Scene Understanding Using Context-based Conditional Random Field

Esfandiar Zolghadr
PhD candidate, CECS
Florida Atlantic University
777 Glades Road, Boca Raton, USA
ezolghad@fau.edu

Borko Furht
Professor, CECS
Florida Atlantic University
777 Glades Road, Boca Raton, USA
bfurht@fau.edu

ABSTRACT

In this paper, a new framework for scene understanding using multi-modal high-ordered context-model is introduced. Spatial and semantical interactions are considered as sources of context which are incorporated in the model using a single object-scene relevance measure that quantifies high-order object relations. This score is used to minimize semantical inconsistencies among objects in dense graph representation of the scene category during the object recognition process. New context model is later incorporated in a Conditional Random Fields (CRF) framework to combine contextual cues with appearance descriptors in order to increase object localization and class prediction accuracy. A novel context-based non-central hypergeometric unary potential is defined to maximize the semantical coherence in the scene. Further refinement is performed using context-based pairwise and high-order potentials which use alpha-expansion and graph-cut to find optimal configuration. Comparison between the purposed approach and state-of-art algorithms shows effectiveness of this approach in annotation and interpretation of scenes.

Keywords

Context-based scene recognition, supervised classification, generative model, representative feature

1. INTRODUCTION

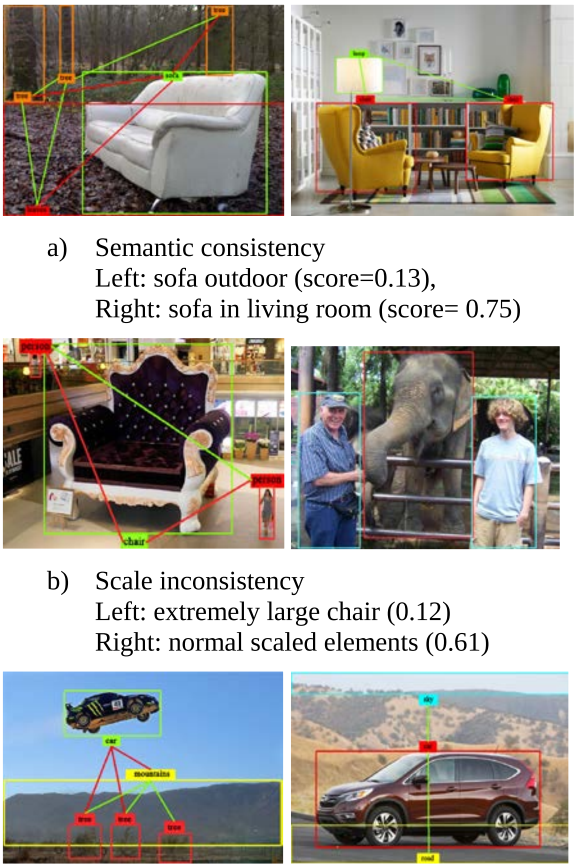
Scene understanding has been studied for decades in areas such as content annotation, object and event recognition and media retrieval engines. The main objective is to provide more precise and accurate description of the scene in order to better respond to user queries. Media search engines currently use manually entered tags in metadata to recognize the content of images. Automated annotation frameworks are preferred methods for high volume contents. They utilize audio visual features to localize and classify candidate objects in an image which assigns class labels to components of a scene [1].

In a top-down approach, a global feature is extracted and used to classify the image as high-level categories such as (indoor/outdoor). Later, more specialized object detectors in that category are applied for detail analysis of the image. Torralba et al. [2, 3] presents a coarse scene-level global feature called ‘gist’ and use it to classify a scene as indoor or outdoor. This approach also allows checking for presence of an object types without running an object detector.

In a bottom up approach, individual objects detectors are applied to the extracted features from homogenous regions of the image to identify a matching object class. Key disadvantages of these paradigms are high dimensionality of the required detectors and region analysis in isolation. Previous work shows that performance of recognition systems could be significantly improved when scene-level knowledge known as “*context*” is exploited [4, 5, 6, 7, 8]. By definition, any information that can be used in accurate semantical recognition of scene elements and its underlying concepts is called context [9]. Contextual models can capture semantical properties, relationships and interactions among image components which can be used to infer higher level meaning of a scene.

Object detectors face many challenges due to poor image quality or overall image content complexity in cluttered images. Incorporating contextual information is used to disambiguate, refine and improve the recognition results. Additionally, context cueing can reduce dimensionality of required object detectors. Traditionally, contexts are constructed using semantical and spatial relations of objects in a scene [8]. More recent applications use additional types of contextual information provided by acquisition system such as sensory context (GPS) and 3D geometric scene context (orientations, support surface, horizon line) [9].

Semantic context is among most studied types of contexts which models object relationships such as co-occurrence statistics in an image. Spatial context captures inter-objects scale or location relationships based on various taxonomies such as horizontal or vertical relative positions [10].



a) Semantic consistency
Left: sofa outdoor (score=0.13),
Right: sofa in living room (score= 0.75)

b) Scale inconsistency
Left: extremely large chair (0.12)
Right: normal scaled elements (0.61)

c) Spatial Inconsistency
Left: car flying in the sky (0.29)
Right: car on the road (0.72)

Figure 1. These images show sample scenes and their contextual relevance scores obtained using higher-order relationship. Images on the right show objects consistent with scene context. Images on the left demonstrate contextual inconsistency (higher scores signifies more consistency).

The previous work on context-based detectors is mostly limited to studying the objects in pairwise relations with less attention to higher order relations and structure of scene layouts. An image of real life scene configuration can be rather complex where multiple types of contexts should be considered to explain high-ordered relationships.

In this paper, a novel multi-source high-order contextual scene recognition framework is introduced to represent realistic scene configurations. This framework measures contextual consistency among the composing elements of an image using a measurement called “*object-to-scene relevance score*” which measures contribution of an object type to overall semantical meaning of the scene for a given context. The object relevance score is used in modeling underlying scene representation based on high-ordered relationships in the form of undirected graph.

In summary contributions of this work is as follows:

- Definition of a context-based conditional random field designed to incorporate multiple source of context in high-order relationship.
- Definition of object-scene relevance score that encodes layout, relations and interactions of an object to conform to scene consistent context.
- Use of a novel unary potential based on non-central hypergeometric distribution to predict the object labels in a context-based generative process.
- We define a high-order potential to encode high-order contextual relationship of the objects.

The rest of this paper is organized as follows. Section II is an overview of the related work. In Section III we present our framework in detail. Section IV presents our experimental results. Section V is concluding remarks.

2. RELATED WORK

Contextual scene understanding frameworks have been studied in many of the previous work [11, 12]. Wolf and Bileschi [13] 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 indicate the presence of a particular object in the image at a semantic layer.

Galleguillos *et al.* [14] explored pairwise interactions between pixels, regions and objects to extract and learn three source of context semantic, boundary support and contextual neighborhood.

Torralba *et al.* [3] introduced a simple framework for modeling the relationship between context and object properties. Scale context was used to provide a strong cue for scale 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.

Choi *et al.*[15] used tree-structure graphical model to encode hierarchical dependencies among object categories and scenes. They used contextual score to quantify pairwise information such as position and scale relationship.

Jones and Shao [16] 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 [17, 18, 19] on single source of context such as co-occurrence. The shortcoming of these methods is when discriminative contextual cues may appear in other contextual modalities such as scale or spatial context as is illustrated in Figure 1.

On the other side of spectrum, generative models such as [20] are widely used to model multi-context relations. The limitation of these frameworks and generally the generative process is the independence assumption on observed data to make the inference tractable which is very restrictive.

Previous work shows success of context-based methods in improving performance of object localization and recognition. We extend previous work to exploit high-order multimodal contextual relationships instead of pair-wise approach. We propose a high-order context framework that learns appearance, structure and semantical consistency of the scene and infers its parameters based on multi-modal context sources for domain 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 set of children toys.

Spatial and scale contexts are sources of layout topology. Location and scale information is obtained from bounding box information in training dataset and transformed into the set of contextual spatial attributes during learning process. Bounding box information is acquired from the image annotations provided in Sun397[1] dataset. This dataset provides a better alternative to Google Sets or web documents used in some related works [21].

Context model represents a scene with fully connected graph consisting objects at each node. These nodes are connected with undirected edges. Each edge is assigned with contextual relevance measurement that quantifies the relations between two objects given the dominated context in that clique. As shown in Figure 1, contextual relevance is defined to maximize semantical consistencies including scale and location in a scene. Contextual inconsistencies may not manifest in pairwise relations where in ternary relation a clear violation is evident. The object-scene score is scalable and extendible to other datasets since it not dependent on visual primitives. Contextually related objects form semantically coherent cliques in our graph representation and are labeled according.

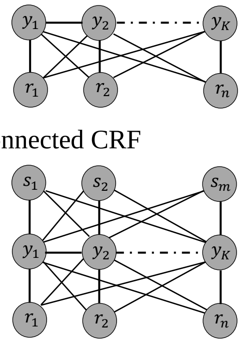
Conditional random fields (CRF) [22], a discriminative framework is used to incorporate contextual cues along with appearance features in a single model. This allows to model intrinsic and extrinsic structure of an image for better understanding of its underlying concepts [23]. Given observed variable X, CRFs model the conditional distribution of Y given X to encode complex dependencies of Y on X. In this paper definition of CRF is extended by conditioning on visual features and context which is called Context-based CRF (CBCRF). CBCRF combines appearance descriptors, contextual relations and layout structure of the objects likely to be present in that scene category.

Our experiments show that contextual relations of high-order can improve object detection, scene classification and can be used in many other applications such as detection of out-of-context or black-boxed object.

3. MODEL AND ALGORITHMS

A conditional random field (CRF) model [22] is used to learn the conditional distribution over the set of class labels given an image. The following is formalization of the model:

With K being total objects in our dataset, let's consider 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 contexts 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.



a) Fully Connected CRF

b) Context-based CRF

Figure 2. Graphical representation of the CRF model (top) and context-based CRF model (below)

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 $P(Y|R)$:

$$P(Y|R,\theta) = P(Y|\,R,S,\theta)\;P(S|R,\theta)$$

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:

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

where Z is normalization partition function.

Our model is fully connected CRF with unary, pairwise and high-order potentials with following Gibbs Energy:

$$E(Y|\,R,S) = \sum_{n \in N} \psi_u(y_n) + \sum_{(i,j) \in P} \psi_p\big(y_i, y_j\big) + \sum_{i \in H} \psi_h(y_i) \qquad (2)$$

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

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

Scene Relevance Score

Context-based conditional random field model builds on the "*Scene Relevance Score*" (SRS) which is calculated using high-ordered interaction of the objects in each scene category.

The high-order pure independence rule [24] is used to define spatial context probabilities. Let $C = \{c_1, c_2, \dots, c_n\}$ be set of object types in our dataset, then A^K represents the set of possible combinations of object types with K object present and $n - K$ not present. For example considering set of $C = \{c_1, \dots, c_4\}$ with four object classes, the set of object configurations with only two objects present could be expressed as $A^2 = \{0011, 0101, 0110, 1010, 1100\}$.

Scene relevance-score is defined as posterior probability which is log likelihood of spatial, scale and semantic contexts:

$$\tau_{1\dots n} = log \prod_{k=0}^{n} \prod_{a \in A^K} \left(P_{L_{1\dots n}^{v}|a}\right)^{(-1)^{n-k}} \left(P_{X_{1\dots n}|\mathrm{a}}\right)^{(-1)^{n-k}} \qquad (3)$$

where $P_{L_{1\dots n}^{v}|a}$ is spatial context and is defined as posterior probability of vertical location of an object in respect to others in a high-order relation. $L_{1\dots n}^{v}$ is high order relative vertical location configuration defined as joint probability distribution of $L_1, L_2, \dots, L_n$. $L_i \in \{above, below, even\}$ and is determined be comparing centroids of each object's bounding box. For example expected relative location of "*Sky*" is "*above*" the object "*Grass*".

$P_{X_{1\dots n}|\mathrm{a}}$ is high-ordered 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 obtained from relative horizontal location does not offer discriminative information and is not modeled.

The semantical relationship is implicitly encoded in scene relevance score in Equation (3) with shows strong semantical correlation for positive values and negative correlation for negative values of $\tau_{1\dots n}$ and zero for no relation.

$$\tau_{1\dots n} = \theta_{1\dots n} + \gamma_{1\dots n} \qquad (4)$$

Normalizing τ transforms the value to zero and one range which more suitable to our context model. The following function transforms the τ to normalized form:

$$\overline{\tau_{1\dots n}} = \frac{1}{1 + exp(-\tau_{1\dots n})}$$

The normalized value of $\overline{\tau_{1\dots n}}$ is interpreted as follows:

$$\begin{cases} 0.5 < \bar{\tau} \leq 1 & \text{semantical related} \\ \bar{\tau} = 0.5 & \text{no relation} \\ 0 \leq \bar{\tau} < 0.5 & \text{negative relation} \end{cases} \qquad (5)$$

Strength of relationship increases with the value of $\bar{\tau}$ from 0 (impossible) up to 1 (strongly coupled).

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 misclassified labels can be refined. Unary potential is defined as:

$$\psi_u(y_n) = p(Y|R, S) \qquad (6)$$

where $S_i = \{s_1, \cdots, s_m\}$ is all possible context graphs and the term $p(Y|R, S)$ is probability that object i^{th} would be assigned label y given the relevance score of the object.

Wallenius Latent Dirichlet Allocation (WLDA) [28] is a generative process for object localization. An image is partitioned into related groups of visual words which represent candidate objects and assigns best annotation label to the image category. In this process each label is associated with image feature data as response variable which is influenced by contextual constraints as bias weight parameter in Wallenius distribution.

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 $R_n, n \in \{1, 2, \dots, N\}$:
  - Draw topic assignment $z_n|\theta \sim Mult(\theta)$
  - Draw region visual word $r_n|z_n \sim Mult(\beta_r)$
- For each object class label
  - draw a Wallenius c_i conditioned on contextual constraints given by $p(Y|Z, S) \sim Wall(Y, Z, S)$

where Y is response variable, Z is a set of topics equivalent to candidate objects, and S is context.

The objective is to obtain probability of most semantically consistent labeling configuration Y given topic distribution:

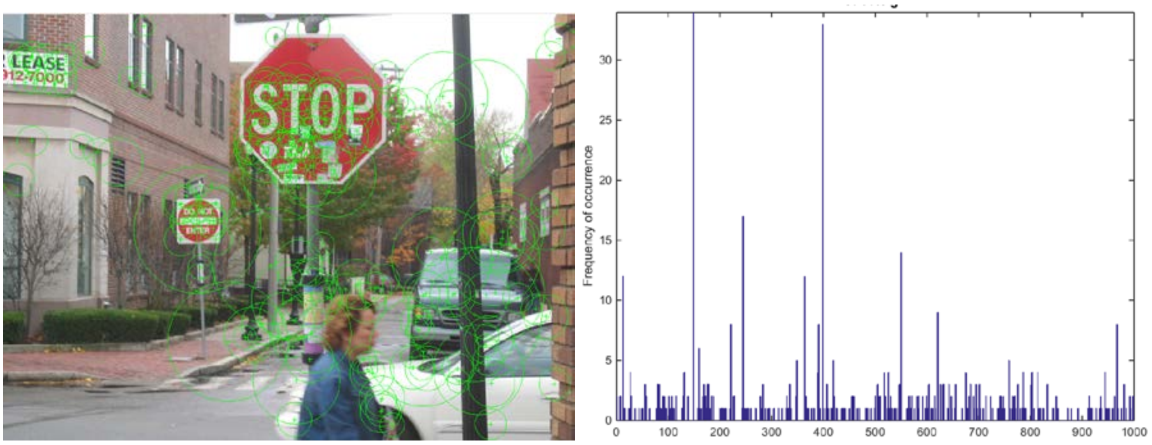
$$\begin{aligned} p(Y|Z, S) &= \Lambda(Y, Z) I(Y, Z, S) \\ \Lambda(Y, Z) &= \prod_{i=1}^{k} \binom{y_i}{z_i} \\ I(Y, Z, S) &= \int_0^1 \prod_{i=1}^{M} \left(1 - t^{\frac{s_i}{\sum_{i=1}^{M} s_i}}\right)^{y_i} dt \end{aligned} \qquad (7)$$

Computing integral in Equation (7) is intractable because of the fractional exponent and must be approximated. First we simplify the formula for binary variables $(\Lambda(Y, Z) = 1)$ as follows:

$$I(Y, Z, S) = \int_0^1 \prod_{j=1}^{M} \left(1 - t^{\frac{s_j}{s}}\right) dt \qquad (8)$$

where $s = \sum_{i=1}^{M} s_i$.

Integrand in Equation (8) can be transformed to an easier to solve polynomial using variable substitution described in Equation (9).



a) Original image **b) Encoded image visual words histogram for the image shown in (a)**

Figure 3- Encoding of a sample image in corresponding visual word frequencies.

The polynomial of Equation (9) can be easily solved by scaling the context values to integer and reducing them by dividing to the greatest denominator.

$$
\begin{aligned}
\mathrm{t} &= \mathrm{u}^{s} \\
dt &= s.\mathrm{u}^{s-1}du \\
p(Y|Z,S) &= \int_0^1 \prod_{j=1}^{M}\left(1 - t^{\frac{s_j}{s}}\right)dt \\
&= \int_0^1 \prod_{j=1}^{M}\left(1 - (\mathrm{u}^{s})^{\frac{s_j}{s}}\right).s.\mathrm{u}^{s-1}du \qquad (9)\\
&= s.\int_0^1 \mathrm{u}^{s-1}.\prod_{j=1}^{M}(1 - u^{s_j})dt
\end{aligned}
$$

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 (or pixels) adjacent to each other are more likely to have the same label or be semantically related. Probability of label assignment follows the given context. This potential takes the form of Potts model ($\mathrm{V_i} \neq \mathrm{V_j}$) to penalize semantically incompatible labels:

$$\psi_p(y_i, y_j) = \begin{cases} 0 & if\ y_i = y_j \\ \lambda_p exp\left(-|l_i - l_j|^2\right) & otherwsie \end{cases} \qquad (10)$$

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.

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 [19]. 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 following:

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

Where C is the clique, $|C_L|$ is cardinality of 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 (2), high-order potential is defined as following:

$$\psi_h(y_i) = \begin{cases} N\lambda_h\vartheta_C\frac{1}{Q} & if\ N \leq Q \\ \lambda_h\vartheta_C & otherwise \end{cases} \qquad (11)$$

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 at from 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 (1), alpha-expansion graph-cut optimization algorithm [25] is applied to get the optimal configuration $y^* = [y_1^*, y_2^*, \ldots, y_n^*]^T$.

$$y^* = arg\,max\,P(y|D) = arg\,max\,min\,E(y) \qquad (12)$$

where y_n^* is unit basis vector that represents the result of object localization for n^{th} object in the image.

Contextual relevance is used during the optimization to eliminate false positives and keep correct detections.

4. EXPERIMENTS

Dataset

Object recognition algorithm was evaluated on subset of SUN397 datasets with 2152 images randomly selected as training set and 2010 images selected as test set from 62 object categories. The metadata of labeled images were used to extract images of objects according to their bounding box information. In pre-processing phase, images were scaled to meet a minimum dimensional constraint.

Training

Image feature space was represented as Bag-of-Features (*BoF*)[26]. Each code-word in the dictionary is a visual appearance feature which was constructed based on “Speeded Up Robust Features” (SURF) [27] algorithm. SURF feature points were obtained 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 3-(b) illustrates BoF representation of an image in (Figure 3-(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 [28]. 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 [19].

Evaluation Methods

For evaluation of context-based CRF framework, multiclass support vector machine (SVM) [29] classification method was used as baseline and compared to the state-of-the-art tree-based contextual model [15] using code provided at their site.

Metrics

Normalized mutual information (NMI) [30] is a metric used to evaluate 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:

$$NMI = \frac{\sum_{h,l} |x_{h,l}| \, log\left(\frac{|X|.\,|x_{h,l}|}{|x_h|.\,c_l}\right)}{\sqrt{\left(\sum_h |x_h| \, log\left(\frac{|x_{h,l}|}{|X|}\right)\right) \sum_l log\left(\frac{c_l}{|X|}\right)}} \tag{13}$$

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 4 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 serious 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 enough to be neglected. The optimum value of K is determined empirically and set to 150.

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

$$F1 = \frac{2 \times Precession \times Recall}{Precession + Recall} \tag{14}$$

Classification performance was evaluated using objects labels in Ground-truth.

Figure 5 shows how this mettric was used in finding optimum parameter values for pairwise and high-order potentials.

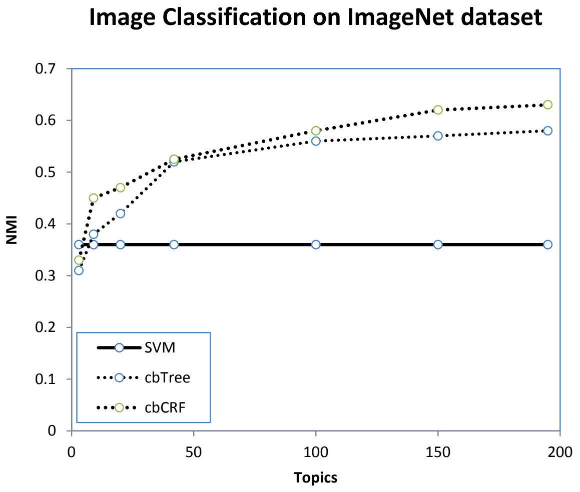


Figure 4 Object detection performance using NMI metric (1.0 is most accurate)

Model Parameters

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

The tuning result on SUN397 is given in the top chart of Figure 5. 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 5, 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.

Result of Empirical Study

To build the framework, a graph was constructed for each scene type to maximize contextual consistency. First scale and location context scores were calculated for all object pairs ($\overline{\tau_{IJ}}$) in that image using Equation (5). Pairwise relations with $\overline{\tau_{IJ}} > 0.5$ were added to the graph and others were ignored. Next, high order context for all cliques combinations (i.e. $\overline{\tau_{1..k}}$) were computed and the clique with highest average score was selected as dominance context. The context model was fitted using Gaussian distribution for each context type which later was used in building CRF model to predict correct label assignment for candidate objects.

Table 1 shows 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.

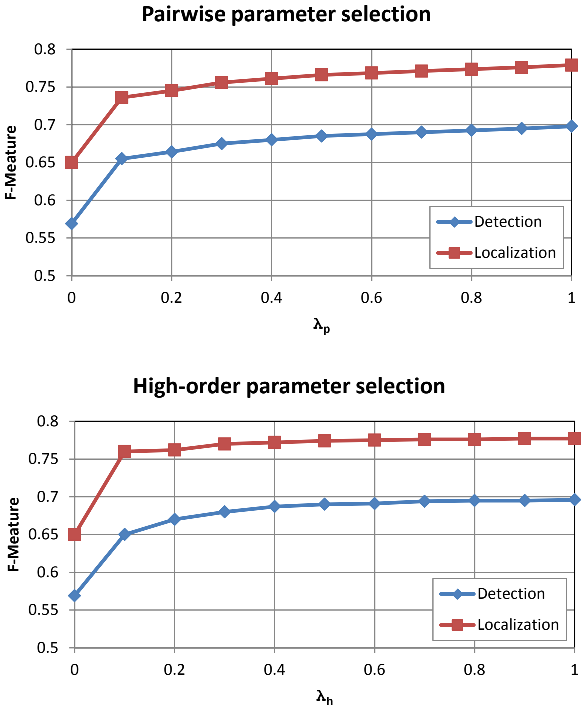


Figure 5. Parameter selection for pairwise and high-order potentials.

The localization improvement over baseline detector algorithm is about 37.2% and the improvement over the state-of-the-art context model (tree context) is 7.5%. For the presence, the corresponding improvements are 37.6%, 14% respectively.

Metric	Localization	Presence
SVM	50.2	57.7
Tree-Based	64.1	69.6
Context-based CRF	**68.9**	**79.4**

Table 2. Object localization and presence performance comparison

Performances of proposed framework for object detection is illustrated in Figure 6 which shows improvement over the tree-context model for most object categories.

Table 2 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 performance of the SVM and CRF in classification of the objects.

5. CONCLUSIONS

In this paper, we presented a discriminative model that combined the power of a generative model as unary potential and used an object-scene relevance score to encode pair-wise and high-order semantic contexts. We showed how to encode the high-order relationship among objects and build a robust models to enforce location, scale and semantical constrains. We compared our framework with other 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 polynomial time. We then applied our framework to distinguish the contextual consistency of the candidate objects using various contextual cues.

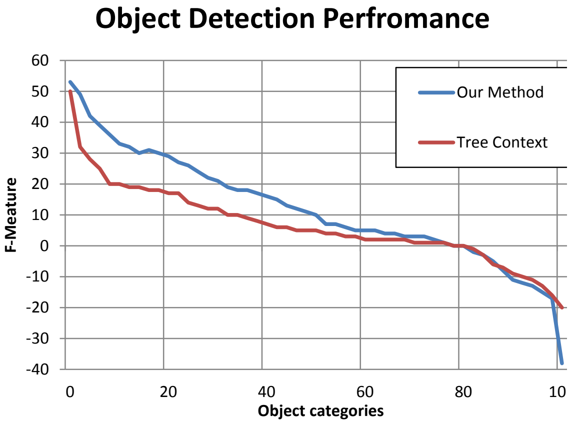
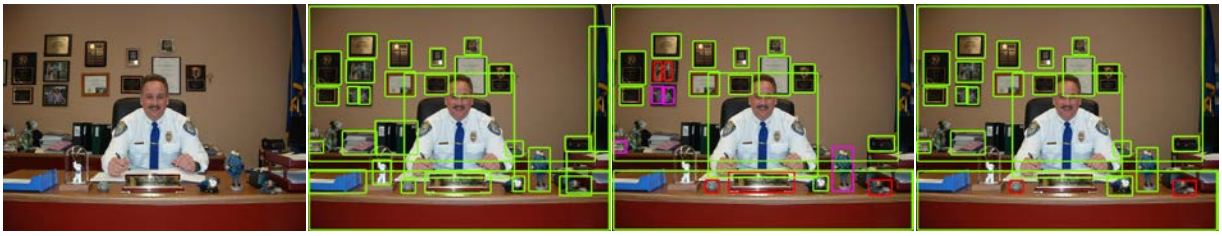


Figure 6. Object detection performance of CBCRF method compare to tree context method using SUN397 dataset.

During our experiments we observed two main weaknesses. First, building a meaningful contextual relevance score requires presence of large number of objects in a scene category with ternary or more interactions. This is limiting factor that restricts choice of training dataset. Second drawback is relatively high computation requirement of this method, which is a side effect of WLDA generative process.

	SVM	**Tree Context**	**Context-based CRF**
Bed	0.53	0.64	0.71
Bicycle	0.59	0.72	0.78
Cabinet	0.44	0.53	0.54
Car	0.58	0.80	0.88
Keyboard	0.52	0.64	0.72
Monitor	0.46	0.63	0.66
Street sign	0.52	0.68	0.72
Table	0.37	0.49	0.50

Table 1- Object detection performance comparison



(a) Original image **(b) Ground Truth** **(c) Baseline Detector** **(d) CBCRF**

Figure 7- Sample annotation result. Green bounding boxes are correctly classified labels.

Our results demonstrated that use of context-based high-ordered potentials has outperformance advantages over the base-line and the state-of-the-art context based object detectors.

6. REFERENCES

[1] Wang C., Blei D. , Fei-Fei L. (2009), Simultaneous Image Classification and Annotation Computer Vision and Pattern Recognition, 2009. Paper presented at the ISSN: 1063-6919 , IEEE CVPR 2009. IEEE Conference on 20-25 June 2009 10.1109/CVPR.2009.5206800

[2] Torralba, A., Contextual Priming for Object Detection, Int'l J. Computer Vision, vol. 53, pp. 169-191, 2003.

[3] Murphy, K.P. Torralba, A. and Freeman, W.T., Using the Forest to See the Trees: A Graphical Model Relating Features, Objects and Scenes, Proc. Neural Information Processing Systems, 2003.

[4] Wang, J., Chen, Z., & Wu, Y. (2011). Action recognition with multiscale spatio-temporal contexts. Paper presented at the 3185-3192. doi:10.1109/CVPR.2011.5995493

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

[6] Zhu, Y., Nayak, N. M., & Roy-Chowdhury, A. K. (2013). Context-aware modeling and recognition of activities in video. Paper presented at the 2491-2498. doi:10.1109/CVPR.2013.322

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

[8] Galleguillos, G. and Belongie, S., Context based object categorization: A critical survey, Comput. Vis. Image Underst., vol. 114, no. 6, pp. 712–722, Jun. 2010.

[9] Marques, O., Barenholtz, E., Charvillat, V.. Context modeling in computer vision: techniques, implications, and applications, Multimedia Tools and Applications, 2011) 51:303–339

[10] Fink M, Perona P (2003), Mutual boosting for contextual inference. In: Thrun S, Saul L, Schökopf B (eds) Advances in neural information processing systems (NIPS). MIT Press, Cambridge, MA

[11] Tang, J Shao, L., and Zhen, X., Robust point pattern matching based on spectral context, Pattern Recognit., vol. 47, no. 3, pp. 1469–1484, 2014.

[12] Jones, S. and Shao, L., Unsupervised spectral dual assignment clustering of human actions in context, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Columbus, OH, USA, 2014, pp. 1–8.

[13] Wolf, L., & Bileschi, S. (2006). A critical view of context. International Journal of Computer Vision, 69(2), 251-261. doi:10.1007/s11263-006-7538-0

[14] Galleguillos, C., McFee, B., Belongie, S., and Lanckriet, G., Multi-class object localization by combining local contextual interactions, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), San Francisco, CA,USA, Jun. 2010, pp. 113–120.

[15] Choi, M. J., Torralba, A., & Willsky, A. S. (2012; 2011). A tree-based context model for object recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(2), 240-252. doi:10.1109/TPAMI.2011.119

[16] Shao. L., Jones. S. , and Xuelong, L., Efficient search and localization of human actions in video databases, IEEE Trans. Circuits Syst. Video Technol., vol. 24, no. 3, pp. 504–512, Mar. 2014.

[17] Chen, G., Ding, Y., Xiao, J., and Han, T., Detection evolution with multi-order contextual co-occurrence, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Portland, OR, USA, Jun. 2013,pp. 1798–1805.

[18] Myeong, H. & Lee, K. M., Tensor-based high-order semantic relationtransfer for semantic scene segmentation, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Portland, OR, USA, Jun. 2013,pp. 3073–3080.

[19] Kohli, P. & Torr, P. H., Robust higher order potentials for enforcing label consistency, Int. J. Comput. Vis., vol. 82, no. 3, pp. 302–324,2009.

[20] Fergus, R., Perona, P., & Zisserman, A., Object class recognition by unsupervised scale-invariant learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, volume 2, pages 264-271, June 2003.

[21] Bengio, S., Dean, J., Erhan, D., Ie, E., Le, Q., Rabinovich, A., Singer, Y. (2013). Using web co-occurrence statistics for improving image categorization.

[22] Lafferty, J., McCallum, A. and Pereira F.. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning, 2001.

[23] Torralba, A., Murphy, K.P., and Freeman, W.T., Contextual Models for Object Detection Using Boosted Random Fields, Advances in Neural Information Processing Systems, MIT Press, 2005.

[24] Hou, Y., He, L. Zhao, X., and Song, D., Pure high-order word dependence mining via information geometry, in Proc. Adv. Inf. Retrieval Theory, Bertinoro, Italy, 2011, pp. 64–76

[25] Boykov, Y. and Veksler, O., Graph cuts in vision and graphics: Theories and applications, in Handbook of Mathematical Models in Computer Vision. New York, NY, USA: Springer, 2006, pp. 79–96.

[26] Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C., Visual categorization with bags of keypoints, in Proc. ECCV Workshop Stat. Learn. Comput. Vis., 2004, pp. 1–22.

[27] Bay, H., Tuytelaars, T., and Van Gool, L.. Surf: Speeded up robust features. In Computer Vision–ECCV 2006, pages 404–417. Springer, 2006.

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

[29] Desai, C., Ramanan, D., and Fowlkes, C. C. Discriminative models for multi-class object layout, Int. J. Comput. Vis., vol. 95, no. 1, pp. 1–12, 2011

[30] Strehl, A., Ghosh, J., and Mooney, R. Impact of similarity measures on Web-page clustering, in Proc. the Workshop on Artificial Intelligence for Web Search, pp. 58-64, Austin: AAAI Press, 2000.