descriptors that previously studied.

7.2 Object Recognition

In this section, we describe a model based on visual features and contextual features as refinement. Then we measure the amount information that was added by applying these contextual features.

7.2.1 Unsupervised GMM

A generalization of scene semantics using low-level features can be expressed as Gaussian Mixture Models (GMM) with M distributions as sum of M components as following(Hadfield & others, 2010):

$$p(X|\lambda) = \sum_{i=1}^{M} w_i f(x_i|\mu_i, \Sigma_i) \tag{26}$$

where λ is set of model parameters $\lambda = \{w_i, \mu_i, \Sigma_i\}$, X is a d-dimensional data feature vector, $w_{i:M}$ are the mixture weights s.t. $w_i \geq 0 \wedge \sum_{i=1}^{M} w_i = 1$, μ_i is mean, Σ_i is covariance and $f(x|\mu_i, \Sigma_i)$ are the Gaussian densities of each component and can written as Gaussian function of the form (Hadfield & others, 2010)

$$f(x|\mu_i, \sigma_i) = \sum_{i=1}^{k} \frac{1}{\sqrt{(2\pi)^{d/2}|\sigma_i|^{1/2}}} exp\left\{-\frac{1}{2}(x-\mu_i)^T {\sigma_i}^{-1}(x-\mu_i)\right\} \tag{27}$$

Given a set of feature vectors x , GMM is built for each concept using maximum likelihood estimation (Ma & Gao, n.d.)

$$\theta_{ML} = arg\max_{\theta} L(\theta|x_{1:n}) = arg\max_{\theta} \sum_{i=1}^{n} log\big(f(x_i|\theta)\big) \tag{28}$$

Concept classifiers must first calculate $\theta_{ML} = \theta_c$ and then calculate $f(x|\theta_c)$ to verify presence the of the concept c in the image. Parameter estimation aims at finding model parameters that maximize the likelihood of GMM given the sequence of independent observations $x = \{x_1, x_2, \dots, x_T\}$. Likelihood of GMM is

$$p(X|\lambda) = \prod_{i=1}^{T} p(x_T|\lambda) \tag{29}$$

Equation (29) is a non-linear function, and maximization is intractable. To find the parameters one possible solution is ML. In an iterative process, a new model $\bar{\lambda}$ is estimated from previous model λ such that $p\big(X\big|\bar{\lambda}\big) \geq p(X|\lambda)$ and continues until convergence. To guarantee monotonic increase of the model's likelihood value the following re-estimation updates are used to estimate the parameters(Ma & Gao, n.d.):

$$\begin{aligned} \bar{w}_i &= \frac{1}{T}\sum_{t=1}^{T} Pr(i|x_t, \lambda) \\ \bar{\mu}_i &= \frac{\sum_{t=1}^{T} Pr(i|x_t, \lambda)x_t}{\sum_{t=1}^{T} Pr(i|x_t, \lambda)} \\ \bar{\sigma^2}_i &= \frac{\sum_{t=1}^{T} Pr(i|x_t, \lambda)x_t}{\sum_{t=1}^{T} Pr(i|x_t, \lambda)} \end{aligned} \tag{30}$$

where the posterior probability for component i is given as

$$Pr(i|x_t, \lambda) = \frac{w_i g(x_t|\mu_i, \Sigma_i)}{\sum_{k=1}^{M} w_k g(x_t|\mu_k, \Sigma_k)} \tag{31}$$

7.2.2 Supervised GMM

Supervised extension of the EM algorithm aims at finding maximum θ_{ML} that maximizes $L(\theta) = logf(X, Y|\theta)$ by iteratively finding a sequence of θ' and $\theta^{"}$such that $L(\theta') > L(\theta^{"})$. EM algorithm is performed in two steps of *Estimation* or calculating $E_{f(X,Y|\theta')}[logf(X, Y|\theta^{"})]$ and *Maximization* of $\theta^{"}$ s.t. $\theta^{"} = argmax\left(E_{f(X,Y|\theta')}[logf(X, Y|\theta^{"})]\right)$ where $X = \{x_i\}$ is a set of unobserved data and Y is observed labeled data used for classification. We also assume each instance of x_i can have a label $c_i \in \{0,1, \dots, k\}$, k is number of Gaussian components and $C = \{c_i\}$ is set of all corresponding labels. To show which component x_i is drawn from, a binary variable z_{ij} is used such that $z_{ij} = 1$ implies j^{th} component is used.

Likelihood function (Fernando, Fromont, Muselet, & Sebban, 2012a) can be written as

$$L_0(\theta) = P(x, z|\theta) = \prod_{j=1}^{k}\sum_{i=1}^{M} z_{ij} w_i f(x_i|\mu_i, \Sigma_i) = \sum_{j=1}^{k}\sum_{i=1}^{M} z_{ij} log[w_i f(x_i|\mu_i, \Sigma_i)] \quad (32)$$

Therefore, for $z_{ij} > 0$ the expectation is computed as

$$E(z_{ij}|x_i, \theta') = \frac{w_i f(x_i|\mu_i, \Sigma_i)}{\sum_{i=1}^{M} w_i f(x_i|\mu_i, \Sigma_i)} \quad (33)$$

Substituting z_{ij} with $E(z_{ij}|x_i, \theta')$ the likelihood can be written as

$$L(\theta, \theta') = \sum_{j=1}^{k}\sum_{i=1}^{M} E(z_{ij}|x_i, \theta') log[w_i f(x_i|\mu_i, \Sigma_i)] \quad (34)$$

A derivative of this likelihood function maximized the function with following parameter updates (Fernando, Fromont, Muselet, & Sebban, 2012b)

$$w_j^{"} = \sum_{i=1}^{n} E\left(z_{ij}|x_i, \theta'\right)/n$$

$$\mu_{i=}^{"} \frac{\sum_{i=1}^{n} E\left(z_{ij}|x_i, \theta'\right).x_i}{\sum_{i=1}^{n} E\left(z_{ij}|x_i, \theta'\right)} \tag{35}$$

$$\Sigma_{i=}^{"} \frac{\sum_{i=1}^{n} E\left(z_{ij}|x_i, \theta'\right).\left(x_i - \mu_j^{"}\right).\left(x_i - \mu_j^{"}\right)^T}{\sum_{i=1}^{n} E\left(z_{ij}|x_i, \theta'\right)}$$

7.2.3 Context-based GMM

To incorporate contextual features, we use Maximum a Posteriori (MAP) estimation. The parameter estimation step begins with probabilistic alignment of the training data into mixture components by computing $Pr(i|x_t, \lambda_{Prior})$ for mixture i in the prior model. To compute desired parameters we need to compute sufficient statistic which are weight $n_i = \sum_{t=1}^{T} Pr(i|x_t, \lambda_{Prior})$, mean $E_i(x) = \frac{1}{n_i}\sum_{t=1}^{T} Pr(i|x_t, \lambda_{Prior})\, x_t$ and variance $E_i(x^2) = \frac{1}{n_i}\sum_{t=1}^{T} Pr(i|x_t, \lambda_{Prior})\, x_t^2$.

Next, we need to update the prior sufficient statistic for mixture i using the sufficient statistics:

$$\begin{aligned} \widehat{w}_i &= [\overline{s_\iota}\, n_i/T + (1 - \alpha_i^w) w_i]\gamma \\ \widehat{\mu_\iota} &= \overline{s_\iota} E_i(x) + (1 - \overline{s_\iota})\mu_i \\ \hat{\sigma}_i^2 &= \overline{s_\iota} E_i(x^2) + (1 - \overline{s_\iota})(\sigma_i^2 + \mu_i^2) - \mu_i^2 \end{aligned} \tag{36}$$

The coefficients $\overline{s_\iota} = \sum_{j=1}^{K} s_{ij}$ are parameters that use contextual scores given in Equation (5) and control contextual improvement between old and new estimates. In these equations $\overline{s_\iota}$ is called *relevance factor* and determines if new parameter is

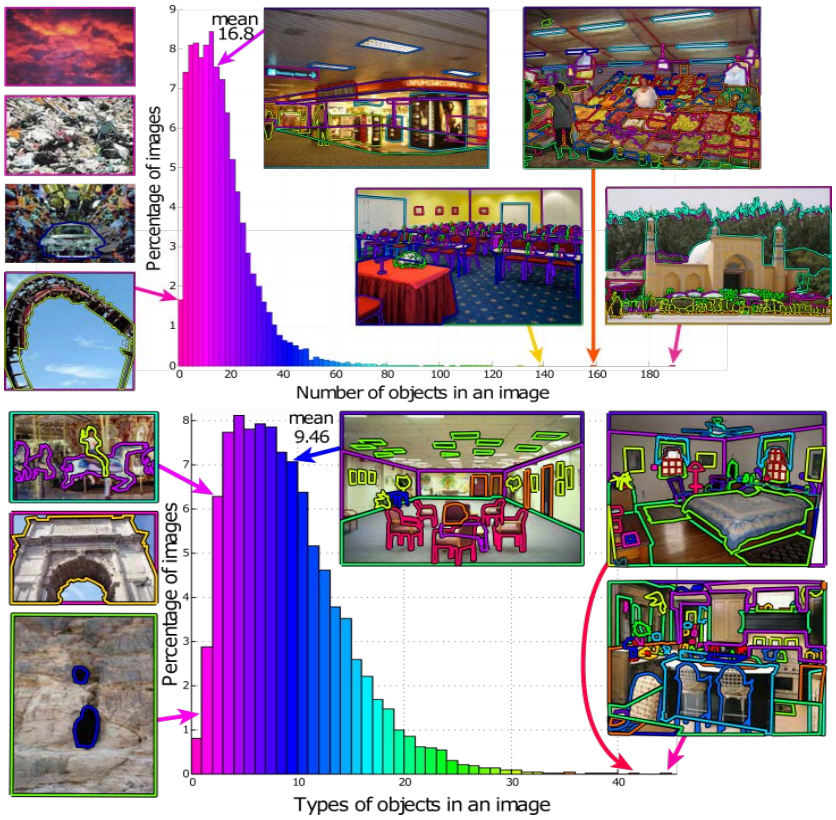


Figure 27. Sun397 database statistics (Xiao et al., 2016).

good enough to replace old one. If $\overline{s_i} \longrightarrow 0$ then old parameter is not going to be replaced and in case of $\overline{s_i} \longrightarrow 1$ then new parameter will be replaced with the old one. The factor γ is computed over all adapted mixture weights to make sure they sum up to one.

7.3 Object Saliency

What makes a visual object salient? Is Saliency in images a subjective criterion or there are semantically standard constraints that can result is constant selection for salient objects in the scene. Let's define salient objects as *the smallest group of objects that produce the most meaning of scene semantics*. The hypothesis is that very popular objects such as windows or walls lack differentiating property, so relative exclusiveness is more descriptive. We also examine various object statistics

such as presence and coverage of an object over the majority of the scene category instances and show it is a good indicator of object semantical saliency.

Let's define *Object Saliency Score* (OSS) as

$$v_{i,j} = \frac{p(f|c_i, w_j)p(c_i|w_j)}{p(w|c_i)p(f|c_i)} \tag{37}$$

where

- $c_i \in C, C = \{c_1, \dots, c_n\}$ is set of all known objects categories in dataset,
- $w_i \in \mathcal{W}, \mathcal{W} = \{w_1, \dots, w_n\}$ is set of all scene categories in the dataset,
- $p(f|c_i)$ (frequency) probability of a frame containing i^{th} object in dataset,
- $p(w|c_i)$ (range) probability that a scene contains i^{th} object in dataset,
- $p(f|c_i, w_j)$ (coverage) probability a frame f contains i^{th} object, given j^{th} scene,
- $p(c_i|w_j)$ (dispersion) probability that i^{th} object appears in j^{th} scene,
- $v_{i,j} \in \Upsilon$ is sparse saliency matrix.

Performance of salient object is measured by comparing highest ranking objects in $v_{i,j}$ to the one picked by human subjects using context-based extension of semantical distance and Kolmogorov complexity (Cilibrasi & Vitanyi, 2007)

$$D(x, y) = \frac{max\{log\, f(x), log\, f(y)\} - log\, f(x, y) . exp(\omega_{xy})}{log\, N - min\{log\, f(x), log\, f(y)\}} \tag{38}$$

where x is saliency of object detected by our framework described by Equation (37), and object y is object picked by user during annotation process, and $f(x), f(y)$ are number of the images that contain object x and y respectively. $f(x, y)$ is number of images with both x and y present, ω_{xy} is CRS of two objects as defined in Equation (4) and finally N is total number of images in the dataset.

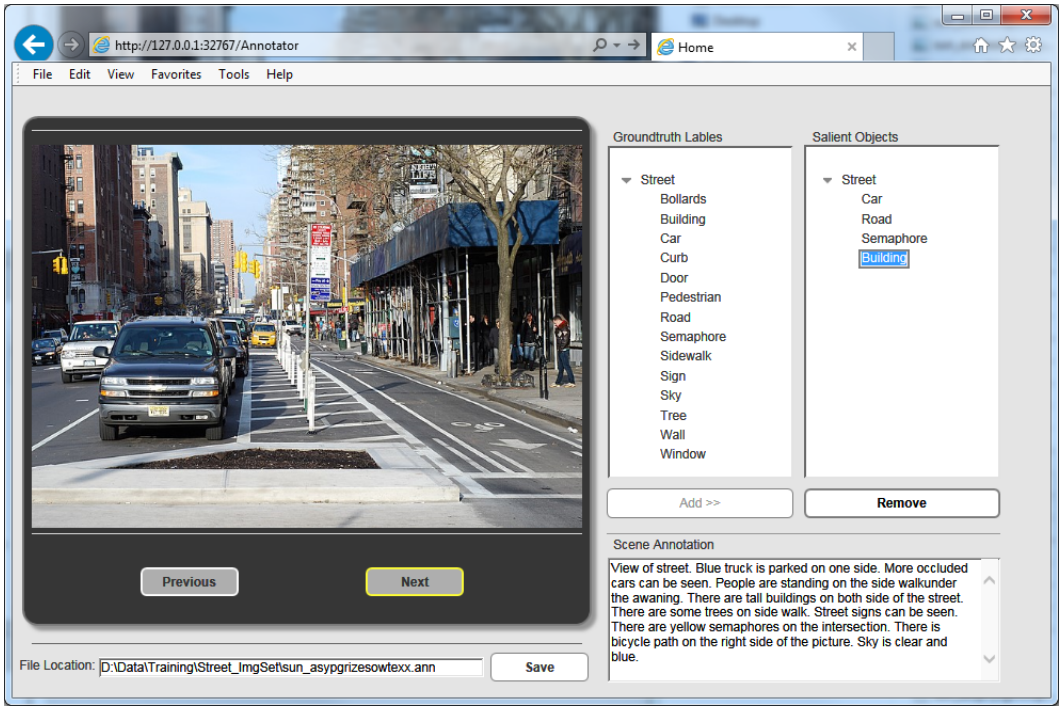


Figure 28. Manual annotation tool used to capture user metadata.

The term ω_{xy} is zero when there are no semantical correlations and as a result $\exp(\omega_{xy})$ has no effect on distance measure allowing co-appearance in the images to take over. However, negative or positive relationships represented by positive or negative values can influence the co-appearance measure.

7.4 Annotation Performance Analysis

To evaluate the annotation performance, a context-based form of *Mean Average Precision* (MAP) is used to measure the semantical annotation accuracy for a given scene

$$\Delta = \frac{1}{M}\sum_{m=1}^{M}\left(\sum_{i=1}^{I}\frac{P(c_i^m)}{I} - \sum_{k=1}^{K}\frac{(v_{k,m}^G)'}{K}\right) \tag{39}$$

where M is the number of the scene categories, I is total number of captured annotation words in the scene category m, K is number of annotation captions in

Table 5. Top 3 Object recognition results

Scene	Manual	sGMM	cGMM
Airport	Airplane, Tower, Runway	Sky, Airplane, Trees	Airplane, Runway, Grass
Bedroom	Bed, Lamp, Nightstand	Bed, Dresser, Book	Bed, Pillow, Lamp
Living room	Sofa, Coffee Table, Chair	Chair, Book shelf, wall	Sofa, Chair, Window
Office	Desk, Chair, Monitor	Person, Book shelf, Chair	Desk, Monitor, Keyboard
Street	Car, Road, Sidewalk	Building, Car, Trees	Car, Road, Semaphores
Bathroom	Toilet, Bathtub, Sink	Wall, Sink, Faucet	Bathtub, Toilet, Faucet
Kitchen	Cabinets, Stove, Refrigerator	Cabinets, Person, Faucet	Refrigerator, Cabinets, Oven
Coast	Ocean, Rocks, Sunset	Sky, Mountain, Ocean	Rock, Ocean, Beach

m^{th} scene of the ground truth and c_i^m is ranked annotation captions of m^{th} scene and $P(c_i^m) = 1$ if c_i^m is found in captions of ground truth and $P(c_i^m) = 0$ otherwise.

The second term is a penalty for ground truth annotation captions that are missing and not discovered. $v_{k,m}^G$ corresponds to saliency score of the k^{th} ground truth object for the m^{th} scene.

7.5 Experiments

In these experiments, eight scenes categories were selected from Sun397 dataset with almost 100 images under each category. There are average nine object types in each image with average 16 objects present (see Figure 27) Object images were extracted based on their bounding box information to train object classifiers. The

image sets containing were equally divided among human annotators, and their input was collected and stored.

A simple annotation tool shown in Figure 28 was used to manage training images annotation. This tool queries annotation in Sun397 database and retrieves the ground truth information. Annotators, who are not specialized in computer vision, used this tool to select and arrange an ordered list of salient objects which to their opinion better reflects the meaning of the scene.

Training of baseline GMM model begins with extracting low-level features. Similar to our previous experiments we used SURF feature descriptors and BoF representation to build and train our models. There are two sets of object classifiers: baseline classifiers based on the method described in section 7.2.2 and context-based extension of section 7.2.3. To achieve our objectives, we trained the context-based GMM model with contextual features and measured refinements performed at every step. Features and attributes that have been studied are as following:

Low-level Features: Surf feature descriptors were used to train baseline object classifiers as described in 7.2.

Semantic Context: GMM is very efficient in capturing co-location relation among objects.

Relative Spatial and Scale Contexts: These contexts were incorporated as high order interactions described in 6.6.

Absolute Spatial Context: Images are divided into three regions of $\{top, middle, bottom\}$ where some objects often appear with very little variation, for instance sky is expected to appear on top, and road always on the bottom of an image. An unsupervised GMM with 3 components is use to model this feature (see section 7.2.1).

Color Context: To model colors first RGB colors were transformed into CIELAB[2] system and then using *k*-mean, the A and B color components of each pixel are quantized into 32 bins corresponding to primary colors. The color feature is then represented by a vector of top 3 dominant colors in an object. For instance, the sky is white and blue, or grass and tree are brown and yellow and green.

7.6 Results

The top three objects recognition results in each selected category are listed in Table 5. The "manual" column shows the human subject caption selections with their ranks adjusted by majority voting. *Supervised GMM* (sGMM) results are ranked according to the recognition confidence. *Contextual GMM* (cGMM) results are ranked based on the object saliency score described in Equation (37).

To calculate the saliency scores, first, we computed the semantical distance of the top ranking scene saliency object pick from manually selected object using the Equation (38). The result shown in Table 6 indicates the Context-based GMM (cGMM) has achieved better performance by being semantically closer to manual selection of human subjects with an average distance of 0.0925 or 90% accuracy versus distance of .4163 or 58% accuracy of supervised GMM.

[2] https://en.wikipedia.org/wiki/Lab˙color˙space

Table 6. Salient object semantical distance(lower is better)

Scene	Manual(z)	sGMM(x)	$D(x,z)$	cGMM(y)	$D(y,z)$
Airport	Airplane	Sky	0.89	Airplane	0.00
Bedroom	Bed	Bed	0.00	Bed	0.00
Living room	Sofa	Chair	0.67	Sofa	0.00
Office	Desk	Person	0.56	Desk	0.00
Street	Car	Building	0.76	Car	0.00
Bathroom	Toilet	Toilet	0.00	Bathtub	0.13
Kitchen	Cabinets	Cabinets	0.00	Refrigerator	0.29
Coast	Beach	Sky	0.45	Rock	0.32

To measure overall performance of the cGMM versus GMM, Equation(39) was calculated for each scene in our dataset with following results (Table 7):

The result shown in Table 7 show an adjustment for missing annotation of terms related to salient object and compensation for overall semantical coverage of discovered terms. The comparison between the two methods demonstrates the superiority of cGMM with 79% overall accuracy versus 48% performance of sGMM in the discovery of the important concepts associated with salient objects.

7.7 Discussion and Concluding Remarks

In this study, we demonstrated the influence of contextual semantics on the annotation results and particularly selection of the salient object. Our experiment clearly shows using context improves the performance of the object recognition by refining salient object selection which plays a critical role in finding scene point of interest and ultimately uncovers the real meaning of the scene. We provided a semantical distance measurement that can quantify the similarity of visual annotation words in video sequences based on saliency score and showed how our

Table 7. Overall performance of the annotation sGMM vs. cGMM

Scene	sGMM	cGMM
Airport	51%	83%
Bedroom	62%	88%
Living room	28%	69%
Office	16%	82%
Street	44%	87%
Bathroom	53%	75%
Kitchen	65%	80%
Coast	68%	71%

method improved by producing annotation that is close to ground truth compare to the baseline method.

REFERENCES

Adams, A., Gelfand, N., Dolson, J., & Levoy, M. (2009). Gaussian KD-trees for fast high-dimensional filtering. In *ACM SIGGRAPH 2009 papers* (pp. 1–12). ACM. https://doi.org/10.1145/1576246.1531327

Asher, M. F., Tolhurst, D. J., Troscianko, T., & Gilchrist, I. D. (2013). Regional effects of clutter on human target detection performance. *Journal of Vision*, *13*(5), 25–25. https://doi.org/10.1167/13.5.25

Babaguchi, N., Kawai, Y., & Kitahashi, T. (2002). Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE Transactions on Multimedia*, *4*(1), 68–75. https://doi.org/10.1109/6046.985555

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, *110*(3), 346–359. https://doi.org/10.1016/j.cviu.2007.09.014

Berg, A. C., Berg, T. L., Daume, H., Dodge, J., Goyal, A., Xufeng Han, … Yamaguchi, K. (2012). Understanding and predicting importance in images (pp. 3562–3569). IEEE. https://doi.org/10.1109/CVPR.2012.6248100

Biederman, I., Rabinowitz, J. C., Glass, A. L., & Stacy, E. W. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology*, *103*(3), 597–600. https://doi.org/10.1037/h0037158

Bileschi, S. M. (2006). *StreetScenes : towards scene understanding in still images* (Thesis). Massachusetts Institute of Technology. Retrieved from http://dspace.mit.edu/handle/1721.1/37896

Bioucas-Dias, J. M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., & Chanussot, J. (2012). Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *5*(2), 354–379.

Blei, D. M., & Jordan, M. I. (2003). Modeling annotated data (p. 127). ACM Press. https://doi.org/10.1145/860435.860460

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, *3*, 993–1022.

Boykov, Y., & Veksler, O. (2006). Graph Cuts in Vision and Graphics: Theories and Applications. In N. Paragios, Y. Chen, & O. Faugeras (Eds.), *Handbook of Mathematical Models in Computer Vision* (pp. 79–96). New

York: Springer-Verlag. Retrieved from http://link.springer.com/10.1007/0-387-28831-7˙5

Boykov, Y., Veksler, O., & Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(11), 1222–1239. https://doi.org/10.1109/34.969114

Bravo, M. J., & Farid, H. (2008). A scale invariant measure of clutter. *Journal of Vision*, *8*(1), 23. https://doi.org/10.1167/8.1.23

Cai, Y., & Leung, H. (2011). Formalizing object membership in fuzzy ontology with property importance and property priority (pp. 1719–1726). IEEE. https://doi.org/10.1109/FUZZY.2011.6007511

Cao, X., Wei, X., Han, Y., & Chen, X. (2015). An Object-Level High-Order Contextual Descriptor Based on Semantic, Spatial, and Scale Cues. *IEEE Transactions on Cybernetics*, *45*(7), 1327–1339. https://doi.org/10.1109/TCYB.2014.2350517

Chang, S.-F., He, J., Jiang, Y.-G., Yanagawa, A., Zavesky, E., Khoury, E., & Ngo, C.-W. (2008). Columbia University/VIREO-CityU/IRIT TRECVID2008 High-Level Feature Extraction and Interactive Video Search. In *TRECVID*. Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.329.1713&rep=rep1&type=pdf

Chang, S.-F., Hsu, W., Kennedy, L., Xie, L., Yanagawa, A., Zavesky, E., & Zhang, D. (2005). Video Search and High-Level Feature Extraction. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.104.6182

Chen, G., Ding, Y., Xiao, J., & Han, T. X. (2013). Detection Evolution with Multi-order Contextual Co-occurrence. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1798–1805). https://doi.org/10.1109/CVPR.2013.235

Chen, H., Gallagher, A., & Girod, B. (2012). Describing clothing by semantic attributes. In *European Conference on Computer Vision* (pp. 609–623). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-33712-3˙44

Chesson, J. (1976). A Non-Central Multivariate Hypergeometric Distribution Arising from Biased Sampling with Application to Selective Predation. *Journal of Applied Probability*, *13*(4), 795. https://doi.org/10.2307/3212535

Chiverton, J., Xie, X., & Mirmehdi, M. (2012). Automatic Bootstrapping and Tracking of Object Contours. *IEEE Transactions on Image Processing*, *21*(3), 1231–1245. https://doi.org/10.1109/TIP.2011.2167343

Choi, M. J., Torralba, A., & Willsky, A. S. (2012a). A Tree-Based Context Model for Object Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(2), 240–252. https://doi.org/10.1109/TPAMI.2011.119

Choi, M. J., Torralba, A., & Willsky, A. S. (2012b). Context models and out-of-context objects. *Pattern Recognition Letters*, *33*(7), 853–862. https://doi.org/10.1016/j.patrec.2011.12.004

Cilibrasi, R. L., & Vitanyi, P. M. B. (2007). The Google Similarity Distance. *IEEE Trans. on Knowl. and Data Eng.*, *19*(3), 370–383. https://doi.org/10.1109/TKDE.2007.48

Clifford, P. (1990). Markov random fields in statistics. *Disorder in Physical Systems: A Volume in Honour of John M. Hammersley*, 19–32.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV* (pp. 1–22).

Desai, C., Ramanan, D., & Fowlkes, C. (2009). Discriminative models for multi-class object layout. In *2009 IEEE 12th International Conference on Computer Vision* (pp. 229–236). https://doi.org/10.1109/ICCV.2009.5459256

Dubey, R., Peterson, J., Khosla, A., Yang, M.-H., & Ghanem, B. (2015). What Makes an Object Memorable? (pp. 1089–1097). IEEE. https://doi.org/10.1109/ICCV.2015.130

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*(3), 433–458. https://doi.org/10.1037/0033-295X.96.3.433

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, *8*(3), 3. https://doi.org/10.1167/8.3.3

Endo, N., & Takeda, Y. (2005). Use of spatial context is restricted by relative position in implicit learning. *Psychonomic Bulletin & Review*, *12*(5), 880–885. https://doi.org/10.3758/BF03196780

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every Picture Tells a Story: Generating Sentences from Images. In *Proceedings of the 11th European Conference on Computer Vision: Part IV* (pp. 15–29). Berlin, Heidelberg: Springer-Verlag. Retrieved from http://dl.acm.org/citation.cfm?id=1888089.1888092

Fei-Fei Li, & Perona, P. (2005). A Bayesian Hierarchical Model for Learning Natural Scene Categories (Vol. 2, pp. 524–531). IEEE. https://doi.org/10.1109/CVPR.2005.16

Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (pp. 1–8). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs˙all.jsp?arnumber=4587597

Fergus, R., Perona, P., & Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings* (Vol. 2, p. II). IEEE. https://doi.org/10.1109/CVPR.2003.1211479

Fernando, B., Fromont, E., Muselet, D., & Sebban, M. (2012a). Supervised learning of Gaussian mixture models for visual vocabulary generation. *Pattern Recognition*, *45*(2), 897–907. https://doi.org/10.1016/j.patcog.2011.07.021

Fernando, B., Fromont, E., Muselet, D., & Sebban, M. (2012b). Supervised learning of Gaussian mixture models for visual vocabulary generation. *Pattern Recognition*, *45*(2), 897–907.

Fink, M., & Perona, P. (2004). Mutual Boosting for Contextual Inference. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems 16* (pp. 1515–1522). MIT Press. Retrieved from http://papers.nips.cc/paper/2520-mutual-boosting-for-contextual-inference.pdf

Fischler, M. A., & Strat, T. M. (1989). *Recognizing objects in a natural environment: a contextual vision system (CVS)*. DTIC Document. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA460946

Fleischman, M., & Roy, D. (2008). Grounded Language Modeling for Automatic Speech Recognition of Sports Video. In *ACL* (pp. 121–129). Retrieved from http://www.aclweb.org/website/old˙anthology/P/P08/P08-1.pdf#page=165

Fleming, M. L. (1977). The picture in your mind. *AV Communication Review*, *25*(1), 43–62. https://doi.org/10.1007/BF02799310

Fog, A. (2008). Calculation Methods for Wallenius' Noncentral Hypergeometric Distribution. *Communications in Statistics - Simulation and Computation*, *37*(2), 258–273. https://doi.org/10.1080/03610910701790269

Freeman, W. T., Murphy, K. P., & Torralba, A. (n.d.). *Contextual models for object detection using boosted random fields*. Retrieved from http://hdl.handle.net/1721.1/30482

Galleguillos, C., & Belongie, S. (2010). Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, *114*(6), 712–722. https://doi.org/10.1016/j.cviu.2010.02.004

Galleguillos, C., McFee, B., Belongie, S., & Lanckriet, G. (2010). Multi-Class Object Localization by Combining Local Contextual Interactions. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2010)*.

Galleguillos, C., Rabinovich, A., & Belongie, S. (2008). Object categorization using co-occurrence, location and appearance. In *2008 IEEE Conference on*

Computer Vision and Pattern Recognition (pp. 1–8). IEEE. https://doi.org/10.1109/CVPR.2008.4587799

Gould, S., Rodgers, J., Cohen, D., Elidan, G., & Koller, D. (2008). Multi-Class Segmentation with Relative Location Prior. *International Journal of Computer Vision*, *80*(3), 300–316. https://doi.org/10.1007/s11263-008-0140-x

Gronau, N., Neta, M., & Bar, M. (2008). Integrated Contextual Representation for Objects' Identities and Their Locations. *Journal of Cognitive Neuroscience*, *20*(3), 371–388. https://doi.org/10.1162/jocn.2008.20027

Hadfield, J. D., & others. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software*, *33*(2), 1–22.

Han, Z., Ye, Q., & Jiao, J. (2008). Online Feature Evaluation for Object Tracking Using Kalman Filter. In *19th International Conference on Pattern Recognition (ICPR 2008)*.

Heitz, G., & NBKollerNB, D. (2008). Learning Spatial Context: Using Stuff to Find Things. In *10th European Conference on Computer Vision (ECCV 2008)*.

Hemingway, E. (2008). *A Moveable Feast*. London: Vintage/Ebury.

Henderson, J. M., Chanceaux, M., & Smith, T. J. (2009). The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements. *Journal of Vision*, *9*(1), 32–32. https://doi.org/10.1167/9.1.32

Hock, H. S., Gordon, G. P., & Whitehurst, R. (1974). Contextual relations: The influence of familiarity, physical plausibility, and belongingness. *Perception & Psychophysics*, *16*(1), 4–8. https://doi.org/10.3758/BF03203242

Hoiem, D., Efros, A. A., & Hebert, M. (2005). Geometric context from a single image (p. 654–661 Vol. 1). IEEE. https://doi.org/10.1109/ICCV.2005.107

Hoiem, D., Efros, A. A., & Hebert, M. (2008). Putting Objects in Perspective. *International Journal of Computer Vision*, *80*(1), 3–15. https://doi.org/10.1007/s11263-008-0137-5

Hou, Y., He, L., Zhao, X., & Song, D. (2011). Pure High-Order Word Dependence Mining via Information Geometry. In G. Amati & F. Crestani (Eds.), *Advances in Information Retrieval Theory* (pp. 64–76). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-23318-0˙8

Huang, R., Zhou, P., & Zhang, L. (2014). A LDA-Based Approach for Semi-Supervised Document Clustering. *International Journal of Machine Learning and Computing*, *4*(4), 313.

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? (pp. 145–152). IEEE. https://doi.org/10.1109/CVPR.2011.5995721

Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database (pp. 248–255). IEEE. https://doi.org/10.1109/CVPR.2009.5206848

Jiang, L., Yu, S.-I., Meng, D., Yang, Y., Mitamura, T., & Hauptmann, A. G. (2015). Fast and Accurate Content-based Semantic Search in 100M Internet Videos (pp. 49–58). ACM Press. https://doi.org/10.1145/2733373.2806237

Jiang, W., Chang, S.-F., & Loui, A. C. (2007). Context-Based Concept Fusion with Boosted Conditional Random Fields. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07* (Vol. 1, p. I-949-I-952). IEEE. https://doi.org/10.1109/ICASSP.2007.366066

Jiang, Y., Lim, M., & Saxena, A. (2012). Learning Object Arrangements in 3D Scenes using Human Context. *arXiv:1206.6462 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1206.6462

Jones, S., & Shao, L. (2014). Unsupervised Spectral Dual Assignment Clustering of Human Actions in Context. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 604–611). https://doi.org/10.1109/CVPR.2014.84

Juneja, M., Vedaldi, A., Jawahar, C. V., & Zisserman, A. (2013). Blocks that shout: Distinctive parts for scene classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 923–930). Retrieved from http://www.cv-foundation.org/openaccess/content˙cvpr˙2013/html/Juneja˙Blocks˙That˙Shout˙2013˙CVPR˙paper.html

Khan, Z. H., Gu, I. Y. H., & Backhouse, A. (2011). Robust Visual Object Tracking using Multi-Mode Anisotropic Mean Shift and Particle Filters. Retrieved from http://publications.lib.chalmers.se/publication/126617-robust-visual-object-tracking-using-multi-mode-anisotropic-mean-shift-and-particle-filters

Kohli, P., Ladický, L., & Torr, P. (2009). Robust Higher Order Potentials for Enforcing Label Consistency. *International Journal of Computer Vision*, *82*(3), 302–324. https://doi.org/10.1007/s11263-008-0202-0

Kok, P., & de Lange, F. P. (2014). Shape Perception Simultaneously Up- and Downregulates Neural Activity in the Primary Visual Cortex. *Current Biology*, *24*(13), 1531–1535. https://doi.org/10.1016/j.cub.2014.05.042

Kolmogorov, V., & Zabin, R. (2004). What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(2), 147–159. https://doi.org/10.1109/TPAMI.2004.1262177

Krähenbühl, P., & Koltun, V. (2012). Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. Retrieved from http://arxiv.org/abs/1210.5644

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., … Berg, T. L. (2013). BabyTalk: Understanding and Generating Simple Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2891–2903. https://doi.org/10.1109/TPAMI.2012.162

Ladicky, L., Russell, C., Kohli, P., & Torr, P. H. S. (2010). Graph Cut Based Inference with Co-occurrence Statistics. In K. Daniilidis, P. Maragos, & N. Paragios (Eds.), *Computer Vision – ECCV 2010* (pp. 239–253). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-15555-0˙18

Lafferty, J. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data (pp. 282–289). Morgan Kaufmann.

Lai, H., Pan, Y., Liu, Y., & Yan, S. (2015). Simultaneous feature learning and hash coding with deep neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3270–3278). IEEE. https://doi.org/10.1109/CVPR.2015.7298947

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories (Vol. 2, pp. 2169–2178). IEEE. https://doi.org/10.1109/CVPR.2006.68

Leutenegger, S., Chli, M., & Siegwart, R. Y. (2011). BRISK: Binary Robust Invariant Scalable Keypoints. In *Proceedings of the 2011 International Conference on Computer Vision* (pp. 2548–2555). Washington, DC, USA: IEEE Computer Society. https://doi.org/10.1109/ICCV.2011.6126542

Li, L.-J., & Fei-Fei, L. (2007). What, where and who? classifying events by scene and object recognition. In *2007 IEEE 11th International Conference on Computer Vision* (pp. 1–8). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs˙all.jsp?arnumber=4408872

Liu, T., Huang, X., & Ma, J. (2016). Conditional Random Fields for Image Labeling. *Mathematical Problems in Engineering*, *2016*, 1–15. https://doi.org/10.1155/2016/3846125

Lowe, D. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, *60*(2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

Lu, W.-L., & Little, J. . (2006). Simultaneous Tracking and Action Recognition using the PCA-HOG Descriptor. In *The 3rd Canadian Conference on Computer and Robot Vision (CRV'06)* (p. 6). IEEE. https://doi.org/10.1109/CRV.2006.66

Ma, J., & Gao, W. (n.d.). The supervised learning Gaussian mixture model. *Journal of Computer Science and Technology*, *13*(5), 471–474. https://doi.org/10.1007/BF02948506

Mcauliffe, J. D., & Blei, D. M. (2008). Supervised topic models. In *Advances in neural information processing systems* (pp. 121–128). Retrieved from http://papers.nips.cc/paper/3328-supervised-topic

Mettes, P., Koelma, D. C., & Snoek, C. G. M. (2016). The ImageNet Shuffle: Reorganized Pre-training for Video Event Detection. https://doi.org/10.1145/2911996.2912036

Morik, Katharina, & Piatkowski, Nico. (2012). Parallel Loopy Belief Propagation in Conditional Random Fields. https://doi.org/10.17877/DE290R-7508

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological Cybernetics*, *66*(3), 241–251.

Murphy, K. P., Torralba, A., & Freeman, W. T. (2004). Using the Forest to See the Trees: A Graphical Model Relating Features, Objects, and Scenes. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems 16* (pp. 1499–1506). MIT Press. Retrieved from http://papers.nips.cc/paper/2386-using-the-forest-to-see-the-trees-a-graphical-model-relating-features-objects-and-scenes.pdf

Myeong, H., & Lee, K. M. (2013). Tensor-Based High-Order Semantic Relation Transfer for Semantic Scene Segmentation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3073–3080). https://doi.org/10.1109/CVPR.2013.395

Nematollahi, M., & Zhang, X.-P. (2014). A new robust context-based dense CRF model for image labeling (pp. 5876–5880). IEEE. https://doi.org/10.1109/ICIP.2014.7026187

Pan, H., Wang, B., & Jiang, H. (2015). Deep Learning for Object Saliency Detection and Image Segmentation. *arXiv:1505.01173 [Cs]*. Retrieved from http://arxiv.org/abs/1505.01173

Pandey, M., & Lazebnik, S. (2011). Scene recognition and weakly supervised object localization with deformable part-based models. In *2011 International Conference on Computer Vision* (pp. 1307–1314). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6126383

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Putthividhya, D., Attias, H. T., & Nagarajan, S. S. (2010). Supervised topic model for automatic image annotation (pp. 1894–1897). IEEE. https://doi.org/10.1109/ICASSP.2010.5495341

Qi, G.-J., Hua, X.-S., Rui, Y., Tang, J., Mei, T., & Zhang, H.-J. (2007). Correlative multi-label video annotation. In *Proceedings of the 15th international conference on multimedia* (pp. 17–26). ACM. https://doi.org/10.1145/1291233.1291245

Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., & Belongie, S. (2007). Objects in Context. In *2007 IEEE 11th International Conference on Computer Vision* (pp. 1–8). IEEE. https://doi.org/10.1109/ICCV.2007.4408986

Rasiwasia, N., & Vasconcelos, N. (2013). Latent Dirichlet Allocation Models for Image Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(11), 2665–2679. https://doi.org/10.1109/TPAMI.2013.69

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF (pp. 2564–2571). IEEE. https://doi.org/10.1109/ICCV.2011.6126544

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … others. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Seddati, O., Dupont, S., & Mahmoudi, S. (2015). DeepSketch: Deep convolutional neural networks for sketch recognition and similarity search. In *2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)* (pp. 1–6). IEEE. https://doi.org/10.1109/CBMI.2015.7153606

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 411–426.

Shafieyan, F., Karimi, N., Nasr-Esfahani, E., & Samavi, S. (2014). Image seam carving based on content importance and depth maps (pp. 1786–1791). IEEE. https://doi.org/10.1109/IranianCEE.2014.6999828

Shotton, J., Winn, J., Rother, C., & Criminisi, A. (2006). TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation. In A. Leonardis, H. Bischof, & A. Pinz (Eds.), *Computer Vision – ECCV 2006* (pp. 1–15). Springer Berlin Heidelberg. https://doi.org/10.1007/11744023˙1

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv Preprint arXiv:1409.1556*. Retrieved from http://arxiv.org/abs/1409.1556

Smeaton, A., Over, P., & Kraaij, W. (2006). Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM international workshop on multimedia information retrieval* (pp. 321–330). ACM. https://doi.org/10.1145/1178677.1178722

Strat, T. M. (1993). Employing contextual information in computer vision. *DARPA93*, 217–229.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2014). Going Deeper with Convolutions. *arXiv:1409.4842 [Cs]*. Retrieved from http://arxiv.org/abs/1409.4842

Tang, J., Shao, L., & Zhen, X. (2014). Robust point pattern matching based on spectral context. *Pattern Recognition*, *47*(3), 1469–1484. https://doi.org/10.1016/j.patcog.2013.09.017

Torralba, A. (2003). Contextual Priming for Object Detection. *International Journal of Computer Vision*, *53*(2), 169–191. https://doi.org/10.1023/A:1023052124951

Torralbo, A., Walther, D. B., Chai, B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2013). Good Exemplars of Natural Scene Categories Elicit Clearer Patterns than Bad Exemplars but Not Greater BOLD Activity. *PLoS ONE*, *8*(3), e58594. https://doi.org/10.1371/journal.pone.0058594

Tu, Z. (2008). Auto-context and its application to high-level vision tasks. In *2008 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8). IEEE. https://doi.org/10.1109/CVPR.2008.4587436

Van De Sande, K. E., Gevers, T., & Smeulders, A. W. (2009). The university of amsterdam's concept detection system at imageclef 2009. In *Workshop of the Cross-Language Evaluation Forum for European Languages* (pp. 261–268). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-15751-6˙32

Vogel, J., & Schiele, B. (2007). Semantic Modeling of Natural Scenes for Content-Based Image Retrieval. *International Journal of Computer Vision*, *72*(2), 133–157. https://doi.org/10.1007/s11263-006-8614-1

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*(9), 1395–1407. https://doi.org/10.1016/j.neunet.2006.10.001

Wang, C., Blei, D., & Li, F.-F. (2009). Simultaneous image classification and annotation. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1903–1910). IEEE. https://doi.org/10.1109/CVPR.2009.5206800

Wang, J., Chen, Z., & Wu, Y. (2011). Action recognition with multiscale spatio-temporal contexts. In *CVPR 2011* (pp. 3185–3192). https://doi.org/10.1109/CVPR.2011.5995493

Wang, S., Joo, J., Wang, Y., & Zhu, S.-C. (2013). Weakly Supervised Learning for Attribute Localization in Outdoor Scenes (pp. 3111–3118). Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Retrieved from http://www.cv-

foundation.org/openaccess/content˙cvpr˙2013/html/Wang˙Weakly˙Supervise d˙Learning˙2013˙CVPR˙paper.html

Wang, Y., Sabzmeydani, P., & Mori, G. (2007). Semi-Latent Dirichlet Allocation: A Hierarchical Model for Human Action Recognition. In A. Elgammal, B. Rosenhahn, & R. Klette (Eds.), *Human Motion – Understanding, Modeling, Capture and Animation* (Vol. 4814, pp. 240–254). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://link.springer.com/10.1007/978-3-540-75703-0˙17

Wolf, L., & Bileschi, S. (2006). A Critical View of Context. *International Journal of Computer Vision*, *69*(2), 251–261. https://doi.org/10.1007/s11263-006-7538-0

Xiao, J., Ehinger, K. A., Hays, J., Torralba, A., & Oliva, A. (2016). SUN Database: Exploring a Large Collection of Scene Categories. *International Journal of Computer Vision*, *119*(1), 3–22. https://doi.org/10.1007/s11263-014-0748-y

Xu, Z., Yang, Y., & Hauptmann, A. G. (2014). A Discriminative CNN Video Representation for Event Detection. Retrieved from http://arxiv.org/abs/1411.4006

Yakimovsky, Y., & Feldman, J. A. (1973). A Semantics-Based Decision Theory Region Analyser. In *IJCAI* (pp. 580–588). Retrieved from http://www.ijcai.org/Proceedings/73/Papers/062.pdf

Yang, J., Jiang, Y.-G., Hauptmann, A., & Ngo, C.-W. (2007). Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on workshop on multimedia information retrieval* (pp. 197–206). ACM. https://doi.org/10.1145/1290082.1290111

Yang, Y., Teo, C. L., Daumé, H., III, & Aloimonos, Y. (2011). Corpus-guided Sentence Generation of Natural Images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 444–454). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2145432.2145484

Zhang, L., Kalashnikov, D., Mehrotra, S., & Vaisenberg, R. (2014). Context-based person identification framework for smart video surveillance. *Machine Vision and Applications*, *25*(7), 1711–1725. https://doi.org/10.1007/s00138-013-0535-8

Zhu, H., Meng, F., Cai, J., & Lu, S. (2016). Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. *Journal of Visual Communication and Image Representation*, *34*, 12–27. https://doi.org/10.1016/j.jvcir.2015.10.012

Zhu, J., Wu, T., Zhu, S.-C., Yang, X., & Zhang, W. (2012). Learning reconfigurable scene representation by tangram model. In *Applications of Computer Vision (WACV), 2012 IEEE Workshop on* (pp. 449–456). IEEE.

Retrieved from http://ieeexplore.ieee.org/xpls/abs˙all.jsp?arnumber=6163023

Zhu, S., & Yung, N. H. C. (2014). Improve scene categorization via sub-scene recognition. *Machine Vision and Applications*, *25*(6), 1561–1572. https://doi.org/10.1007/s00138-014-0622-5

Zhu, Y., Nayak, N. M., & Roy-Chowdhury, A. K. (2013). Context-aware modeling and recognition of activities in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2491–2498). Retrieved from http://www.cv-foundation.org/openaccess/content˙cvpr˙2013/html/Zhu˙Context-Aware˙Modeling˙and˙2013˙CVPR˙paper.html

Zolghadr, E., & Furht, B. (2016a). Context-Based Scene Understanding: *International Journal of Multimedia Data Engineering and Management*, *7*(1), 22–40. https://doi.org/10.4018/IJMDEM.2016010102

Zolghadr, E., & Furht, B. (2016b). Scene Understanding Using Context -based Conditional Random Field. In *ResearchGate*. Retrieved from https://www.researchgate.net/publication/306047240˙Scene˙Understanding˙Using˙Context˙-based˙Conditional˙Random˙Field