Building a Real-Time System for High-Level Person Recognition

Denver Dash
Intel Embedded Systems Science and Technology Center
Carnegie Mellon University
Pittsburgh, PA 15213
denver.h.dash@intel.com

Long Quoc Tran
Georgia Institute of Technology

Abstract

We describe a system that uses graphical models to perform real-time high-level perception. Our system uses Markov Logic Networks to relate entities in images via first-order logical sentences to perform real-time semi-supervised person recognition. The system is a collection of "commodity-level" vision algorithms such as the Viola-Jones face detector, histogram matching and even low-level pixel comparisons, together with logical relationships such as mutual exclusivity and entity confusion combined with a small number of labeled examples into a Markov random field which can be solved to provide labels for faces in the images. We describe the methodology for constructing the logical relations used for the system, and the (many) pitfalls we encountered despite the small number of relations used. We also discuss several future approaches to achieve interactive speeds for such a system, including bounding the size of the graph using temporal weighting of instances, approximating the structure of the graphical model, parallelizing graphical model inference, and low-level hardware acceleration.

Introduction

In this paper, we describe a real-time system[1] (Figure 1) for recognition using a small labeled dataset plus first-order logic relations. The system assumes a constrained environment, i.e., one in which the same people generally occur and that the instances that we want to classify are localized in time.

[1]A batch version of this system along with empirical results was described by Chechetka et al. (2010). Here we focus more on the implementation details as well as the knowledge management of the system.

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 et al., 2005, Heitz and Koller, 2008, Heitz et al., 2008, Gupta and Davis, 2008, Rabinovich and Belongie, 2009, Gould et al., 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 few if any practical real-time systems exist that attempt to do high-level reasoning by integrating context at a high-level. In this paper, we discuss our attempts at building such a system using Markov Logic Networks (MLNs) and by constructing a database of logical relationships that are useful for relating entities to be identified.

This paper also makes the point that MLNs provide a uniform, intuitive and modular interface for performing high-level perception. 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 (c.f., Fergus et al., 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

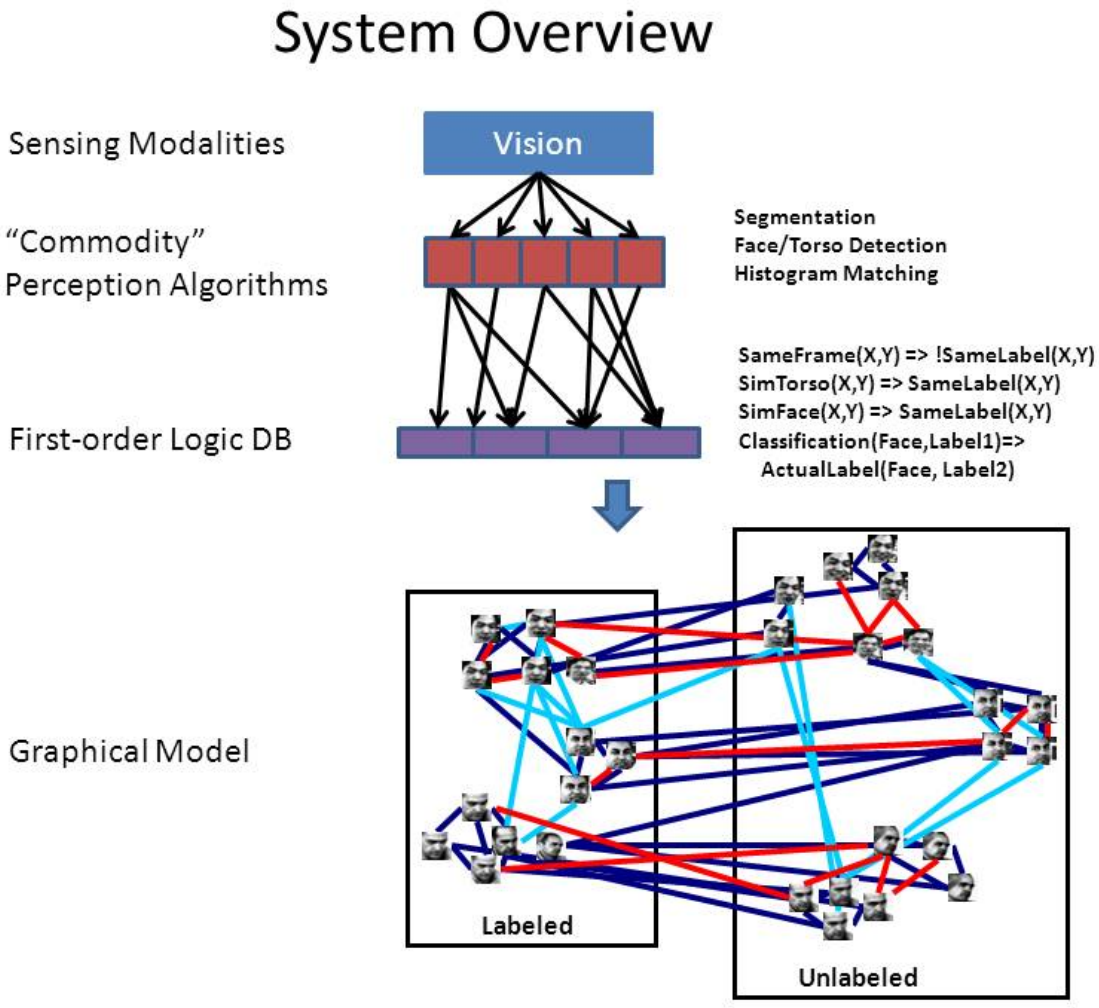

Figure 1: System Overveiw

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 every first-order clause. The joint probability distribution over the ground atom variables is defined as

$$P(x) = \frac{1}{Z} \exp \left\{ \sum_f \sum_{x_i} w_f f(x_i) \right\}, \qquad (1)$$

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_i 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 et al., 2009) along with a lazy instantiation of MLN structure as recommended by Poon et al. (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 et al., 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 $\texttt{Label}(\texttt{b}, \texttt{o})$, which is true if and only if blob $\texttt{b}$ has label $\texttt{o}$. 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 $\texttt{b}_\texttt{i}$ and $\texttt{b}_\texttt{j}$ respectively. Denote $\|x_i - x_j\|$ to be the distance between x_i and x_j. We define the evidence predicate $\texttt{SimilarFace}(\texttt{b}_\texttt{i}, \texttt{b}_\texttt{j})$ that is true if and only if $\|x_i - x_j\| < \Delta_f$, where Δ_f is a threshold. Then the rule to favor matching labels for similar faces is simply

$$\texttt{SimilarFace}(\texttt{b}_\texttt{i}, \texttt{b}_\texttt{j}) \wedge \texttt{Label}(\texttt{b}_\texttt{i}, \texttt{o}) \Rightarrow \texttt{Label}(\texttt{b}_\texttt{j}, \texttt{o}) \qquad (2)$$

We selected threshold Δ_f so as to get precision 0.9 on the training data: $\frac{\sum_{\texttt{i},\texttt{j}} I(\texttt{SimilarFace}(\texttt{b}_\texttt{i},\texttt{b}_\texttt{j})=\texttt{true})}{\sum_{\texttt{i},\texttt{j},\texttt{o}} I(\texttt{Label}(\texttt{b}_\texttt{i},\texttt{o})=\texttt{Label}(\texttt{b}_\texttt{j},\texttt{o}))} = 0.9$, where $I(\cdot)$ is the indicator function. For simplicity of implementation, we used 16-bin color histograms as representations for x_i and χ^2 distance $\|x_i - x_j\|_{\chi^2} \equiv \sum_{k=1}^{\#bins} \frac{(x_i(k)-x_j(k))^2}{x_i(k)+x_j(k)}$. 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 an-

other useful channel of information, as was demonstrated by Sivic et al. (2006). In our approach, information about clothing appearance similarity is used in the same way to the face similarity: for every face blob $\mathtt{b_i}$, we define the corresponding torso blob $\mathtt{t_i}$ to be a rectangle right under $\mathtt{b_i}$; the scale of the rectangle is determined by the size of $\mathtt{b_i}$. Let y_i be the appearance representation of $\mathtt{t_i}$. We define the evidence predicate $\mathtt{SimilarTorso}(\mathtt{b_i}, \mathtt{b_j})$ which is true if and only if $\|y_i - y_j\| < \Delta_t$ and introduce the corresponding label smoothing rule

$$\mathtt{SimilarTorso}(\mathtt{b_i}, \mathtt{b_j}) \wedge \mathtt{Label}(\mathtt{b_i}, \mathtt{o}) \Rightarrow \mathtt{Label}(\mathtt{b_j}, \mathtt{o}) \quad (3)$$

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 et al., 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, $\Delta_f^{(1)} < \Delta_f^{(2)}$, for face blob similarity. Then we would introduce two similarity predicates, $\mathtt{SimilarFace}^{(1)}(\mathtt{b_i}, \mathtt{b_j})$, which is true if and only if $\|x_i - x_j\| < \Delta_f^{(1)}$, and analogously $\mathtt{SimilarFace}^{(2)}(\mathtt{b_i}, \mathtt{b_j})$, for $\Delta_f^{(2)}$. Then for highly similar blobs, those with $\|x_i - x_j\| < \Delta_f^{(1)}$, both versions of the formula in Eq. 2 for $\mathtt{SimilarFace}^{(1)}$ and $\mathtt{SimilarFace}^{(2)}$ will have the left-hand side to be true, providing a higher reward for matching the labels. On the other hand, for weakly similar blobs, those with $\Delta_f^{(1)} < \|x_i - x_j\| < \Delta_f^{(2)}$, only the version of Eq. 2 corresponding to $\mathtt{SimilarFace}^{(2)}$ will have the LHS to be true, providing a weaker reward for matching labels.

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.

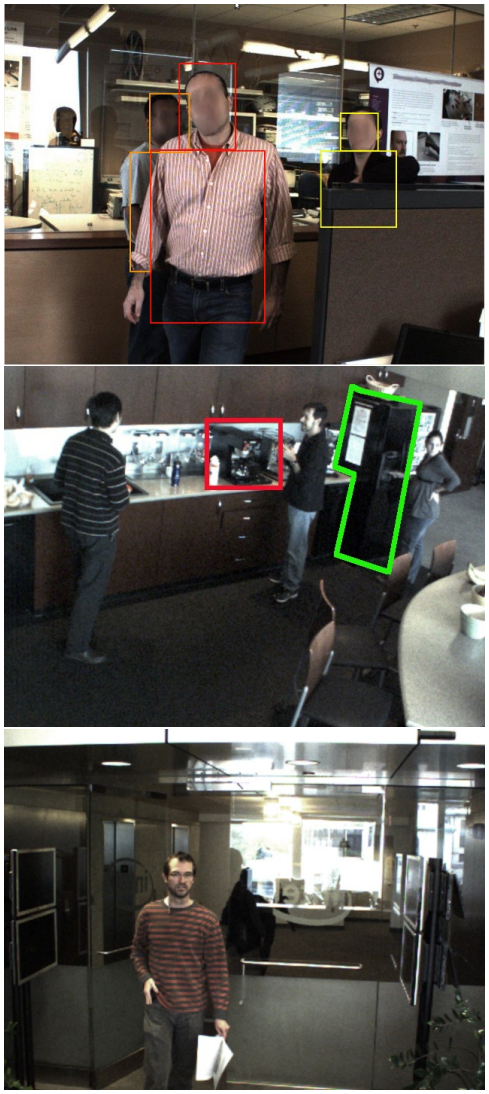

Figure 2: Example security images for datasets 1–3 (top to bottom). The top image shows an example of torso extraction (faces on this set have been blurred by subjects' request). The middle image shows a view of a kitchen area where a coffee machine (red) is in the middle of the frame, while the refrigerator (green) is on the right; thus coffee drinkers might be more likely to appear in the middle.

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

$\texttt{SameImage}(\texttt{b}_\texttt{i}, \texttt{b}_\texttt{j})$ which is true if and only if $\texttt{b}_\texttt{i}$ and $\texttt{b}_\texttt{j}$ are in the same image, and the following MLN rule:

$$\texttt{SameImage}(\texttt{b}_\texttt{i}, \texttt{b}_\texttt{j}) \Rightarrow !\texttt{Label}(\texttt{b}_\texttt{j}, \texttt{o}_1) \vee !\texttt{Label}(\texttt{b}_\texttt{j}, \texttt{o}_2) \vee (\texttt{o}_1 != \texttt{o}_2) \vee (\texttt{o}_1 == \texttt{Unknown}) \quad (4)$$

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. 2 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 $\texttt{InTile}(\texttt{b}, \texttt{tile})$ and an MLN rule capturing the spacial prior:

$$\texttt{InTile}(\texttt{b}, +\texttt{tile}) \Rightarrow \texttt{Label}(\texttt{b}, +\texttt{o})$$

Notice we use the Alchemy convention $+\texttt{tile}$ and $+\texttt{o}$, 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 $\texttt{TimeOfDay}(\texttt{b}, \texttt{time})$ and the corresponding MLN rule:

$$\texttt{TimeOfDay}(\texttt{b}, +\texttt{time}) \Rightarrow \texttt{Label}(\texttt{b}, +\texttt{o})$$

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 $\texttt{ObservedLabel}(\texttt{b}, \texttt{observedLabel})$, which is true if and only if the external face recognition system assigned $\texttt{observedLabel}$ as the label for blob b. The MLN rule

$$\texttt{ObservedLabel}(\texttt{b}, +\texttt{observedLabel}) \Rightarrow \texttt{Label}(\texttt{b}, +\texttt{o}) \quad (5)$$

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, $\texttt{clf}_1$ and $\texttt{clf}_2$, and both label blob $\texttt{b}_1$ as $\texttt{John}$, then one would set two ground predicates to true: $\texttt{ObservedLabel}(\texttt{b}_1, \texttt{John\_clf}_1)$ and $\texttt{ObservedLabel}(\texttt{b}_1, \texttt{John\_clf}_2)$. 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.

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 Sim et al. (2000) to be generally superior to the more common method based on PCA for face classification in single images. For evaluating torso similarity for $\texttt{SameTorso}$ evidence predicate, simple torso occlusion handling was performed by assuming that larger faces were in the foreground. Thus, larger-faced torsos were assumed to lie in front of smaller-faced torsos, and the resulting torso bounding boxes did not intersect (see Fig. 2 for an example).

Results

Quantitative results for a batch version of this model were presented by Chechetka et al. (2010). Here, for some added context, we just present some of the qualitative lessons learned from those experiments.

Exploiting additional information channels dramatically improves accuracy. Classification error is reduced by our approach by a factor from 1.35 to 5.2 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 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. 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 weights 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.

Building a Real-time system

In this section, we explore how the ideas in this paper can be augmented into a real-time system. There are two broad objectives that need to be addressed:

1. Updating the model as new instances come in (online learning).
2. Performing graphical model inference at interactive speeds (online inference).

To perform these two tasks simultaneously, we chose an asynchronous architecture (Figure 3) where learning and inference are performed in separate processes. This provided a natural parallelism for the whole system. Even with this parallelized approach, both learning and inference components required special enhancements to enable real-time operation. Online learning was necessary whenever a new labeled instance was observed by the system. This could happen whenever Incorporating new instances caused the graph structure to change.

Online inference was triggered whenever a new unlabeled instance was observed. In this case, the structure of the graph was altered. Specifically, new nodes corresponding to new instantiations of all propositions involving the new observed faces will be added to the network. At this point, since the structure of the network has changed, the beliefs of the network are necessarily invalidated. Thus, an exact algorithm would run belief propagation over the entire graph after such events occurred. In our system, we avoided this with the following heuristic: we maintained the current beliefs of the network (as of the last iteration), and we pushed the beliefs of the new nodes on the top of the priority queue in the Residual BP calculation. This had the effect of focusing the next round of computation on the new nodes until convergence was reached.

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 et al. (2009) and Torralba et al. (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 et al. (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

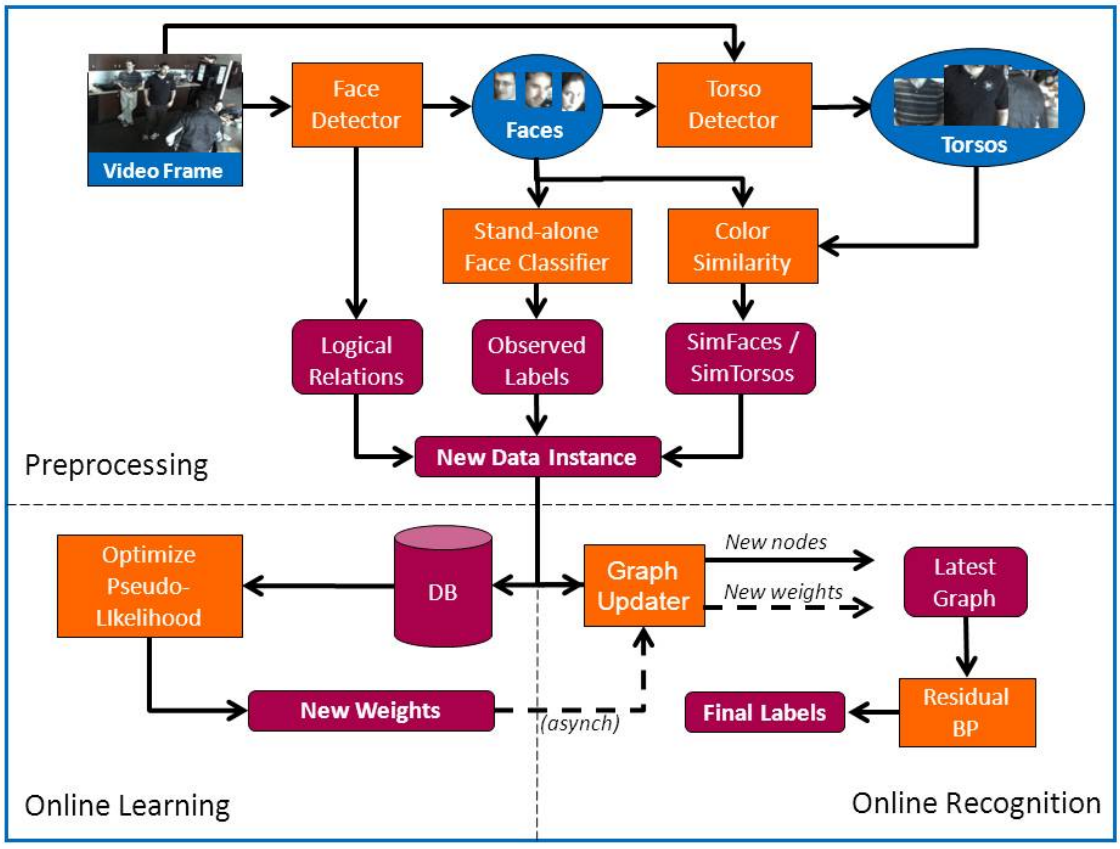

Figure 3: Real time system architecture

to similarity-based classification; whereas we can incorporate much more general relations such as our mutual exclusivity. Furthermore, our approach can easily incorporate any of these classifiers (as we do in this paper by taking the classifier of Kveton et al. (2010)) and use them as core face recognizers in an object model. Finally, 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 present a real-time perception system that incorporates Markov Logic 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. Although much work remains to be done, we present some of the specific modeling issues involved with this system, as well as some of the obstacles to making the system operate at interactive speeds.

References

A. Chechetka, D. Dash, and M. Philipose. Relational learning for collective classification of entities in images. In *Workshop on Statistical Relational AI in conjunction with the Twenty-Fourth Conference on Artificial Intelligence (AAAI-10)*, Atlanta, Georgia, 2010.

R. Fergus, Y. Weiss, and A. Torralba. Semi-supervised learning in gigantic image collections. In *NIPS*. 2009.

J. Gonzalez, Y. Low, and C. Guestrin. Residual splash for optimally parallelizing belief propagation. In *AISTATS*, 2009.

S. Gould, T. Gao, and D. Koller. Region-based segmentation and object detection. In *NIPS*. 2009.

A. Gupta and L. S. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*, 2008.

G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV*, 2008.

G. Heitz, S. Gould, A. Saxena, and D. Koller. Cascaded classification models: Combining models for holistic scene understanding. In *NIPS*. 2008.

S. Kok, M. Sumner, M. Richardson, P. Singla, H. Poon, D. Lowd, and P. Domingos. The alchemy system for statistical relational AI. Technical report, Department of Computer Science and Engineering, University of Washington, Seattle, WA., 2009. URL `http://alchemy.cs.washington.edu/`.

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

B. Kveton, M. Valko, A. Rahimi, and L. Huang. Semi-supervised learning with max-margin graph cuts. In *to appear, AISTATS*, 2010.

H. Poon, P. Domingos, and M. Sumner. A general method for reducing the complexity of relational inference and its application to mcmc. In *AAAI*. AAAI Press, 2008.

A. Rabinovich and S. Belongie. Scenes vs. objects: a comparative study of two approaches to context based recognition. In *International Workshop on Visual Scene Understanding (ViSU)*, Miami, FL, 2009.

M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62(1–2):107–136, Feb 2006.

T. Sim, R. Sukthankar, M. Mullin, and S. Baluja. Memory-based face recognition for visitor identification. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, 2000.

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

J. Sivic, C. L. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In *Proceedings of the British Machine Vision Conference*, 2006.

A. Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53(2):169–191, July 2003.

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

M. A. Turk and A. P. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

P. Viola and M. Jones. Robust real-time object detection. In *International Journal of Computer Vision*, 2001.