Context Modeling by Discriminative Random Fields for Robust Target Identification From Multi-Modal Sensors

Pranab Banerjee*

Abstract

Target identification is crucial in many defense and national security domains, such as target tracking, surveillance, and gaining situational awareness. In practice, a variety of multi-modal sensors are deployed to support such applications. A critical challenge is to exploit multi-sensor information to robustly identify targets in noisy and imperfect data, such as low contrast imagery under adverse weather conditions. For example, a sensor may identify a set of targets with high confidence, but another sensor may poorly resolve a potential target. The intent is to combine information from all available sensors for robust identification. We tackle the problem by exploiting contextual information based on inter-target relationships in different scenarios of interest, such as desert, forest, urban terrain, under-water zones, cultural festivals, etc. In this paper, we develop discriminative random field based scenario specific contextual models using past labeled sensor data. Subsequently we use these models to quantify the likelihood of an ill-identified target, and determine the most likely identification of a current low confidence target by probabilistic inference using the random field models.

Key Words: Discriminative random field, contextual model, target identification, Conditional random field

1. Introduction

Target identification is a fundamental and crucial task in many application domains relevant to defense and national security. Examples of such applications are surveillance, perimeter security, target tracking from air-borne and space-borne sensors, autonomous navigation, etc. The fundamental goal of target identification is to classify a target as belonging to a particular class or category, such as a tank, a helicopter, a car, a building, and so on. In some cases, the targets need to be identified with a finer level of specificity, such as a T-90 model tank, a T-80 model tank, an M1A1 tank, an M1A2 tank, etc. It is not uncommon to deploy and exploit a variety of sensors of multiple modalities (such as visible band optical, infrared, lidar etc.) for this task. The advantage of exploiting multi-modal sensors is that different modalities can detect different targets, or different aspects of the same target, thus providing a more comprehensive situational awareness. For example, a visible band camera can provide information about the color and texture of an object while a lidar sensor can provide information about the 3D profile and the distance of the object from the sensor. Having more sensor measurements about a target potentially enhances the robustness of identifying the target.

In general, the task of target identification can be decomposed into two high-level sub-tasks: (i) target detection, and (ii) target classification. The goal of the first sub-task, target detection, is to detect the presence of a target within the sensors' field of view, without attempting to determine the type or class of the target. The second task, target classification,

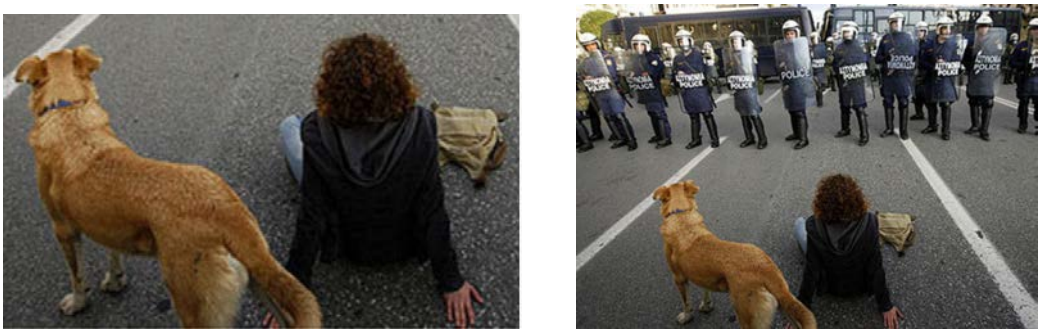
*Boston Fusion Corp., 70 Westview St, Lexington, MA 02421

takes as input the sensor measurements corresponding to the sub-areas of the sensor's field of view where the presence of potential targets were detected by the first sub-task, and classifies the targets using appropriate feature space computation. The focus of this paper is this second sub-task of target classification. We assume that target detection has already been carried out, and bounding boxes or contours for potential target regions are available.

Traditionally, the task of target identification considers each target in isolation, and ignores contextual information such as target co-occurrence, spatial relationships among various targets, relative size, temporal relationships, the surrounding environment, and other such context. However, context of a scene can provide valuable information to facilitate disambiguation of target categories for identification.

1.1 Importance of Context

The importance of contextual information for target/object identification has been well documented in literature (Torralba 2003, Mottaghi 2014). Human understanding of the World around us is heavily dependent on contextual models that we develop through observations about our surroundings. For example, even if we look at a very minimal sketch of a square shape on a wall in a room, our mind almost immediately interprets that as a mirror or a framed photograph but not a book or a laptop since it is not common for us to hang the latter two on a wall. Hence our mental contextual model rules them out. But if we see a sketch of a square on table top, then a book or a laptop is more likely than a photo frame or a mirror. One of the most groundbreaking work in this area was by (Hock 1974). The authors demonstrated that our cognitive biases about arrangement of objects in scenes, their relative sizes, relative locations, and other such contextual information play a major role as cues when we detect real world objects. Our biases resulting from our cognitive models can fool us too under insufficient contextual information. Figure 1 illustrates the important role that context plays in our understanding of a scene. The image on the left is a segment of a larger scene shown in the image on the right. Most of us will likely interpret the image on the left as portraying a peaceful situation where a man is relaxing with his dog. Given the larger context on the right, it is obvious how wrong that perception is!



(a) (b)

Figure 1: Illustration of importance of contextual information in object/target identification. The image on the left in isolation will most likely be interpreted as a scene where a person is relaxing with his dog. However, the interpretation changes significantly when we are presented with the larger context as shown in the image on the right (*Image source*: https://i.ytimg.com/vi/6mMLLO5U1AY/sddefault.jpg)

Since context plays such a crucial role in our ability to detect objects in a scene, our goal in this paper is to exploit such contextual knowledge in identifying targets that are difficult to classify individually. Various factors may result in such low-confidence classifications

in isolation. Examples of such factors are limitations in sensor operational sweet spots, occlusion, and challenging environmental conditions that degrade sensor measurements, such as low contrast, shadows, background clutter etc. The focus of our work is to boost the identification confidence of such poorly identified targets by exploiting the knowledge we have about the other targets in the scene that have been classified with high confidence.

We use a discriminative random field to model the contextual knowledge. In particular, we use a discriminative model based on Conditional Random Fields (CRF) (Lafferty 2001). The advantage of a CRF is that it directly models the conditional distribution of a variable L given an observation variable X, encoding complex dependencies between the two. For context enhanced target identification, the CRF framework can incorporate target appearance features as well as their spatial properties in a unified probabilistic graphical model.

1.2 Previous Work

CRF was first introduced by (Lafferty 2001) with primary application in the domain of natural language processing. Since then, the value of CRFs have been recognized by researchers in other fields, and it has seen significant popularity in the area of computer vision. Some of the notable applications of CRF in this domain are image labeling (He 2004, Huang 2011), and object recognition and segmentation (Shotton 2006). These applications primarily use a 2D CRF graph to process imagery at the pixel level for labeling each pixel as belonging to a class, or segmenting an image by exploiting inter-pixel relationship. (Singhal 2003) used CRF to take into account inter-segment properties as higher level pairwise relationships for scene understanding, but did not consider explicit relative spatial relationships. (Galleguillos 2008) considered high level inter-object appearance as well as spatial relationships, and their work is closest to the approach in this paper. The primary difference of this paper from these previous research is that we use a different graph structure to suit the need of our specific focus, which is to enhance the identification of a poorly classified target in a scene when other targets in the scene have been identified with high confidence. The main contributions of this paper are use of additional pairwise spatial relationship such as relative size, and the use of rotation invariant local binary pattern (LBP) as a discriminating feature for the computation of the pairwise potential in the CRF formulation.

Rest of this paper is structured as follows: Section 2 presents a brief introduction to CRF; Section 3 describes the specific CRF structure used in this paper; Section 4 provides details about our experiment and results; and finally we provide a conclusion.

2. Conditional Random Field (CRF)

This paper uses a discriminative random field based on a Conditional Random Field (CRF) as proposed by (Lafferty 2001). (Sutton 2011) presents an excellent and detailed introduction to CRFs, which are a class of undirected probabilistic graphical models. The graphical structure for a CRF in general has two classes of nodes: (i) nodes representing observations, and (ii) nodes representing output labels or categories. In Figure 2, which represents the graphical structure used for the CRF in this paper, the nodes labeled x_i and S are the observation nodes and the nodes labeled l_i are the output nodes. Here the x_i nodes correspond to the observed features of targets, such as size and texture; the node S encodes the scene category, such indoor, outdoor, etc.; and the nodes l_i correspond to the labels of the targets, such as car, dog, book, etc. The CRF graph structure allows the

computation of the output variables conditioned on the input observations. In Figure 2, the input observation is $X = \{x_i\}_{i=1}^{n} \cup S$. The goal is to develop a model to compute $P(L|X)$ where $L = \{l_i\}_{i=1}^{n}$. The graph structure shows that an output l_i (the label of an object) depends on the features of that object (by the existence of an edge between l_i and x_i), the scene category S, and the relationship with all the other output labels $l_j \in L, j \neq i$ (because of the fully connected nature of the CRF graph). The edge between two output variables encode the relationships between them. Under the CRF framework, the desired joint conditional distribution is computed as

$$P(L|X) = \frac{1}{Z(X)} \cdot exp\left(-E\left(L|X\right)\right) \tag{1}$$

where $Z(X)$ is a partition function and $E\left(L|X\right)$ is a Gibbs potential computed as

$$E\left(L|X\right) = \sum_{i=1}^{n} \lambda_i \phi_u(l_i) + \sum_{i=1}^{n} \sum_{j \in \mathcal{N}(i)} \mu_{ij} \phi_p(l_i, l_j) \tag{2}$$

where $\mathcal{N}(i)$ is the neighborhood of the node l_i in the CRF graph, and λ, μ are the parameters of the model which are learned during training. The optimal output labels conditioned on the input are obtained by minimizing this Gibbs potential.

Here $\phi_u(l_i)$ is a *unary* potential, and $\phi_p(l_i, l_j)$ is a *pairwise* potential. $\phi_u(l_i)$ is computed independently for each object and primarily captures information about the probability of that i-th object appearing in a given scene context. The pairwise potential $\phi_p(l_i, l_j)$ captures the relationship between the i-th and j-th object in a given context.

3. Context Model For Target Identification

The goal of this paper is to build a context model using CRFs for enhancing the identification of individually poorly classified targets. Our goal is to capture scene specific contextual information in the model. In particular, we incorporate object features, co-occurrence, relative spatial information, and scene category in our model. Co-occurrence provides rich knowledge about objects in a scene. For example, we commonly find dogs with people but it is rare for bears to co-occur with people. If this knowledge is captured in the context model, it can be used to disambiguate a target with general characteristics of an animal that appears in spatial proximity of a person. The model will assign a high probability to a dog than a bear based on the contextual knowledge.

Besides co-occurrence, relative spatial relationships provide powerful contextual cues as well. For example, signal lights are normally *above* a car, whereas fire hydrants are usually *at the same level* as cars. Relative size of targets also provide highly useful contextual information. For example, birds are *smaller* than cars, and houses are *larger* than cars. The model presented here takes into account these contextual aspects: object specific features, co-occurrence, relative spatial positions, and relative size among the objects in a scene. Note that only relative vertical spatial relationship is used, since horizontal spatial information does not provide useful information for scenes captured with a ground based sensor (as is the case for the dataset used in our experiment (see Section 4)). The model framework is general enough to allow incorporation of additional contextual information in the future.

A notable strength of our approach is that with proper feature alignment, sensors with multiple modalities can be seamlessly used in this CRF framework for enhanced target

identification. Such feature alignment across modalities is not always trivial. But once it is achieved, the object nodes in the CRF model can be described using components from this aligned feature space, and the rest of the context modeling process will have no difference from the case of uni-modal sensors.

To formally define the problem, let us introduce the following notations. Let $X = \{x_i\}_{i=1}^{n}$ be the set of objects in the training dataset; $S = \{s_i\}_{i-1}^{N_s}$ be the set of scene categories under consideration; given a scene s_i, let $X_H \subset X$ be the set of objects recognized with a high confidence to have category labels $L(X_H)$; and $X_L \subset X$ be the set of objects in the scene that are classified with low confidence. The goal is to enhance the classification of the objects X_L by taking into account the contextual relationships among in X_L and X_H. The contextual model developed here allows this in a probabilistic setting by computing the optimal classification $\hat{L}\left(X_L\right)$ as

$$\arg\max_{\hat{L}} P\left(\hat{L}\left(X_L\right)|L\left(X_H\right), S\right) \tag{3}$$

To construct the probabilistic context model, a CRF with the graph structure shown in Figure 2 was employed. It is a fully connected CRF, where the nodes l_i representing the object category labels are fully connected, i.e., each $l_i \in L$ is connected to every $l_j \in L, j \neq i$. This allows encoding the relationships between any two objects in the scene. The optimal joint probability of output labels conditioned on the input observations and scene category is obtained by minimizing the Gibbs potential described in equation (2). The *unary* potential $\phi_u(l_i)$ is given by the negative log of the probability $P(l_i|x_i, S)$, and the *pairiwse* potential $\phi_p(l_i, l_j)$ is computed as the negative log of $P(l_i, l_j|x_i, x_j, S)$.

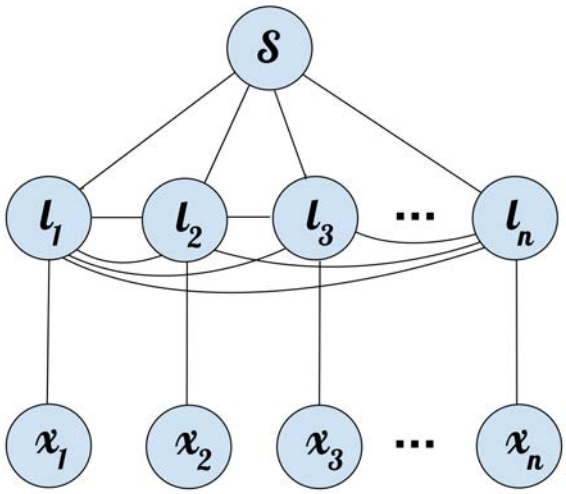


Figure 2: Graph structure for the CRF based context model

4. Experimental Results

Due to unavailability of suitable open, labeled, multi-modal, co-located, ground based sensor data, the openly available COCO 2017 dataset [1] was used to train and test the contextual model. Only the *2017 Train Images* dataset [2] was downloaded, and then split

[1] http://cocodataset.org

[2] http://images.cocodataset.org/zips/train2017.zip

into training and testing subsets. This dataset contains over 118 thousand images. The motivation behind choosing the COCO dataset in our experiment was that it corresponds to complex everyday scenes containing common objects in their natural context, and comes with interesting objects in each image hand-segmented and annotated with category and super-category labels. Objects are labeled using per-instance segmentations to aid in precise object localization. These features align well with our needs since the focus of this paper is to build a model of contextual knowledge about mutual relationships among known (labeled) targets appearing in the context of a scene (such as indoor, outdoor, dinner party, etc.) The dataset spans 91 easily recognizable objects categories, such as person, bicycle, car, motorcycle, airplane, cup, fork, baseball etc. Each category is further clustered into 12 higher level *super-categories*.

Figure 3: An image of dining table scene from the COCO dataset and the corresponding segmented objects

Figure 4: An image of an idoor scene from the COCO dataset and the corresponding segmented objects

To get a glimpse of the COCO dataset, Figure 3 shows an example image from the set. The image on the left is a raw imgae and the one on the right is the version with hand segmented objects. Figure 4 is another example of a COCO image. The one on the left is an image showing an indoor scene, and the one on the right shows the version with the objects in the scene manually segmented with contours and color masks. For each such image, the dataset provides various attributes about the segmented objects. Among those, the following were used in this experiment: (i) the area of objects, (ii) the bounding boxes of objects, (iii) category labels of objects, and (iv) super-category labels of objects.

Since the intent is to build a contextual model that takes into account the overall scene category in addition to relationships among the objects/targets within the context of a scene, it was necessary to have training data corresponding to multiple scene categories. CCOO dataset, however, does not come with subsets aggregated by scene types. So, for this experiment, a simple heuristic was used to classify a set of images as belonging to one of two scene categories: (i) indoor and (ii) outdoor. If an image had any object belonging to the "outdoor" super-category, the image was marked as portraying an outdoor scene, otherwise if the image contained any object belonging to the "indoor" super-category, it was assumed to portray an indoor scene. If neither of these super-categories were present in an image, it was not included in our training data. While this scheme was not perfect and produced some anomalous category memberships, overall it created reasonably accurate clusters as was verified manually via random inspection. Since our context model is a probabilistic one, a few erroneous image memberships would not have a significant impact on the model's predictive capability.

For training the CRF model, the following features were used for the graph nodes (representing the objects in a scene): (i) area, (ii) aspect ratio of the bounding box, and (iii) computed feature vector of length eight corresponding to rotation invariant Local Binary Pattern (LBP) (Ojala 2002) for an object. Each edge in the CRF graph had two features: (i) relative vertical spatial locations (above/below) between the two objects for the edge, and (ii) the relative size (larger/smaller). 2500 images from each of indoor and outdoor scene categories were used for training the CRF. While much larger number of images were available, selection of this training data size was driven primarily by current computational resource constraints. It took about six hours to train this CRF on a desktop with Intel i7 2.67 GHz CPU with 8 cores and 16 GB RAM. We used six of the eight cores for training, leaving the other two cores for auxiliary tasks and essential system processes. This training data size was sufficient to validate the approach presented in this paper. The open source python package *pystruct*[3] was used to build and train the CRF based contextual model, which was then used for inferring the categories of the originally poorly identified objects.

To test the performance of our context enhanced target identification algorithm, the labels of a subset of the objects were ignored, and the algorithm was used to predict them. Since the ground truths were available, it was easy to verify the accuracy of the predictions. Since the contextual model developed here is a probabilistic one, the output of the inference algorithm is a probability distribution over all the categories. The top three categories were considered for evaluating the performance. If the ground truth appeared in the top three of the possible 91 categories, it was considered to be a successful prediction. The top three instead of the single highest category are considered because the currently used object features are relatively coarse for computational efficiency, and we hypothesize that the performance of the contextual model can be further improved by enhancing and fine tuning the node features. Use of other features in addition to the currently used rotation invariant LBP will likely increase the accuracy, and this will be researched in the future. Note that the basic assumption behind the current model is that majority of the target objects in a scene are already recognized with high confidence, and a small number of targets with low confidence need to be recognized with higher accuracy. This is because there is no useful context to exploit where most of the targets have poor classifications to begin with.

Here are a set of results from our experiment. Figure 5 shows an image of a room with bookcases along the farthest wall. A segment of the bookcase, enclosed inside the red box was assumed to be unknown. The true manual annotation for this target object was *book*.

[3]https://pystruct.github.io

Figure 5: An indoor scene from the COCO dataset. The object inside the red box toward the top right was treated as unknown, and inferred using the context model. The top three inferred categories were *book*, *keyboard*, and *clock*.

Figure 6: Same scene as in Figure 5, but with a different target, inside the red box in the center, assumed to be unknown. It was inferred using the context model. The top three inferred categories were *bottle*, *potted plant*, and *apple*.

The top three inferences for this object by the context model were *book*, *keyboard*, and *clock*. In this case, the topmost inference matched the ground truth.

Figure 6 shows the same scene as in Figure 5, but a different target, enclosed by the red box toward the center of the image, was assumed to be unknown. The ground truth, in this case, was *vase*, and the top three inferences by the CRF model were *bottle*, *potted plant*, and *vase* - in that order. Here, the third inferred item matched the ground truth, but the first two were closely related.

Figure 7: An image from the COCO dataset portraying the interior of a store. The teddy bear in the bottom left was treated as an unknown object and predicted using the context model

It was found that objects that appear relatively rarely in the training dataset could result in relatively higher rate of inaccurate inferences. Figure 7 shows an image of a store

interior. When teddy bear inside the red box was treated as an unknown object, the top three inferences were *clock*, *teddy bear*, and *toothbrush* respectively. Based of the limited availability of features for teddy bear, the model did not learn about this object well, and produced relatively visually different matches, such as a toothbrush.

5. Conclusion

This paper has presented a formulation of a discriminative context model based on Conditional Random Field (CRF) for identifying targets in a scene by exploiting domain specific contextual knowledge. The experimental results show that a fully connected CRF with appropriate node and edge features can encode useful contextual knowledge to facilitate target identification. It was shown that rotation invariant Local Binary Pattern (LBP) can be an effective object feature to use for visible band imagery. The context model presented here also exploited inter-object spatial relationships such as relative vertical location and relative size in addition to pure object co-occurrence which has been used by other researchers in the past.

One of the drawbacks of CRFs is the computational complexity. It took over six hours to train the context model using relatively modest number (5000) training images. In the future, we are going to explore heuristics to speed-up this computation.

REFERENCES

Galleguillos, C., Rabinovich, A., and Belongie, S. (2008), "Object categorization using co-occurrence, location and appearance", *Proceedings of the 2008 Conference on Computer Vision and Pattern Recognition, CVPR 2008*.

He, X., Zemel R. S., and Carreira-Perpinan, M. A. (2004), "Multiscale conditional random fields for image labeling", *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition, CVPR 2004*.

Hock, H., Gordon, G. P., and Whitehurst R. (1974), "Contextual relations: The influence of familiarity, physical plausibility, and belongingness", *Perception & Psychophysics*.

Huang, Q., Han, M., Wu, B., and Ioffe, S. (2011), "A hierarchical conditional random field model for labeling and segmenting images of street scenes", *Proceedings of the 2011 Conference on Computer Vision and Pattern Recognition, CVPR 2011*.

Mottaghi, R., Chen, X., Liu, X., Cho, NG, Lee, SW, Fidler, S., Urtasun, R., and Yuille2, A. (2014), "The Role of Context for Object Detection and Semantic Segmentation in the Wild", *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014*.

Ojala, T., Pietikainen, M., and Maenpaa, T. (2002), "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns", *IEEE Transactions on Pattern Analysis and Machine Intelligence* VOL. 24, NO. 7, pp. 971-987.

Shotton, J., Winn, J., Rother, C., and Criminisi, A. (2006), "TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation", *European Conference on Computer Vision, ECCV 2006*, pp, 1-15.

Singhal, A., Luo, J. and Zhu W. (2003), "Probabilistic spatial context models for scene content understanding", *Proceedings of the 2003 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2003*.

Sutton, C. and McCallum, A. (2011), "An Introduction to Conditional Random Fields", *Foundations and Trends in Machine Learning*, Vol. 4, No. 4, pp. 267-373.

Torralba, A.. Murphy, K., Freeman, W, and Rubin, M. A. (2003), "Context-based vision system for place and object recognition", *Proceedings Ninth IEEE International Conference on Computer Vision*, 273-280, DOI: 10.1109/ICCV.2003.1238354.