Relational Learning for Collective Classification of Entities in Images

Anton Chechetka  
Carnegie Mellon University  
Pittsburgh, PA 15213  
antonc@cs.cmu.edu

Denver Dash  
Intel Labs Pittsburgh  
Pittsburgh, PA 15213  
denver.h.dash@intel.com

Matthai Philipose  
Intel Labs Seattle  
Seattle, WA 98105  
matthai.philipose@intel.com

Abstract

We consider the problem of discrete multi-label entity classification in images. We argue that the framework of Markov Logic can provide a unified, well-grounded mechanism to incorporate arbitrary logical relationships between entities to improve classification in images, and thus generalizes much of the recent work on exploiting local and global context in object recognition and scene understanding. Furthermore, we show that Markov Logic can provide a powerful new set of contexts that can relate entities across images in a database for joint classification of all entities in a test set simultaneously. We relate this collective classification of images to graph-based semi-supervised learning approaches, and show that Markov Logic can effectively provide a method to unify context-related work with semi-supervised approaches in a way that neither techniques could easily do on their own. Finally, we show the efficacy of these techniques on a face recognition task on three datasets showing that adding contextual relations dramatically improves accuracy over semi-supervised learning approaches alone.

Introduction

In this paper, we consider the case where one has collected a large sequence of images from one’s daily life, and one wants to collectively label entities in all those images given a small labeled subset. In the process, one wants to take advantage of several varieties of high-level context based on background knowledge, intuition or just common sense. For the sake of concreteness, we consider the problem of person identification, although the ideas in this paper could apply to other types of entity classification in images as well.

In recent years there has been an explosion of work on exploiting in-frame context for entity classification in images (Torralba 2003; Kumar and Hebert 2005; Torralba, Murphy, and Freeman 2005; Heitz and Koller 2008; Heitz et al. 2008; Gupta and Davis 2008; Rabinovich and Belongie 2009; Gould, Gao, and Koller 2009). The work typically involves finding some useful relations for the specific domain at hand, e.g., “the sky is usually above the ground” (Gupta and Davis 2008), building a customized conditional random field model over the entities in a frame and jointly classifying each entity in an image given the observed pixel values. Despite these successes, at present no general purpose method has been reported for contextual reasoning in arbitrary images. It will be no surprise to some in the field of Relational Learning that Markov Logic Networks (MLNs) can be used precisely as this general purpose tool.

This paper makes the above point by showing that MLNs provide a uniform, intuitive and modular interface for specifying relations in first-order logic. More importantly, we show that MLNs can provide a newer more global sense of context that allows them to jointly classify an entire dataset of images (entities), using meaningful relations between these entities, in a manner similar to the collective classification of citation entries done by (Singla and Domingos 2006). The image representation provides a wealth of relations that can be brought to bear on the problem, such as mutual exclusivity of multiple faces in an image, temporal and spatial stratification, personal traits that may relate people to various objects or distinctive clothes, etc. We thus expect that this application is even more suited for the use of a powerful tool like MLNs than the case of citation matching.

This use of MLNs for collective classification resembles graph-based semi-supervised learning (SSL) approaches (Fergus, Weiss, and Torralba 2009), which relate entities across a corpus via a distance or similarity measure. However, compared to SSL approaches, MLNs provides a much richer way of connecting labeled/unlabeled instances, allowing one to combine multiple similarity metrics at the same time as well as incorporate arbitrary logical relationships. In fact we argue that MLNs can provide an approximate generalization to some of the standard SSL approaches by discretizing a distance/similarity measure and incorporating them into the MLN. In addition one can continue to exploit other relations that would not fit well within the SSL framework, such as contextual information that relates entities within frames. We show empirically that this approach yields favorable results for face recognition in images of three datasets collected by us, and that the use of the additional logical relations, which would be difficult in standard SSL, is crucial for the best classification accuracy.

MLN Background

Markov logic (Richardson and Domingos 2006) is a probabilistic generalization of finite first-order logic. A Markov logic network (MLN) consists of a set of weighted first-order clauses. Given a set of constants, an MLN defines a Markov network with one binary variable for every ground atom and one potential for every possible grounding of ev-
very first-order clause. The joint probability distribution over the ground atom variables is defined as

\[
P(x) = \frac{1}{Z} \exp \left\{ \sum_f \sum_{x_f} w_f f(x_f) \right\},
\]

where \( f \) is an indicator function corresponding to a first-order clause (1 if that clause is true and 0 otherwise), \( w_f \) is a weight of that clause, and \( x_f \) is the set of ground atom variables in a particular grounding of that clause. The inner summation in (1) is over all possible groundings. Therefore, for every grounding of every first-order clause, the higher the weight for that clause, the more favored are assignments to \( x \) where that grounding is true.

Two fundamental problems in Markov logic that apply to our application are those of learning optimal weights for the known set of first-order clauses given the knowledge base of known ground atoms, and inference, or finding the most likely assignment to unknown ground atoms given the knowledge base. Even though both problems are intractable in general, well-performing approximate algorithms are available. For weight learning, we used preconditioned conjugate gradient with MC-SAT sampling implemented in the Alchemy package (Kok et al. 2009). For inference, we used a high-performance implementation of residual belief propagation (Gonzalez, Low, and Guerin 2009) along with a lazy instantiation of MLN structure as recommended by (Poon, Domingos, and Sumner 2008).

Model Description

In the existing literature, many types of very different features have been shown to be useful for face recognition (and object recognition more generally). In particular, SSL approaches (Fergus, Weiss, and Torralba 2009) exploit similarity in object appearances in different images to propagate label information from labeled to unlabeled blobs, and between unlabeled blobs. In a supervised setting, typically a low-dimensional representation of blob appearances is extracted, e.g. (Turk and Pentland 1991) and a standard technique such as a support vector machine (Vapnik 1995) is then applied. Besides the blob appearance information, it has been shown that taking context in which the blob appears, such as blob location within the frame or labels of other objects in the scene, is crucial to accurate object recognition. In this section, we show that all the above sources of information can be combined efficiently using a Markov logic network. Our approach thus combines the advantages of the diverse existing approaches to improve face recognition accuracy. In the MLN described below, we will use the query predicate Label(b, o), which is true if and only if blob b has label o. The evidence predicates will be introduced gradually, as they are needed for the MLN rules.

We assume that face detection has already been performed by some standard approach, such as that of (Viola and Jones 2001). The input to our system thus consists of a set of images, and for each image, a set of bounding boxes for the detected faces, some of which are labeled with people’s names. The goal is to assign labels to the remaining unlabeled face blobs.

Label propagation: semi-supervised component

A key idea of the SSL approaches is to classify all the objects of the test set simultaneously by rewarding the cases of similar-looking objects having the same label (equivalently, penalizing labels mismatches for similarly looking objects). Let \( x_i \) and \( x_j \) be the appearances of blobs \( b_i \) and \( b_j \) respectively. Denote \( \|x_i - x_j\| \) to be the distance between \( x_i \) and \( x_j \). We define the evidence predicate SimilarFace(b_i, b_j) that is true if and only if \( \|x_i - x_j\| < \Delta_f \), where \( \Delta_f \) is a threshold. Then the rule to favor matching labels for similar faces is simply

\[
\text{SimilarFace}(b_i, b_j) \land \text{Label}(b_i, o) \Rightarrow \text{Label}(b_j, o)
\]

(2)

We selected threshold \( \Delta_f \) so as to get precision 0.9 on the training data: \[ \frac{\sum_i \sum_j I(\text{SimilarFace}(b_i, b_j)=\text{true})}{\sum_i \sum_j I(\text{Label}(b_i, o)=\text{Label}(b_j, o))} = 0.9, \]

where \( I(\cdot) \) is the indicator function. For simplicity of implementation, we used 16-bin color histograms as representations for \( x_i \) and \( x_j \) and distance \( \|x_i - x_j\|^2 = \sum_k (x_i(k) - x_j(k))^2 \). Naturally, any other choice of representation and distance can be used instead.

Observe that similar face appearance is not the only possible clue that two image fragments actually depict the same person. For example, similar clothing appearance is another useful type of information, as was demonstrated by (Sivic, Zitnick, and Szeliski 2006). In our approach, information about clothing appearance similarity is used in the same way to the face similarity: for every face blob \( b_i \), we define the corresponding torso blob \( t_i \) to be a rectangle right under \( b_i \); the scale of the rectangle is determined by the size of \( b_i \). Let \( y_i \) be the appearance representation of \( t_i \). We define the evidence predicate SimilarTorso(b_i, b_j) which is true if and only if \( \|y_i - y_j\| < \Delta_t \) and introduce the corresponding label smoothing rule

\[
\text{SimilarTorso}(b_i, b_j) \land \text{Label}(b_i, o) \Rightarrow \text{Label}(b_j, o)
\]

(3)

into the MLN. One can see that we have two versions of essentially the same rule exploiting different channels of information for label propagation. Even though it is possible in principle to achieve the same effect in standard graph Laplacian-based SSL approaches (Fergus, Weiss, and Torralba 2009), one would need to use costly cross-validation to find a good way to combine the two separate distance metrics into one (alternatively, find the relative importance of the torso distance and face distance metrics). In contrast, standard algorithms for MLN weight learning provide our approach with the relative importance of the two rules automatically.

More fine-grained label smoothing. One advantage of the graph Laplacian-based unsupervised methods over our approach is that the former naturally support real-valued blob similarity values, while our approach requires thresholding. However, our approach can also be adapted to handle varying degrees of similarity: instead of a single similarity threshold, one can use multiple different similarity thresholds and introduce corresponding similarity predicates. For example, suppose we want to use two different thresholds,
Exploiting single-image context

In addition to the appearance of the blob of interest itself and the labels of similar blobs in other images, powerful contextual cues often exist in the image containing the blob. In the broader context of object recognition, spacial context (e.g., sky is usually in the top part of an image), co-occurrence (computer keyboards tend to occur together with monitors) and broad scene context (fridges usually occur in kitchen scenes) have all been shown to enable dramatic improvements in recognition accuracy. Here, we describe the MLN rules used by our system to take single-image context into account.

A person only occurs once in an image. In the absence of mirrors, for every person at most one occurrence of their respective face is possible in a single image. Therefore, if two faces are present in the same image, they necessarily have to either have different labels, or be both labeled as unknown. Hence we introduce an evidence predicate SameImage(b₁, b₂) which is true if and only if b₁ and b₂ are in the same image, and the following MLN rule:

\[
\text{SameImage}(b₁, b₂) \Rightarrow \text{Label}(b₁, o₁) \lor \text{Label}(b₂, o₂) \lor (o₁ == \text{Unknown})
\] (4)

Face location. For multiple images taken with the same camera pose, such as images from a security camera, often different people will tend to occupy different parts of the frame. For example, in the middle image of Fig. 1 the refrigerator is in the right part of the frame, and the coffee machine is in the middle. Therefore, faces of coffee drinkers may be more likely to appear in the middle of the frame, while those preferring soft drinks may spend more time in the right part. In addition, false-positive face detections (which are given the label “junk”) will appear randomly whereas actual faces appear in more constrained locations. Using the spacial prior in such settings will benefit the recognition accuracy. In our approach, we subdivide every image into 9 tiles of the same size, arranged in a 3 × 3 grid and introduce an evidence predicate InTile(b, tile) and an MLN rule capturing the spacial prior:

\[
\text{InTile}(b, +\text{tile}) \Rightarrow \text{Label}(b, +o)
\]

Notice we use the Alchemy convention +tile and +o, meaning that for every combination of the tile and label a separate formula weight will be learned, yielding different priors over the face labels for different regions of the image.

Time of the day. Similar to face location, a time-dependent label prior is also useful when processing images from security cameras: “early birds” will be more likely to occur in images taken earlier in the day and vice versa. We subdivide the duration of the day into 3 intervals: morning (before 11AM), noon (11AM to 2PM) and evening (after 2PM), introduce an evidence predicate TimeOfDay(b, time) and the corresponding MLN rule:

\[
\text{TimeOfDay}(b, +\text{time}) \Rightarrow \text{Label}(b, +o)
\]
Again, to obtain a time-dependent label prior we force the system to learn a separate weight for every combination of the time interval and face label.

One can see that extracting the relations introduced in this section requires little preprocessing, and it is possible to come up with similar common-sense relations to improve accuracy for settings other than security camera image sequences.

**Plugging in existing face recognizers**

The relations and predicates described so far only use simple representations and similarity metrics. However, there is a large amount of existing literature and expert knowledge dealing with design of representations, distance metrics and integrated face recognition systems that improve accuracy significantly over simpler baselines in a supervised setting. If such a recognition system is available, it is desirable to be able to leverage its results in our framework instead of completely discarding the existing system and replacing it with the MLN model. Fortunately, it is easy to combine any existing face recognition system with our approach by using the face labels produced by the existing system as observations in our model. Formally, we use an evidence predicate $\text{ObservedLabel}(b, \text{observedLabel})$, which is true if and only if the external face recognition system assigned $\text{observedLabel}$ as the label for blob $b$. The MLN rule

$$\text{ObservedLabel}(b, +\text{observedLabel}) \Rightarrow \text{Label}(b, +o)$$

then provides the observation model. Observe that several different external classifiers can be used as observations simultaneously, by mapping the labels produced by different classifiers to disjoint sets of atoms. For example, if there are two different classifiers, $\text{clf}_1$ and $\text{clf}_2$, and both label blob $b_1$ as John, then one would set two ground predicates to true: $\text{ObservedLabel}(b_1, \text{John}_{\text{clf}_1})$ and $\text{ObservedLabel}(b_1, \text{John}_{\text{clf}_2})$. Again, as in the case of multiple measures of blob similarity, MLN weight learning would automatically determine the relative importance and reliability of the two classifiers by assigning corresponding weights to the groundings of the observation model.

**Empirical Results**

In this section, we evaluate our approach on 3 real-world datasets. We show that exploiting the similarity relations and single-image context yields significant accuracy improvements over a state-of-the-art face SSL recognizer (Kveton et al. 2010) that only uses similarity of face blobs for label smoothing. We also demonstrate that all the MLN formulas described in the previous section are important in the sense that removing any one of them from the model reduces accuracy in at least one dataset.

We used a boosted cascade of Haar features as given by (Viola and Jones 2001) for face detection, and face recognizer of (Kveton et al. 2010) as observations for the MLN rule in Eq. 5. This classifier is based on calculating the $L_2$ distance in pixel space for down-sampled (92 × 92 resolution) and normalized images. This method was shown by

**Datasets**

To evaluate our approach, we have collected 3 datasets that are briefly summarized in the following table:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Images</th>
<th>Unique</th>
<th>Timespan</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1720</td>
<td>4</td>
<td>40 min.</td>
<td>1024 × 768</td>
</tr>
<tr>
<td>2</td>
<td>161</td>
<td>9</td>
<td>1 day</td>
<td>640 × 480</td>
</tr>
<tr>
<td>3</td>
<td>268</td>
<td>24</td>
<td>3 days</td>
<td>640 × 480</td>
</tr>
</tbody>
</table>

Dataset 1 was extracted from a video sequence at approximately 1 frame per second (see Fig. 1 for an example image). Datasets 2 and 3 were extracted from two different
security cameras, facing a lab kitchen area (see Fig. 1 for an example image) an entrance door respectively. One can see that images in dataset 1 are of much higher quality, with most faces in a frontal perspective and with a high level of detail. Datasets 2 and 3 have lower resolution faces in a skewed perspective. Also, in dataset 1 the images have been taken much more frequently in time than in datasets 2 and 3. As a result, dataset 1 is the easiest one for face recognition, while datasets 2 and 3 are more challenging, allowing us to check the robustness of our approach. For example, dataset 3 spans the time of 3 days, and contains instances of the same person wearing different clothes, which makes extracting SimilarTorso evidence predicates less reliable. In our experiments, we used 100 images for training for dataset 1, 50 for dataset 2, and 70 for dataset 3. All results have been averaged over 10 random train/test splits. Inference took 30 minutes on dataset 1, and under 10 seconds on datasets 2 and 3.

Results

In Fig. 2 we present the classification error rates on the three datasets described above. We plot the results on the full relational model (leftmost bars) and the baseline semi-supervised approach of (Kveton et al. 2010) (rightmost bars) that only takes face blob similarities, but not context, into account. We also plot the error rates for simplified versions of our model obtained by removing a single evidence predicate and the corresponding relation from the full model to evaluate the importance of individual relations on the classification error. From our results, we make the following conclusions.

Exploiting additional information channels dramatically improves accuracy. Classification error is reduced by our approach by a factor from 1.35 (dataset #3) to 5.2 (dataset #1) compared to the baseline of (Kveton et al. 2010). Such an improvement confirms the long-standing observation that using the context, such as time of the day, is crucial for achieving high recognition accuracy. It also shows that the framework of Markov logic is an efficient way to combine the multiple sources of information, both within a single image, and multiple types of relations between different images, for the goal of face recognition.

No single relation accounts for the majority of the improvement. Over all the dataset, the most extreme single-relation accuracy improvement over the baseline of (Kveton et al. 2010) (InTile predicate and the corresponding location prior on dataset #3) is less than 40% of the total performance improvement of the full model over the baseline. Therefore, the multiple relations of our full model are not redundant and represent information channels that complement each other. It is the interaction of multiple relations that enables significant accuracy improvements.

Relation importance is not uniform across datasets. One can see that the effect of the same relation can be dramatically different for different datasets, depending on those datasets’ properties. Only label propagation via the SimilarTorso relations provides a consistently significant performance improvement, the effect of other relations is much more varied.

For example, InTile predicate and the corresponding location prior that provide almost 40% of total accuracy improvement on dataset #3 have almost no effect on accuracy on dataset #1. The reason for such a disparity is that in dataset #1 all of the subjects move in and out of the same areas of the frame, so the spacial prior is almost uniform, while in dataset #1 people whose offices are on the left of the door usually turn left, and vice versa, so the spacial prior contains a significant amount of information.

Another example is the SameImage predicate and the corresponding mutual exclusivity constraint (4). On dataset #1, where high-quality images lead to very few false-positive face detections, almost every detected face blob corresponds to an actual face and thus a unique label within a frame. On lower-resolution security camera images in datasets #2 and #3, however, there is a significant amount of false positives among detected faces, all having the ground truth label of GarbageFace. When two false positives occur in the same image, the mutual exclusivity constraint prevents both of them from getting a correct GarbageFace label, and labels at least one blob incorrectly. Thus, the accuracy gain from the mutual exclusivity constraint for correctly detected faces is offset by the accuracy decrease on false positives.

The varying degree of relation importance for different datasets makes it important for a face recognition approach to be easily adjustable to emphasize important relations and ignore the unimportant ones. Fortunately, the Markov logic framework makes such adjustability extremely easy on two levels. First, learning the wights of the formulas automatically assigns large weights to important formulas and close to zero weight to irrelevant ones. Second, any relation or formula can be easily taken out of the model or put back in, enabling the search for the optimal set of relations using cross-validation.

Related Work

There exists quite a lot of work now on incorporating relations into image classification. (Rabinovich and Belongie 2009) provides a good overall review of this work, and contrasts “scene-based” and “object-based” context. The former methods are represented by (Torralba 2003; Kumar and Hebert 2005; Heitz and Koller 2008; Heitz et al. 2008), which all attempt to understand the scene (“the gist”) before trying to recognize objects. (Gould, Gao, and Koller 2009) and (Torralba, Murphy, and Freeman 2005) use MRFs to do joint segmentation and object recognition by exploiting physical relations between entities. (Gupta and Davis 2008) uses prepositions present in annotated images to help determine relative positions of objects in images. For example, if an image is annotated with “car on the street”, one might infer that a car is above a street in the image. Many of these efforts have a different aim from our work. Namely, they attempt to do object class detection, i.e., detect all the objects of some given classes in an image; whereas in our face recognition application, we are doing object-instance recognition: given the presence of objects of a given type, find specific labels for those objects. On the other hand, these methods have in common with us the intent to exploit physical relations between objects and abstract relations between
a set of objects and the gist of a scene to improve their results. The difference between their application of this principle and ours is that they all attempt to relate entities across a single image; whereas we use cross-image relationships. Second, by using the framework of Markov Logic, we have a unified, automated mechanism to add arbitrary relations and automatically generate the CRF.

(Fergus, Weiss, and Torralba 2009) and (Kveton et al. 2010) present approximations to the graph Laplacian-based semi-supervised learning solution for classifying images. These methods in general have the advantage over our method that they allow continuous similarity measures rather than our discretized version, and they can be solved efficiently. However, these approaches are typically restricted to similarity-based classification; whereas we can incorporate much more general relations such as our mutual exclusivity. Furthermore, our approach can approximate these approaches (albeit much less efficiently) by using a discretized version of a similarity-measure, as we do using face and torso histograms in this work.

Conclusions

Our contributions in this paper are as follows: First, we show that Markov Logic provides a powerful general-purpose interface to modeling in-frame context for multilabel classification in images. Whereas there has been much existing research showing the benefits of exploiting local and global in-frame context, they all have involved custom-made graphical models and therefore are less accessible as a general modeling tool for specific domains. Second, we show that Markov Logic can also provide a powerful new type of context for collective classification across frames, especially when the database is expected to have many repeated shots of the same entity in different circumstances. We have argued that this type of context generalizes graph-based SSL approaches, and adds much to these approaches in the expressibility of the relations across frames that can guide the collective classification of entities. Thus, we show that Markov Logic can provide a beneficial unification of two quite dissimilar cutting-edge techniques for entity classification in images. Finally, for the specific case of person identification, we have shown empirically that relations such as clothing preferences, mutual exclusivity, spatial and temporal stratification as well as multiple similarity channels can dramatically improve face recognition over the state-of-the-art.

References


Kumar, S., and Hebert, M. 2005. A hierarchical field framework for unified context-based classification. In ICCV.


Singla, P., and Domingos, P. 2006. Entity resolution with markov logic. In ICDM.


Torralba, A.; Murphy, K. P.; and Freeman, W. T. 2005. Contextual models for object detection using boosted random fields. In NIPS.


