Discovery of Social Relationships in Consumer Photo Collections using Markov Logic

Parag Singla
University of Washington
parag@cs.washington.edu

Henry Kautz
University of Rochester
kautz@cs.rochester.edu

Jiebo Luo
Eastman Kodak Company
jiebo@kodak.com

Andrew Gallagher
Eastman Kodak Company
gallagher@kodak.com

Abstract

We identify the social relationships between individuals in consumer photos. Consumer photos generally do not contain a random gathering of strangers but rather groups of friends and families. Detecting and identifying these relationships are important steps towards understanding consumer image collections. Similar to the approach that a human might use, we use a rule-based system to quantify the domain knowledge (e.g. children tend to be photographed more often than adults; parents tend to appear with their kids). The weight of each rule reflects its importance in the overall prediction model. Learning and inference are based on a sound mathematical formulation using the theory developed in the area of statistical relational models. In particular, we use the language called Markov Logic [14]. We evaluate our model using cross validation on a set of about 4500 photos collected from 13 different users. Our experiments show the potential of our approach by improving the accuracy (as well as other statistical measures) over a set of two different relationship prediction tasks when compared with different baselines. We conclude with directions for future work.

1. Introduction

Consumer photo collections are pervasive. Mining semantically meaningful information from such collections has been an area of active research in machine learning and vision communities. There is a large body of work focusing on problems of face detection and recognition, detecting objects of certain types such as grass, water, and sky. Most of this work relies on using low level features (such as color or texture.) available in the image. In recent years, there has been an increased focus on extracting semantically more complex information such as scene detection and activity recognition [9, 8]. For example, one might want to cluster pictures based on whether they were taken outdoors or indoors or cluster work pictures from leisure pictures. This work relies primarily on using the derived features such as people present in the image, the presence or absence of certain kinds of objects in the image, and so on. Typically, the power of collective inference [16, 11] is used in such scenarios. For example, it may be hard to tell for a particular picture if it is work or leisure, but by looking at other pictures that are similar in a spatio-temporal sense it becomes easier to arrive at the correct conclusion. This line of research aims to revolutionize the way we perceive the digital photo collection - from a bunch of pixel values to highly complex and meaningful objects that can be queried and organized in meaniungful ways.

Our goal is to automatically detect social relationships in consumer photo collections. For example, given two faces appearing in an image, we would like to infer if the two people are spouses or merely friends. Even when information about age, gender, and identity of various faces is known, this task seems extremely difficult. In related work [12, 3], the intention is to learn the likelihood that particular groups will appear together in images to facilitate recognition. In [17], movies are analyzed to find groups of characters that are associated. In all of these works, the relationships within the groups are left undefined.

Given a single image of two people, it is difficult to determine whether the two are friends or spouses. However, when a group of pictures are considered collectively, the task becomes more tractable. An observer unfamiliar with the photos' subjects might guess based on the rules of thumb such as (a) couples often tend to be photographed just by themselves as opposed to friends who typically appear in groups, and (b) couples with young children often appear with their children in the photos. The beauty of this approach is that one can even infer meaningful things about relationships between people who never (or very rarely) are photographed together. For example, if A (male) appears with a child in a group of images and B (female) appears with the same child in other images, and A and B appear together in a few other images, then is it likely they share a spousal relationship and are the child's parents.

We would somehow like to capture these rules of thumb in a meaningful way. There are several issues that need to be taken into account while doing this: (a) after all, these are "rules of thumb" and may not always be correct; (b) many rules may need to be simultaneously considered; so a mechanism for combining rules must be defined; and (c) in certain scenarios, multiple rules may conflict with each other. To address these issues, we turn to Markov Logic [14], which provides a framework for combining first order logic rules in a mathematically sound way. Each rule is seen as a soft constraint (as opposed to a hard constraint in logic) whose importance is determined by the real-valued weight associated with it; the higher the weight, more important the rule is. In other words, given two conflicting rules, the rule with higher weight should be more strongly believed, other things being equal. The weights can be learned from training data. Further, Markov logic also provides the ability to learn new rules using the data, in addition to the rules supplied by the domain experts, thereby enhancing the background knowledge. These learned rules (and their weights) are then used to perform a collective inference over the set of possible relationships. As we will show, we can also build a model for predicting relationships, age and gender, using noisy predictors (for age, and gender) as inputs to the system. Inference is performed over all images, age, and gender simultaneously to allow each component to influence the others in the context of our model, thereby improving the accuracy of the prediction.

The remainder of the paper is organized as follows. We first present the background on Markov logic. This is followed a detailed description of our model for the problem. We then describe out datasets and experiments. We conclude with the directions for future work.

2. Markov Logic

Statistical relational models [4] combine the power of relational languages such as first order logic and probabilistic models such as Markov networks. These models are capable of explicitly modeling the relations in the domain (for example, various social relationship in our case) and also explicitly take uncertainty (for example, rules of thumb may not always be correct) into account. There has been a large body of research in this area in recent years.

One of the most powerful such models is Markov Logic [14]. It combines the power of first-order logic with Markov networks to define a distribution over the properties of underlying objects (e.g. age, gender, facial features in our domain) and relations (e.g. various social relationships in our domain) among them. This is achieved by attaching a real valued weight to each formula in a first-order theory, where the weight (roughly) represents the importance of the formula. Formally, a *Markov Logic Network L* is defined as a set of pairs (F_i, w_i), F_i being a formula in first-order logic and w_i a real number. Given a set of constants C, the probability of a particular configuration $\mathbf{x}$ of the set of ground predicates $\mathbf{X}$ is given as

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{m} w_i n_i(\mathbf{x})\right) \qquad (1)$$

where the sum is over all the formulas appearing in L, w_i is the weight of the i^{th} formula and $n_i(\mathbf{x})$ is the number of its true groundings under the assignment $\mathbf{x}$. Z is the normalization constant. For further details, see [14].

Inference corresponds to finding the marginal probability of each ground predicate. Exact inference is intractable in such models and sampling is often used. Poon and Domingos[13] propose a novel MCMC-based sampling method called MC-SAT, which uses slice-sampling to efficiently jump between two different regions (possibly disconnected) of the probability space. The results are better than plain Gibbs sampling, especially when deterministic dependencies (rules that are always true) are present. In this case, Gibbs sampling is stuck in one region of the solution space but MC-SAT has no problem in exploring disconnected regions.

Parameter learning in the model corresponds to finding the weights (maximum likelihood or MAP) for each formula. Weights can be learned generatively as detailed by Richardson and Domingos[14]. Since likelihood is diffcult to optimize, they optimize another measure called pseudo-likelihood [1], which is simply the product of the likelihoods of each variable given its Markov blanket. Pseudolikelihood ignores the long-range dependencies among variables and hence gives sub-optimal results when such dependencies are present. A better approach is to optimize the likelihood itself and learn the weights discriminatively as proposed by Singla and Domingos[15], which uses a voting perceptron algorithm for gradient descent or the recently proposed algorithm of Lowd and Domingos[10], which uses scaled conjugate gradient as the optimization scheme. Note that in discriminative learning, we typically optimize a much simpler function as there is no need to model the dependencies between evidence variables (unlike in generative learning). Further, discriminative learning directly optimizes the function to be used at inference time and hence tends to give better results in practice.

Kok and Domingos[5] propose an algorithm for doing structure learning of Markov logic networks, i.e., discovering new rules in the domain. This is similar to learning ILP style rules for first order logic. Rather than first learning structure and then the weights, Kok and Domingos[5] optimize both the structure and weights together, which gives better results for obvious reasons.

Domingos *et al.* [2] is a comprehensive source that details the various algorithms described above. Finally, all the above algorithms are implemented in a software package called “alchemy” [6] which is publicly available on the web. We use this software (with minor changes in the structure learning algorithm) to perform our experiments.

3. Model

In this section, we will describe our model for predicting the social relationships in consumer image collections. Our model is expressed in Markov logic. We will describe our objects of interest, predicates (properties of objects and the relationships among them), and the rules that impose certain constraints over those predicates. We will then pose our learning and inference tasks.

3.1. Objects and Predicates

3.1.1 Objects

We have three kinds of objects in our domain:

- **Person:** A real person in the world.
- **Face:** A specific appearance of a face in an image.
- **Image:** An image in the collection.

We model two kinds of predicates defined over the objects of interest.

3.1.2 Evidence Predicates

The value of these predicates is known at the time of the inference through the data. An example evidence predicate is *OccursIn(face,img)*, which describes the truth value of whether a particular face appears in a given image or not. We use the evidence predicates for the following properties/relations:

- Number of people in an image: *HasCount(img,cnt)*
- The age of a face appearing in an image : *HasAge(face,age)*
- The gender of a face appearing in an image : *HasGender(face, gender)*
- Whether a particular face appears in an image : *OccursIn(face, img)*
- Correspondence between a person and his/her face : *HasFace(person, face)*

The age (gender) of a face is the estimated age (gender) value estimated from the face from an image. Note that this is distinct from the actual age (gender) of a person, which is modeled as a query predicate. The age (gender) associated with a face is inferred from a classification model trained separately on a collection of faces using various facial features. We implemented age and gender classifiers following the examples of [7] and [18]. Note that different faces associated with the same person will likely have different estimated age and gender values. We model the age using five discrete bins: child, teen, youth, adult, and senior adult.

For this application, we assume the presence of face detection and recognition and therefore we know exactly which face corresponds to which person. Relaxing this assumption and folding face detection and recognition into the model is an important piece of future work.

3.1.3 Query Predicates

The value of these predicates is not known at the time of the inference and needs to be inferred. An example of this kind of predicates is *HasRelation(person1, person2, relation)*, which describes the truth value of whether two persons share a given relationship. We use the following query predicates:

- The relationship between two persons: *HasRelation(person1, person2, relation)*
- Age of a person: *PersonHasAge(person, age)*
- Gender of a person: *PersonHasGender(person, gender)*

We model two different kinds of settings.
Two-Class: In this basic setting, we model only two relationships in the domain: friend or relative. All the relations are mapped onto one of these two relationships. Relative follows the intuitive definition of being a relative. Anyone who is not a relative is considered as a friend.
Multi-Class: In the second setting, we model seven different kinds of social relationships: relative, friend, acquaintance, child, parent, spouse, and childfriend. Relative includes any relative (by blood or marriage) not including child, parent, or spouse. Friends are people who are not relatives and satisfy the intuitive definition of the friendship relation. Childfriend models the friends of children. It is important to model the childfriend relationship, as children are pervasive in consumer photo collections and often appear with their friends. In such scenarios, it becomes important to distinguish between children and their friends. Any non-relatives, non-friends are acquaintances. For our purposes and in our data, spouses always have opposite gender.

3.2. Rules

We formulate two kinds of rules for our problem. All the rules are expressed as formulas in first order logic.

3.2.1 Hard Rules

These describe the hard constraints in the domain, i.e., they should always hold true. An example of a hard rule is *OccursIn(face, img1) and OccursIn(face, img2)* $\Rightarrow$ *(img1 = img2)*, which simply states that each face occurs in at most one image in the collection (for brevity, we mention only the hard rules that contain at least one query predicate.)

- Parents are older than their children.
- Spouses have opposite gender.
- Two people share a unique relationship among them.

Note that we model that there is a unique relationship between two people. Relaxing this assumption (e.g. two people can be both relatives as well as friends) is a part of the future work.

3.2.2 Soft Rules

These rules describe the interesting set of constraints that we expect to usually, but not always, be true. An example of a soft rule is *OccursIn(person1, img) and OccursIn(person2, img)* $\Rightarrow$ *!HasRelation(person1, person2, acquaintance)*. This rule states that two people who occur together in a picture are less likely to be mere acquaintances. Each additional instance of their occurring together (in different pictures) further decreases this likelihood. Here are some of the other soft rules that we use:

- Children and their friends are of similar age.
- A young adult occurring solely with a child shares the parent/child relationship.
- Two people of similar age and opposite gender appearing together (by themselves) share spouse relationship.
- Friends and relatives are clustered across photos: if two friends appear together in a photo, then a third person occurring in the same photo is more likely to be a friend. The same holds true for relatives.

Proper handling of the combination of hard and soft constraints together provides the power of our approach. We seek a solution that satisfies all the hard constraints (presumably such a solution always exists) and at the same time satisfies the maximal number (weighted) of soft constraints.

3.2.3 Singleton Rule

Finally, we have a rule consisting of a singleton predicate *HasRelation(person1,person2, +relation)* (+ indicates a unique weight is learned for each relation), that essentially represents the prior probability of a particular relationship holding between any two random people in the collection. For example, the "friends" relationship is more common than the "parent-child" relationship. Similarly, we have the singleton rules *HasAge(person, +age* and *HasGender(person, +gender)*. These represent (intuitively) the prior probabilities of having a particular age and gender, respectively. For example, it is easy to capture the fact that children tend to be photographed more often by giving a high weight to the rule *HasAge(person, child)*.

3.3. Learning and Inference

Given the model (the rules and their weights), inference includes finding the marginal probability of query predicates *HasRelation*, *HasGender* and *HasAge* given all the evidence predicates. Since we have to handle a combination of hard (deterministic) and soft constraints, we use the MC-SAT algorithm of Poon and Domingos [13] as implemented in alchemy.

Given the hard and soft constraints, learning corresponds to finding the optimal weights for each of the soft constraints. We find the MAP weights with a Gaussian prior centered at zero. We use an off-the-shelf learner [10] as implemented in alchemy [6]. We use the structure learning algorithm of Kok and Domingos[5] to refine (and learn new instances) of the rules that help us predict the target relationships. The original algorithm as described by them (and as implemented in alchemy) does not allow the discovery of partially grounded clauses. This is important for us as we need to learn the different rules for different relationships. The rules may also differ for specific age groups and/or gender (for example, one might imagine that males and females differ in terms of with whom they tend to be photographed). The change needed in the algorithm to have this feature is straightforward. We allow the addition of all possible partial groundings of a predicate as we search for the extensions of a clause. (Only certain variables (i.e. relationship, age, and gender) are allowed to be grounded in these predicates to avoid blowing up the search space.) After learning, the inference proceeds as previously described.

4. Experiments

4.1. Dataset

We experimented on a dataset consisting of personal photo collections of 13 different owners. The collectors came from diverse demographics in terms of age, gender, and ethnicity. There are about a couple hundred photos for each collector and the total number of images is 1926 (we account only for photos in which at least one person appears). The faces are detected and recognized, i.e. there is a one-to-one mapping between a face and a person in the real world. Each collection owner labeled their relationship to each person appearing in their photo collection. The relationship label domain is a set of 30 different relationships such as mother, father, son, daughter, friend, son-in-law, and so on. Each collection owner also labeled the current age of each person appearing in the collection. The gender of each person was labeled by looking at their name and appearance.

4.2. Models

We compared five different models as explained below.

- **Random:** This is the simplest of all models. We randomly select, with a uniform distribution, one of the all possible relationships as the detected relationship.
- **Prior:** Using the prior distribution from the training data, one of the all possible relationships is selected as the detected relationship.
- **Hard:** This model picks uniformly at random one of the possible relationships while enforcing the common-sense hard constraints. For example, if two people are of same gender, then the spouse relationship is ruled out and we can pick only among the remaining relationships. Number of hard constraints used depended upon the setting (multi-class/two-class) and will be detailed in the respective sections.
- **Hard-Prior**: This model is a combination of hard and prior models. The true relationships are selected using the prior probability of the relationships while ensuring hard constraints.
- **MLN:** This is the full blown model which uses a combination of soft and hard constraints, with weights learned for the soft rules using the training data. The prior information is incorporated into the model using singleton (soft) constraints which state that a particular kind of relationship holds. The number of rules used for multi-class/two-class cases will be detailed in the respective sections.

4.3. Methodology

We performed two sets of experiments: two-class and multi-class as detailed in the model section. For the two-class setting, no age or gender information was used in the constraints. In the multi-class experiment, we assume the gender and age of each person is known. In this case, everything except the *HasRelation* predicate is treated as evidence (at learning as well as inference time).

For all our experiments, the age attribute was divided into five bands: child ($<$13), teen (13-20), young-adult (20-40), adult (40-60) and senior adult ($>$60). Since we had labeled data for only relationships between the collectors and the people appearing in their collections, we restricted ourselves to working with the relationships of various people with the corresponding photo collectors (instead of predicting all possible pairs of relationships). We assumed that we know the identity of the photo collector during learning as well as inference. The rules described in our model can easily (and in obvious manner) be adapted to this setting. Note that we fixed the above formulation to evaluate our model(s) in a fair manner.

We performed leave-one-out cross validation, learning on 12 photo collectors' data and predicting on the remaining one. Each experiment predicts the relationship between the collection owner and each person in the image collection, for a total of 286 predictions across the 13 collections. For each model on each test set, we measure the accuracy (ACR), conditional log-likelihood (CLL), and area under the precision-recall curve (AUC) for the query predicates. ACR is a basic and intuitive measure of how well the model is doing. CLL directly measures the quality of the probability estimates produced. AUC gives a measure of how well the model works overall in the precision recall space. AUC is especially useful when the distribution of positives and negatives is skewed (like in our case - out of n possible relationships, only one holds). The CLL of a query predicate is the average over all its groundings of the ground atom's log-probability given the evidence. The precision-recall curve for a predicate is computed by varying the threshold CLL above which a ground atom is predicted to be true. We computed the standard deviations of the AUCs using the method of Richardson and Domingos[14].

4.4. Results

4.4.1 Two-Class:

In the two-class setting, four soft constraints were used. No hard constraints were used. The soft constraints simply state the following regularities: acquaintances appear in few photos, friends occur in large groups, friends and relatives are clustered across photos. Note that none of these soft constraints uses age or gender information. The goal of this experiment is to show our approach leverages information from the soft constraints in a very simple setting. Table 1 details the results [1]. Clearly, the MLN model has the best performance followed by the prior model. Figure 1 shows the precision-recall curves that compare all the models. These results establish the benefit of our approach in a very basic setting. Next, we move on to the more involved,

[1]The models "hard" and "hard-prior" are absent as no hard constraints are used in this setting

Table 1. Results comparing the different models for the two-class setting.

Model	ACR	CLL	AUC
Random	48.8±3.1	-0.693±0.000	49.6±0.0
Prior	50.4±3.1	-0.693±0.001	54.4±0.3
MLN	64.7±3.0	-0.709±0.031	66.7±0.4

Table 2. Results comparing the different models for the multi-class setting.

Model	ACR	CLL	AUC
Random	12.4±2.1	-0.410±0.015	9.4±0.2
Prior	26.4±2.7	-0.404±0.015	21.0±0.2
Hard	14.3±2.2	-0.372±0.013	21.0±0.2
Hard-Prior	26.4±2.7	-0.370±0.013	23.3±0.2
MLN	29.9±2.9	-0.352±0.013	29.7±1.0

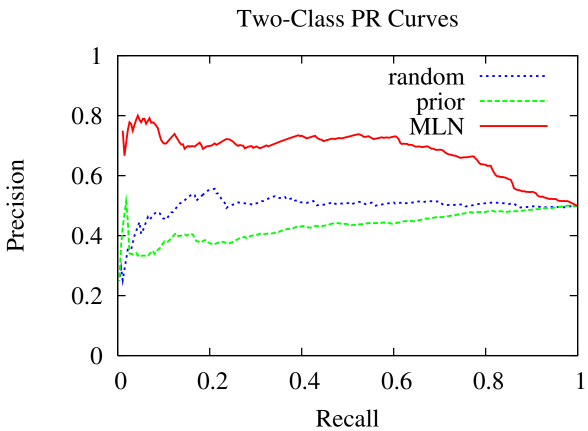

Figure 1. Precision-recall curves for the two-class setting.

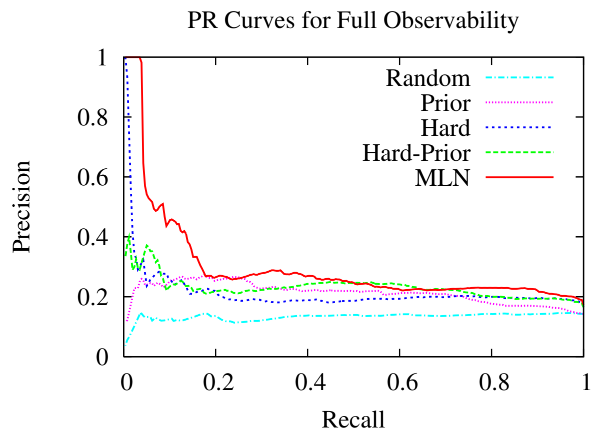

Figure 2. Precision-recall curves for the multi-class setting.

multi-class setting.

4.4.2 Multi-Class:

For the case of full observability, we used 5 hard constraints and 14 soft constraints (see the model section for examples).

Table 2 details the results. As expected, there is a gradual increase in performance as we move from the most basic model (random) to the most sophisticated model (MLN). Figure 2 shows the precision-recall curves for multi-class setting. The MLN model dominates over other models at all points of precision-recall space, clearly demonstrating the benefit of our soft rules.

It is interesting to examine the weights that are learned for the soft rules. As seen in (1), more importance is placed on rules with higher weights. The following three rules received the highest weights (in descreasing order) in our Markov Logic Network:

- For an image with one child and one non-child, then the child shares the "child" relationship with the collection owner.
- Same rule as above, except the image has exactly two people.
- If a person is friends with a child, then the child and the person have the same age.

Our Morkov Logic Network actually learns these intuitive rules from the training data.

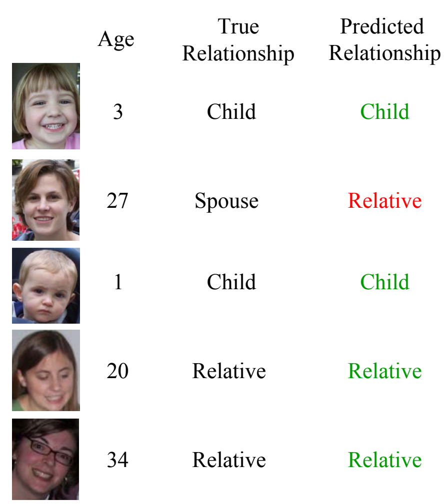

Figure 3. For four of five people from this collection, the relationship predictions are correct using the Markov Logic Network.

Figure 3 shows the predicted social relationships from the Markov Logic Network for five people from an image collection.

5. Future Work

There are a number of directions for future work. Perhaps the most obvious is to apply our model to a more diverse and larger set of photo collections. We would also like to integrate face recognition as part of our experiment. The input to the system will be various features relevant for face recognition and the system will perform a collective inference over recognizing faces, predicting relationships, and other personal characteristics such as age and gender. It would also be interesting to leverage the information available in social relationships to better solve other problems being tackled by the vision community such as detecting the setting in which a set of photos was taken (indoor/outdoor, home/office), various activities being performed by the users, and so on.

6. Conclusion

Automatically predicting social relationships in consumer photo collections is an important problem for the semantic understanding of consumer images and has a wide range of applications. Humans typically use ”rules of thumb” to detect these social relationships. Our contribution in this paper is to use the power of Markov logic to formulate these rules of thumb as soft constraints in a mathematically sound and meaningful way. A combination of hard and soft constraints is used. To the best of our knowledge, out approach is unique in the literature. Our approach performs better than a combination of various basic approaches and suggests further investigation into the problem, leading to a much better understanding of social relationships that can be extracted from consumer photos.

References

[1] J. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24:179–195, 1975.

[2] P. Domingos, S. Kok, D. Lowd, H. Poon, M. Richardson, and P. Singla. Markov logic. In K. K. L. De Raedt, P. Frasconi and S. Muggleton, editors, *Probabilisitc Inductive Logic Programming*, page In press. Springer, New York, NY, 2007.

[3] A. Gallagher and T. Chen. Using group prior to identify people in consumer images. In *Proc. CVPR SLAM workshop*, 2007.

[4] L. Getoor and B. Taskar. *Introduction to Statistical Relational Learning*. MIT Press, 2007.

[5] S. Kok and P. Domingos. Learning the structure of Markov logic networks. In *Proc. ICML-05*, pages 441–448, Bonn, Germany, 2005. ACM Press.

[6] S. Kok, P. Singla, M. Richardson, and P. Domingos. The Alchemy system for statistical relational AI. Technical report, Univ. of WA, Seattle, WA, 2005. http://www.cs.washington.edu/ai/alchemy.

[7] A. Lanitis, C. Taylor, and T. Cootes. Toward automatic simulation of aging effects on face images. *PAMI*, 2002.

[8] L.-J. Li and L. Fei-Fei. What, where and who? classifying event by scene and object recognition. In *Proc. ICCV*, 2007.

[9] J.-H. Lim, Q. Tian, and P. Mulhem. Home photo content modeling for personalized event-based retrieval. *IEEE Multimedia*, 2003.

[10] D. Lowd and P. Domingos. Efficient weight learning for Markov logic networks. In *Proc. PKDD-07*, pages 200–211, Warsaw, Poland, 2007. Springer.

[11] J. Luo, M. Boutell, and C. Brown. Pictures are not taken in a vacuum - an overview of exploiting context for semantic scene content understanding. *IEEE Signal Processing Magazine*, 2006.

[12] M. Naaman, R. Yeh, H. Garcia-Molina, and A. Paepcke. Leveraging context to resolve identity in photo albums. In *Proc. JCDL*, 2005.

[13] H. Poon and P. Domingos. Sound and efficient with probabilistic and determinisitc dependencies. In *Proc. AAAI-06*, pages 458–463, Boston, MA, 2006. AAAI Press.

[14] M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62:107–136, 2006.

[15] P. Singla and P. Domingos. Discriminative training of Markov logic networks. In *Proc. AAAI-05*, pages 868–873, Pittsburgh, PA, 2005. AAAI Press.

[16] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 485–492, Edmonton, Canada, 2002. Morgan Kaufmann.

[17] C.-Y. Weng, W.-T. Chu, and J.-L. Wu. Rolenet: treat a movie as a small society. In *ACM Multimedia MIR Workshop*, 2007.

[18] M.-H. Yang and B. Moghaddam. Support vector machines for visual gender classification. *Proc. ICPR*, 2000.