How to Make Face Recognition Work: The Power of Modeling Context

Ashish Kapoor[1], **Dahua Lin**[2], **Simon Baker**[1], **Gang Hua**[3] **and Amir Akbarzadeh**[1]
[1]Microsoft Research
[2]MIT CSAIL
[3]Stevens Institute of Technology

Abstract

Face recognition in the wild has been one of the longest standing computer vision challenges. While there has been constant improvement over the years, the variations in appearance, illumination, pose etc. still makes it one the hardest task to do well. In this paper we summarize two techniques that leverage context and show significant improvement over vision only methods. At the heart of the approach is a probabilistic model of context that captures dependencies induced via set of contextual relations. The model allows application of standard variational inference procedures to infer labels that are consistent with contextual constraints.

With the ever increasing popularity of digital photos, vision-assisted tagging of personal photo albums has become an active research topic. Existing efforts in this area have mostly been devoted to using face recognition to help tag people. However, current face recognition algorithms are still not very robust to the variation of face appearance in real photos.

This paper explores how contextual cues can aid and improve face recognition in the wild. In particular instead of just focusing on answering *who*, we incorporate methods to consider *what*, *when*, and *where*, and in doing so show how such contextual can immensely help with the task. We consider the domains of people, events, and locations, as a whole and the key insight that the domains are not independent and knowledge in one domain can help the others. For example, if we know the event that a photo was captured in, we can probably infer who was in the photo, or at least reduce the set of possibilities. On the other hand, the identities of the people in a photo may help us infer when and where the photo was taken.

Our approach is based on the principles of probabilistic graphical models where we build a network of classifiers over different domains. Ideally, if a strong classifier is available to recognize the instances in one domain accurately, the labels in this domain can be used to help recognition in the other domains. However, a challenge that arises in real systems is that we often do not have a strong initial classifier in any of the domains. One of the primary contributions of the work presented here is to develop a unified framework that couples the recognition across the domains. A joint learning and inference algorithm allows accurate recognition in all domains by exploiting the statistical dependency between them and reinforces individual classifiers, which alone may be relatively weak.

Related Work

Over the last decade, there has been a great deal of interest in the use of context to help improve face recognition accuracy in personal photos. A recent survey of context-aided face recognition can be found in (Gallagher and Tsuhan 2008). Zhang et al. (Zhang et al. 2003) utilized body and clothing in addition to face for people recognition. Davis et al. (Davis et al. 2005; 2006) developed a context-aware face recognition system that exploits GPS-tags, time-stamps, and other metadata. Song and Leung (Song and Leung 2006) proposed an adaptive scheme to combine face and clothing features based on the time-stamps. These methods treat various forms of contextual cues as linearly additive features, and thus oversimplify the interaction between different domains.

Various methods based on co-occurrence have also been proposed. Naaman et al. (Naaman et al. 2005) leveraged time-stamps and GPS-tags to reduce the candidate list based on people co-occurrence and temporal/spatial re-occurrence. Gallagher and Chen (Gallagher and Tsuhan 2007) proposed an MRF to encode both face similarity and exclusivity. In later work by the same authors (Gallagher and Chen 2007), a group prior is added to capture the tendency that certain groups of people are more likely to appear in the same photo. In addition, Anguelov et al. (Anguelov et al. 2007) developed an MRF model to integrate face similarity, clothing similarity and exclusivity.

There is also a lot of research on use of context in object recognition and scene classification. For example, Torralba et al. (Torralba et al. 2003; Torralba 2003) used scene context as a prior for object detection and recognition. Rabinovich et al. (Rabinovich et al. 2007) proposed a CRF model that utilizes object co-occurrence to help object categorization. Galleguillos et al. (Galleguillos, Rabinovich, and Belongie 2008) extended this framework to use both object co-occurrence and spatial configurations for image segmentation and annotation. Li-Jia and Fei-Fei (Li and Fei-Fei 2007) proposed a generative model that can be used to label scenes and objects by exploiting their statistical depen-

dency. In later work (Li, Socher, and Fei-Fei 2009), the same authors extended this model to incorporate object segmentation. Cao et al. (Cao et al. 2008) employed a CRF model to label events and scenes coherently.

Modeling Context

In this paper we summarize two approaches proposed earlier to model context in face recognition task. First is a model that extends vanilla face recognition setting to incorporate simple match and non-match constraints (Kapoor et al. 2009) that are derived from physical context of individual photographs (e.g. two faces in the same image cannot have the same identity). The second model (Lin et al. 2010) is a more sophisticated version of the first, where the location, time and co-occurrence patterns of individuals in a photograph is considered as a context to influence the classification. The core idea behind both of the models is a power probabilistic framework that enables us to incorporate context via a network of classifiers that depend upon each other. Also note that both the models follow discriminative modeling paradigm as neither of them attempt to model $P(\mathbf{X})$, the high dimensional underlying density of observations. Finally, the inference methodology for both the models follows the principles of variational approximation where the classification and constraint resolution is performed iteratively.

Model 1: Match and Non-match Constraints

This is a fairly simple model (Kapoor et al. 2009) that extends classic supervised classification setting to allow incorporation of additional sources of prior information. In this model, we consider two simple contextual constraints: match and non-match. First, if two faces appeared in the same unedited photo, the two faces cannot have the same identity. We call such constraints nonmatch constraints. Another example is in video. If faces are tracked and two images are from the same track, they must have the same identity. We call such constraints match constraints.

Model: Assume we are given a set of face images $\mathbf{X} = \{\mathbf{x}_i\}$. We partition this set into a set of labeled ones $\mathbf{X}_L$ with labels $\mathbf{t}_L = \{t_i | i \in L\}$ and a set of unlabeled ones $\mathbf{X}_U$. The model consists of a network of predictions that interact with one another such that the decision of each predictor is influenced by the decision of its neighbors. Specifically, given match and non-match constraints we induce a graph where every vertex corresponds to a label t_i, $i \in L \cup U$, and is connected to its neighbors according to the given constraint. We will denote the set of edges corresponding to match and non-match edges as $\mathcal{E}^+$ and $\mathcal{E}^-$ respectively.

Figure 1 illustrates the factor graph corresponding to the proposed model. The class labels $\{t_i : i \in L \cup U\}$ are denoted by squares and influence each other based on different match (green lines) and non-match (dashed red lines) constraints. In addition to these constraints, our model also imposes smoothness constraints using a GP prior (Rasmusen and Williams 2006). We introduce latent variables $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^n$ that use a GP prior to enforce the assumption that *similar* points should have similar prediction. In particular,

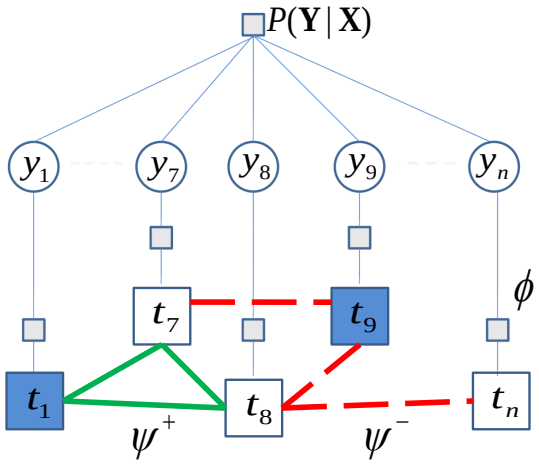


Figure 1: Factor graph depicting the first proposed discriminative model. The shaded nodes correspond to the observed labels (training data) and the thick green and dashed red line correspond to match and non-match constraints respectively.

the latent variables are assumed to be jointly Gaussian and the covariance between two outputs y_i and y_j is typically specified using a kernel function applied to $\mathbf{x}_i$ and $\mathbf{x}_j$. Formally, $p(\mathbf{Y}|\mathbf{X}) \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$ where $\mathbf{K}$ is a kernel matrix[1] with $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and encodes similarity between two different face regions.

Given a pool of images, the model induces a conditional probability distribution $p(\mathbf{t}, \mathbf{Y}|\mathbf{X})$ using the GP prior $p(\mathbf{Y}|\mathbf{X})$ and potential functions ϕ, ψ^+ and ψ^-. Here ϕ encodes the compatibility of a label t and the corresponding latent variable y. Further, ψ^+ and ψ^- encode the pairwise label compatibility according to the match and non-match constraints respectively. Thus, the conditional distribution induced by the model can be written as:

$$p(\mathbf{t}, \mathbf{Y}|\mathbf{X}) = \frac{1}{Z} p(\mathbf{Y}|\mathbf{X}) \prod_{i=1}^{n} \phi(\mathbf{y}_i, t_i) \times \prod_{(i,j)\in\mathcal{E}^+} \psi^+(t_i, t_j) \prod_{(i,j)\in\mathcal{E}^-} \psi^-(t_i, t_j)$$

where Z is the partition function (normalization term) and the potentials ϕ, ψ^+ and ψ^- take the following form:

$$\phi(\mathbf{y}_i, t_i) \propto e^{-\frac{||\mathbf{y}_i - \bar{\mathbf{t}}_i||^2}{2\sigma^2}}$$
$$\psi^+(t_i, t_j) = \delta(t_i, t_j)$$
$$\psi^-(t_i, t_j) = 1 - \delta(t_i, t_j).$$

Here, $\delta(\cdot, \cdot)$ is the Dirac delta function and evaluates to 1 whenever the arguments are equal, and zero otherwise. Also,

[1] This kernel matrix is a positive semidefinite matrix and is akin to the kernel matrix used in classifiers such as SVMs.

$\bar{t}_i$ is the indicator vector corresponding to t_i and σ^2 is the noise parameter and determines how tight the relation between the smoothness constraint and the final label is. By changing the value of σ we can emphasize or de-emphasize the effect of the GP prior. Note that in absence of any match and non-match constraints the model reduces to a multi-class classification scenario with GP models (Kapoor et al. 2007; Rasmusen and Williams 2006).

Inference: Given some labeled data the key task is to infer $p(\mathbf{t}_U|\mathbf{X}, \mathbf{t}_L)$ the posterior distribution over unobserved labels $\mathbf{t}_U = \{t_i|i \in U\}$. The inference is performed by simple message passing to resolve the smoothness, match and non-match constraints and infer the unobserved variables in an efficient manner. Intuitively, the inference is done by alternating between a classification step and a constraint resolution scheme. A simple application of the face recognition classifier on the unlabeled set might lead to labels that are inconsistent with contextual constraints. However, these inconsistencies are resolved by performing belief propagation. This resolution of inconsistencies induces novel beliefs about the unlabeled points and are propagated to the classifier in the next iteration. By iterating between these two updates the model consolidates information from both components, and thus provides a good approximation of the true posterior. This scheme is an approximate inference and derived by maximizing the variational lower bound (see Kapoor et al. (Kapoor et al. 2009) for the details).

Model 2: Location, Co-occurrence and Time

The second model (Lin et al. 2010) is a richer model and specifically consider four kinds of contextual relations: (a) the *people-event relation* models who attended which events, (b) the *people-people relation* models which pairs of people tend to appear in the same photo, (c) the *event-location relation* models which event happened where, and (d) the *people-location relation* models who appeared where. These relations embody a wide range of contextual information, which is modeled uniformly under the same mathematical framework. It is important to note that each pair of related domains are symmetric with respect to the corresponding relation. This means, for example, that utilizing the people-event relation, event recognition can help people recognition, and people recognition can also help event recognition.

Model: The framework, outlined in figure 2, consists of three domains: people, events, and locations. Each domain contains a set of instances. In order to account for the uncertainty due to missing data or ambiguous features, we consider the labels in all three domains as random variables to be inferred. Pairs of domains are connected to each other through a set of cross-domain relations that model the statistical dependency between them.

Suppose there are M domains: $\mathcal{Y}_1, \ldots, \mathcal{Y}_M$. Each domain is modeled as a set of instances, where the i-th instance in $\mathcal{Y}_u$ is associated with a label of interest, denoted as the random variable y_u^i. While the user can provide a small number of labels in advance, most labels are unknown and need to

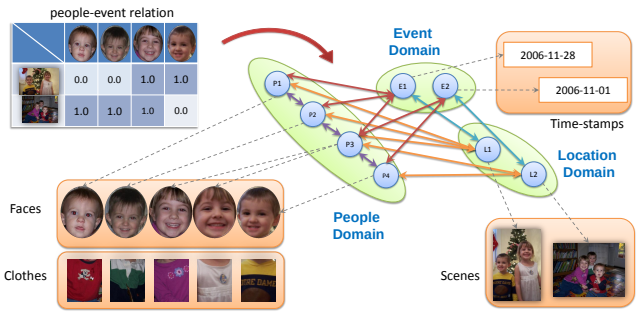


Figure 2: Our framework comprises three types of entity: (1) The people, event, and location **domains**, together with their instances. (2) The **observed features** of each instance in each domain. (3) A set of contextual **relations** between the domains. Each relation is a 2D table of coefficients that indicate how likely a pair of labels is to co-occur. Although only the people-event relation is shown in this figure, we consider four different relations in this paper. See body of text for more details.

be inferred. Specifically, we consider three domains for people, events, and locations. Each detected face corresponds to a person instance in the people domain, and each photo corresponds to both an event instance and a location instance. Each domain is associated with a set of features to describe its instances. In particular, person instances are characterized by their facial appearance and clothing; while events and locations are characterized by time-stamps and the background color distribution respectively.

To exploit the statistical dependency between the labels in different domains, we introduce a relational model R_{uv} between each pair of related domains $\mathcal{Y}_u$ and $\mathcal{Y}_v$. Each R_{uv} is parameterized by a 2D table of coefficients that indicate how likely a pair of labels is to co-occur. Taking advantage of these relations, we can use the information in one domain to help infer the labels in others.

Here, we use Y and $\mathbf{X}$ to represent the labels and features of all domains. The formulation has two parts: (1) $p(Y|R, \mathbf{X})$: the joint likelihood of the labels given the relational models and features and (2) $p(R)$: the prior put on the relations to regularize their estimation.

The joint likelihood of the data labels is directly modeled as a conditional distribution based on the observed features:

$$p(Y|\mathbf{X}; R) = \frac{1}{Z} \times$$
$$\exp\left(\sum_{u=1}^{M} \alpha_u \Phi_u(Y_u; \mathbf{X}_u) + \sum_{(u,v)\in\mathcal{R}} \alpha_{uv} \Phi_{uv}(Y_u, Y_v; R_{uv})\right).$$

This likelihood contains: (1) an *affinity potential* $\Phi_u(Y_u, \mathbf{X}_u)$ for each domain $\mathcal{Y}_u$ to model feature similarity, and (2) a *relation potential* $\Phi_{uv}(Y_u, Y_v; R_{uv})$ for each pair of related domains $(u, v) \in \mathcal{R}$. The terms are combined with weights α_u and α_{uv}.

The affinity potential term Φ_u captures the intuition that two instances in $\mathcal{Y}_u$ with similar features are likely to be in

the same class:

$$\Phi_u(Y_u; \mathbf{X}_u) = \sum_{i=1}^{N_u} \sum_{j=1}^{N_u} w_u(i,j) \mathbb{I}(y_u^i = y_u^j).$$

Here, $w_u(i,j)$ is the similarity between the features of the instances corresponding to y_u^i and y_u^j. $\mathbb{I}(\cdot)$ denotes the indicator that equals 1 when the condition inside the parenthesis holds. The similarity function w_u depends on the features used for that domain. Intuitively, Φ_u considers all instances of $\mathcal{Y}_u$ over the entire collection, and attains large value when instances with similar features are assigned the same labels. Maximizing Φ_u should therefore result in clusters of instances that are consistent with the feature affinity.

The relational potential term $\Phi_{uv}(Y_u, Y_v; R_{uv})$ models the cross-domain interaction between the domains $\mathcal{Y}_u$ and $\mathcal{Y}_v$. The relational model R_{uv} is parameterized as a 2D table of co-occurring coefficients between pairs of labels. For example, for the people domain $\mathcal{Y}_u$ and the event domain $\mathcal{Y}_v$, $R_{uv}(k,l)$ indicates how likely it is that person k attended event l. We define Φ_{uv} to be:

$$\Phi_{uv}(Y_u, Y_v; R_{uv}) = \sum_{i \sim j} \sum_{k,l} R_{uv}(k,l) \mathbb{I}(y_u^i = k) \mathbb{I}(y_v^j = l).$$

Here, $i \sim j$ means that y_u^i and y_v^j co-occur in the same photo. Intuitively, a large value of $R_{uv}(k,l)$ indicates that the pair of labels k and l co-occur often, and will encourage y_u^i to be assigned k and y_v^j be assigned l. Hence, maximizing Φ_u should lead to the labels that are consistent with the relation.

Formally, our goal is to jointly estimate the posterior probability of the labels Y and relations R conditioned on the feature measurements $\mathbf{X}$:

$$p(Y, R|\mathbf{X}) \propto p(Y|R, \mathbf{X}) p(R).$$

Finally, the prior follows standard principles of regularization and sparsity in order to avoid over-fitting:

$$p(R) \propto \exp\left(-\beta_1 \sum_{(u,v) \in \mathcal{R}} ||R_{uv}||_1 - \beta_2 \sum_{(u,v) \in \mathcal{R}} ||R_{uv}||_2^2 \right).$$

Here, $||R_{uv}||_1$ and $||R_{uv}||_2$ are the L1 and L2 norms of the relational matrix. The first term encourages sparsity of the relational coefficients, and therefore can effectively suppress the coefficients due to occasional co-occurrences, retaining only those capturing truly stable relations. The second term inhibits really large coefficients that might otherwise occur when the class sizes are imbalanced.

Inference: The inference is performed via a variational EM algorithm where the goal is to jointly infer the labels of the instances and estimate the relational models. With a few labels in different domains provided in advance by a user, the algorithm iterates between two steps: (1) Infer the distribution of the unknown labels based on both the extracted features and the current relational models R. (2) Estimate and update the relational models R using the labels provided by the user and the hidden labels inferred in the previous iteration. As before the the iterations can be thought of as a message passing scheme between individual classifiers and resolution of labels in order to resolve the constraints imposed by the contextual cues. We refer readers to Lin et al. (Lin et al. 2010) for technical details.

Experiments

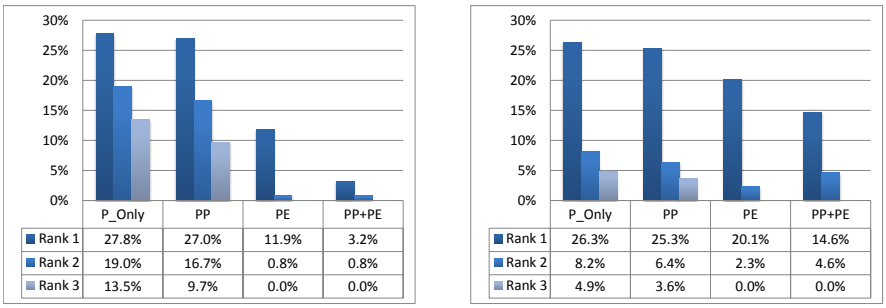
We present experimental results for the Model 2 as its richer and more powerful. The experiments are carried out on two publicly available data sets that are commonly used to evaluate research in personal photo tagging, which we call *E-Album* (Cui et al. 2007) and *G-Album* (Gallagher 2008). We refer readers to Lin et al. (Lin et al. 2010) for details on feature extraction and implementation. We compare the performance of four different variants of our algorithm: (1) using only people affinity (no contextual information), (2) with the people-people relation, (3) with the people-event relation, and (4) with both relations.

The results are shown in Figure 3. We note three observations: First, on both albums the people-people relation alone provides only a limited improvement (rank-1 errors reduced from 27.8% to 27.0% for the E-Album). Second, the people-event relation gives a much bigger improvement (rank-1 errors reduced from 27.8% to 11.9% for the E-Album). Third, the combination of the people-event relation and the people-people relation yields another significant improvement (rank-1 errors down to 3.2% on the E-Album). To illustrate our results visually, we include a collage of all of the errors for the E-Album in Figure 4. With both the people-event and people-people relations used, there are only three errors (3.2%) on the E-Album (see Figure 4).

These results show: (1) that the people-event and people-people relations provide complementary sources of information, and (2) the people-event relation makes the people-people relation more effective than without it. The most likely explanation is that the group-prior and exclusivity are more powerful when used on the small candidate list provided by the people-event relation.

Overall, we found the G-Album to be more challenging. Partly, this is due to the fact that the G-Album contains a very large number of events (117), each with very few photos (3.8 on average.) The people-event relation would be more powerful with more photos per event. Note, however, that our framework still yields a substantial improvement, reducing the rank-1 error rate from 26.3% to 14.6%. Note also, that the rank-3 error rate is reduced to zero on both albums, a desirable property in vision-assisted tagging system where a short-list of candidates is often provided for the user to choose from.

To validate the statistical significance of our results, we randomly generated multiple pre-labeled sets, with the percentage of pre-labeled instances varying from 15% to 55%. Figure 5 contains the median rank-1 results (signified by the central mark) along with the 25th and 75th percentiles (signified by lower and upper bars) obtained on E-Album. We found that the improvement is significant across the entire range of pre-labeling percentage in both data sets.



(a) Error rates on E-Album

(b) Error rates on G-Album

Figure 3: People labeling results for several different configurations of our algorithm.

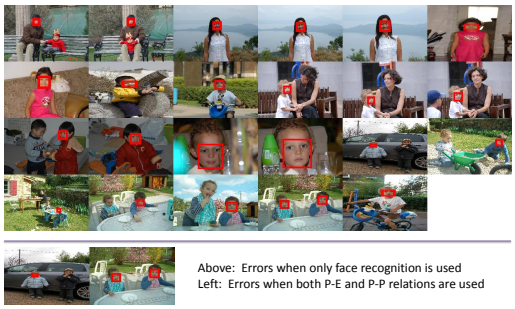


Figure 4: All rank-1 errors for the E-Album. Above the delimiter: Errors with no contextual relations (27.8%). Below the delimiter: Errors with both the people-event and people-people relations (3.2%).

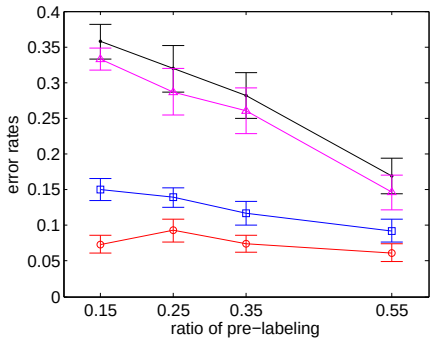


Figure 5: Statistical significance results on the E-Album with different percentages of pre-labeled instances. Curves from top to bottom: using only face, people-people only, people-event only, and using both.

Conclusion

The paper explores how context can be leveraged to boost face recognition in personal photo collections. We have proposed probabilistic model that incorporate cross-domain relations as a mechanism to model context in multi-domain labeling (people, events, locations) and match/non-match constraints. Relation estimation and label inference are unified in a single optimization algorithm. Our experimental results show that probabilistic graphical models provide a elegant, powerful, and general method of modeling context in vision-assisted tagging applications.

References

Anguelov, D.; Lee, K.-c.; Gokturk, S. B.; and Sumengen, B. 2007. Contextual identity recognition in personal photo albums. In *CVPR'07*.

Cao, L.; Luo, J.; Kautz, H.; and Huang, T. S. 2008. Annotating collections of photos using hierarchical event and scene models. In *CVPR'08*.

Cui, J.; Wen, F.; Xiao, R.; Tian, Y.; and Tang, X. 2007. Easyalbum: an interactive photo annotation system based on face clustering and re-ranking. In *SIGCHI*, 367–376.

Davis, M.; Smith, M.; Canny, J.; Good, N.; King, S.; and Janakiraman, R. 2005. Towards context-aware face recognition. In *13th ACM Conf. on Multimedia*.

Davis, M.; Smith, M.; Stentiford, F.; Bamidele, A.; Canny, J.; Good, N.; King, S.; and Janakiraman, R. 2006. Using context and similarity for face and location identification. In *SPIE'06*.

Gallagher, A. C., and Chen, T. 2007. Using group prior to identify people in consumer images. In *CVPR Workshop on SLAM'07*.

Gallagher, A. C., and Tsuhan, C. 2007. Using a markov network to recognize people in consumer images. In *ICIP*.

Gallagher, A. C., and Tsuhan, C. 2008. Using context to recognize people in consumer images. *IPSJ Journal* 49:1234–1245.

Gallagher, A. C. 2008. Clothing cosegmentation for recognizing people. In *CVPR'08*.

Galleguillos, C.; Rabinovich, A.; and Belongie, S. 2008. Object categorization using co-occurrence, location and appearance. In *CVPR'08*.

Kapoor, A.; Grauman, K.; Urtasun, R.; and Darrell, T. 2007. Active learning with Gaussian Processes for object categorization. In *ICCV*.

Kapoor, A.; Hua, G.; Akbarzadeh, A.; and Baker, S. 2009. Which faces to tag: Adding prior constraints into active learning. In *ICCV'09*.

Li, L.-J., and Fei-Fei, L. 2007. What, where and who? classifying events by scene and object recognition. In *CVPR'07*.

Li, L.-J.; Socher, R.; and Fei-Fei, L. 2009. Towards total scene understanding: Classification, annotation, and segmentation in an automatic framework. In *CVPR'09*.

Lin, D.; Kapoor, A.; Hua, G.; and Baker, S. 2010. Joint people, event, and location recognition in personal photo collections using cross-domain context. In *ECCV*.

Naaman, M.; Garcia Molina, H.; Paepcke, A.; and Yeh, R. B. 2005. Leveraging context to resolve identity in photo albums. In *ACM/IEEE-CS Joint Conf. on Digi. Lib.*

Rabinovich, A.; Vedaldi, A.; Galleguillos, C.; Wiewiora, E.; and Belongie, S. 2007. Objects in context. In *ICCV'07*.

Rasmusen, C. E., and Williams, C. 2006. *Gaussian Processes for Machine Learning*. MIT Press.

Song, Y., and Leung, T. 2006. Context-aided human recognition - clustering. In *ECCV'06*.

Torralba, A.; Murphy, K. P.; Freeman, W. T.; and Rubin, M. A. 2003. Context-based vision system for place and object recognition. In *ICCV'03*.

Torralba, A. 2003. Contextual priming for object detection. *Int'l. J. on Computer Vision* 53(2):169–191.

Zhang, L.; Chen, L.; Li, M.; and Zhang, H. 2003. Automated annotation of human faces in family albums. In *11th ACM Conf. on Multimedia*.