Joint People, Event, and Location Recognition in Personal Photo Collections using Cross-Domain Context *

Dahua Lin[1,2], Ashish Kapoor[2], Gang Hua[3], and Simon Baker[2]

[1] Computer Science and Artificial Intelligence Laboratory, MIT
[2] Microsoft Research
[3] Nokia Research Center Hollywood

Abstract. We present a framework for vision-assisted tagging of personal photo collections using context. Whereas previous efforts mainly focus on tagging people, we develop a unified approach to jointly tag across multiple domains (specifically people, events, and locations). The heart of our approach is a generic probabilistic model of context that couples the domains through a set of cross-domain relations. Each relation models how likely the instances in two domains are to co-occur. Based on this model, we derive an algorithm that simultaneously estimates the cross-domain relations and infers the unknown tags in a semi-supervised manner. We conducted experiments on two well-known datasets and obtained significant performance improvements in both people and location recognition. We also demonstrated the ability to infer event labels with missing timestamps (i.e. with no event features).

1 Introduction

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. To address this issue, various methods [1] have been proposed to exploit contextual cues to aid recognition. While obtaining some improvement, these methods focus on the people domain, and neglect other important domains such as events and locations.

The most important questions in regard to personal photo tagging are *who*, *what*, *when*, and *where*. With an aim of answering these questions coherently, we consider the domains of people, events, and locations, as a whole. Our work is motivated by the 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

* The research described in this paper was conducted when all four authors were affiliated with Microsoft Research Redmond.

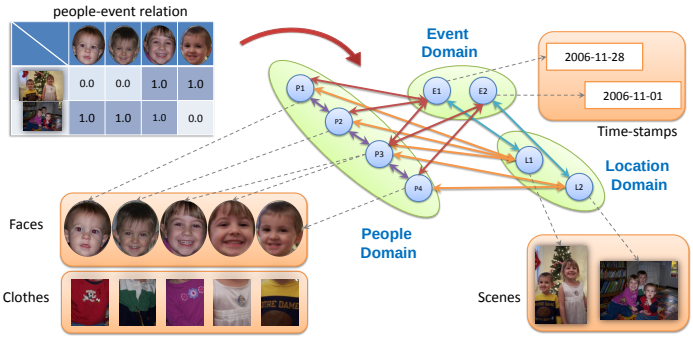


Fig. 1. Our framework comprises three types of entity: (1) The people, event, and location **domains**, together with their instances. (2) The **observed features** of each instance in each domain. (3) A set of contextual **relations** between the domains. Each relation is a 2D table of coefficients 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.

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.

Ideally, if a strong classifier is available to recognize the instances in a domain accurately, one can utilize the labels in this domain to help the recognition in others. However, a challenge arises in real system is that we often do not have a strong classifier to start with in any domain. One of our primary contributions is to develop a unified framework that couples the recognition in these domains. We also derive a joint learning and inference algorithm that would allow us to achieve accurate recognition in all domains by exploiting the statistical dependency between them to reinforce individual classifiers.

Our framework, outlined in figure 1, 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.

In this paper, we specifically consider four 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.

Based on this framework, we formulate a joint probabilistic model to integrate both feature similarity and contextual relations. However, we face a challenge that specially arises in the application of personal photo tagging. Unlike other

classification problem such as object recognition where one can learn the contextual models from training data, the relational models (e.g. people-event relation) estimated from one photo collection are generally not applicable to other collections. In fact, the set of people or events may well be completely different in two different photo collections. It is also infeasible to require a user to prepare training data for each of their albums. Instead, we develop an algorithm that simultaneously estimates the relations and infers the labels across all domains by solving a unified optimization problem in a semi-supervised way.

We tested our approach on two well-known datasets. For people labeling, the error rate is reduced from 27.8% to 3.2% on one data set, and from 26.3% to 14.6% on the other. We also obtained a huge improvement in location labeling (16.7% to 1.3%). Finally, we demonstrate the ability to estimate event labels for photos in the presence of missing timestamps (i.e. with missing event features.)

2 Related Work

Related prior work can be roughly split into two categories: context-aided face recognition, and object/scene classification using context. We now review this related work and clarify the key differences from our approach.

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 [1]. Zhang et al. [2] utilized body and clothing in addition to face for people recognition. Davis et al. [3, 4] developed a context-aware face recognition system that exploits GPS-tags, time-stamps, and other meta-data. Song and Leung [5] 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 oversimplifies the interaction between different domains.

Various methods based on co-occurrence have also been proposed. Naaman et al. [6] leveraged time-stamps and GPS-tags to reduce the candidate list based on people co-occurrence and temporal/spatial re-occurrence. Gallagher and Chen [7] proposed an MRF to encode both face similarity and exclusivity. In later work by the same authors [8], 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. [9] developed an MRF model to integrate face similarity, clothing similarity and exclusivity. Finally, Kapoor et al. [10] proposed a framework that uses Gaussian Processes to capture contextual constraints. Whereas these models provide a more flexible way to capture the interaction between co-occuring instances, they are nearly all formulated within the people domain. An exception is Naaman et al. [6], which uses time and locations, however the model is heuristic and the time and location labels are treated as noiseless.

In contrast to prior contextual face recognition work, our framework treats all three domains in a uniform manner. The labels in each domain (including events and locations) are modeled as random variables, rather than noiseless quantities, and the relation connecting each pair of domains can be utilized for the

inference in both domains. Moreover, instead of using heuristics to utilize time and locations, we develop a principled approach that establishes a joint probabilistic model over these domains. Labeling and estimation are thus performed as a unified optimization process.

Our framework is also related to the use of context in object recognition and scene classification. For example, Torralba et al. [11, 12] used scene context as a prior for object detection and recognition. Rabinovich et al. [13] proposed a CRF model that utilizes object co-occurrence to help object categorization. Galleguillos et al. [14] extended this framework to use both object co-occurrence and spatial configurations for image segmentation and annotation. Li-Jia and Fei-Fei [15] proposed a generative model that can be used to label scenes and objects by exploiting their statistical dependency. In later work [16], the same authors extended this model to incorporate object segmentation. Cao et al. [17] employed a CRF model to label events and scenes coherently.

While these approaches share some technical similarity with our work, three key differences distinguish our work:

(1) As mentioned above, it is infeasible in personal photo tagging to provide a separate training set to estimate the contextual model. To meet this challenge, we designed an algorithm where the model is estimated directly from the photo collection to be tagged, along with inference being performed. This should be contrasted with the conventional approach to object/scene classification, where the models are learned offline on a training set.

(2) The instances to be labeled in object/scene recognition are typically instances (e.g. objects) within a *single image*. The context models the relations (spatial, co-occurrence) within that image. On the other hand, our contextual model is over the *entire photo collection*. It models inter-photo dependencies rather than just intra-image relations. This makes it possible to reliably estimate the relational models without the need of a priori training.

(3) The application domain is different. Rather than considering generic object recognition and scene classification, we consider the problem of context-assisted face, location, and event recognition in personal photo collections.

3 Probabilistic Model Formulation

In this section, we formalize our framework as a Bayesian model. 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, modeled as a random variable y_u^i. While the user can provide a small number of labels in advance, most labels are unknown and to be inferred. Specifically, we consider three domains for people, events, and locations. Each detected face corresponds to a person instance in 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 respectively characterized by time-stamps and the background color distribution.

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 $\mathcal{Y}_v$. It 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.

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

$$p(Y, R|\mathbf{X}) \propto p(Y|R, \mathbf{X})p(R). \tag{1}$$

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 (section 3.1). (2) $p(R)$: the prior put on the relations to regularize their estimation (section 3.2).

3.1 Joint Probability of Labels

We propose to directly model the joint label distribution conditioned on the observed features, rather than assuming a parametric feature distribution for each class as in generative models. This approach is generally more effective when the number of labeled samples in each class is limited. In particular, we propose the following model for $p(Y|R, \mathbf{X})$:

$$p(Y|\mathbf{X}; R) = \frac{1}{Z} \exp\left(\sum_{u=1}^{M} \alpha_u \Phi_u(Y_u; \mathbf{X}_u) + \sum_{(u,v)\in\mathcal{R}} \alpha_{uv} \Phi_{uv}(Y_u, Y_v; R_{uv}) \right). \tag{2}$$

The proposed 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 $(u, v) \in \mathcal{R}$. They are combined with weights α_u and α_{uv}, which can be set by cross-validation in practice.

1. The **affinity potential** Φ_u captures the intuition that two instances in $\mathcal{Y}_u$ with similar features are likely to be in the same class:

$$\Phi_u(Y_u; \mathbf{X}_u) = \sum_{i=1}^{N_u} \sum_{j=1}^{N_u} w_u(i, j) \mathbb{I}(y_u^i = y_u^j). \tag{3}$$

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 (see section 5 for details). If the instances in a domain can be described by different types of features, we define affinity potentials for different features, and use their sum as the overall potential.

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 the feature affinity. This is in contrast to standard CRF models [18] that require learning class-specific feature coefficients for each class.

When clothing is used as one of the features in the people domain, a modification is necessary. As people may change clothes, comparing clothing features is only appropriate when the two person instances were in the same event. To model this, we modify the affinity potential for clothing features to be:

$$\Phi(Y_P; \mathbf{X}_C) = \sum_{i=1}^{N} \sum_{j=1}^{N} w_C(i,j) \mathbb{I}(y_p^i = y_p^j) \mathbb{I}(y_e^{ph(i)} = y_e^{ph(j)}). \tag{4}$$

Here, Y_P and $\mathbf{X}_C$ denote the people labels and clothing features, $w_c(i,j)$ is the similarity between the clothes of the i-th and j-th person instances, and $y_e^{ph(i)}$ and $y_e^{ph(j)}$ are the event labels of the corresponding photos. The factor $\mathbb{I}(y_e^{ph(i)} = y_e^{ph(j)})$ only turns on rest of the term within the same event.

2. The **relational potential** $\Phi_{uv}(Y_u, Y_v; R_{uv})$ 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 people domain $\mathcal{Y}_u$ and event domain $\mathcal{Y}_v$ $R_{uv}(k,l)$ indicates how likely it is that the person k attended the event l. Then, we define Φ_{uv} to be:

$$\Phi_{uv}(Y_u, Y_v; R_{uv}) = \sum_{i \sim j} \sum_{k,l} R_{uv}(k,l) \mathbb{I}(y_u^i = k) \mathbb{I}(y_v^j = l). \tag{5}$$

Here, $i \sim j$ means that y_u^i and y_v^j co-occur in the same photo. Intuitively, large value of $R_{uv}(k,l)$ indicate 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.

3.2 Relational Model Prior

In real application, only a relatively small number of instances are tagged in advance by user (often just one or two per class). The model is estimated from these user-given labels. While the estimation can also use the labels inferred in previous step in our iterative algorithm, the inferred labels could be noisy and actually depend on the user-given labels. To avoid over-fitting, it is important to regularize the relational models. To this end, we incorporate the following prior:

$$p(R) = \frac{1}{Z_{prior}} \exp\left(-\beta_1 \sum_{(u,v) \in \mathcal{R}} ||R_{uv}||_1 - \beta_2 \sum_{(u,v) \in \mathcal{R}} ||R_{uv}||_2^2 \right). \tag{6}$$

Here, $||R_{uv}||_1$ and $||R_{uv}||_2$ are L1 and L2 norm of the relational matrix. Intuitively, 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. Furthermore, it is often the case that a small number of people may appear hundreds of times, while others

only several times. This could result in exceptionally large coefficients for those dominant classes, and as a consequence, some instances in small classes may be incorrectly assigned the labels of large classes. The second term regularizes the coefficients, and thus can help to inhibit such errors that could otherwise occur when class sizes are imbalanced.

4 Joint Inference and Learning

We derive a variational EM algorithm where the goal is to jointly infer the labels of instances and estimate the relational model. With a few labels in different domains provided in advance by user (denoted as Y_L), the algorithm iterates between two steps: (1) Infer the distribution of the unknown labels (denoted as Y_U) based on both the extracted features and the current relational model R. (2) Estimate and update the relational model R using the labels provided by user and the hidden labels inferred in previous iteration.

We can derive such iterative procedure by considering the task of Maximum-a-posteriori (MAP) estimation of R

$$R^* = \underset{R}{\operatorname{argmax}}\, p(R|Y_L; \mathbf{X}), \quad \text{where } p(R|Y_L; \mathbf{X}) \propto p(R) \sum_{Y_U} p(Y_U, Y_L|R, \mathbf{X}). \quad (7)$$

Note that computing $p(R|Y_L; \mathbf{X})$ requires marginalizing over the unknown labels Y_U and is intractable. The variational methods tackle this problem by maximizing a tractable lower a bound of the log posterior. Formally, if q denotes any valid distribution of Y_U, then using Jensen's equality it is easy to obtain a lower bound of $\log[p(R)p(Y_L|R, \mathbf{X})]$, given by

$$J(R, q) = \mathrm{E}_q\{\log p(Y_U, Y_L|R, \mathbf{X})\} + \log p(R) + \mathrm{H}_q(q(Y_U)) \quad (8)$$

Further, it is well known (put some ref here) that equality holds when $q(Y_U) = p(Y_U|Y_L; R, \mathbf{X})$. In other words, maximizing the lower bound $J(R, q)$ with respect to both R and q will not only provide us with an estimate of R but also the posterior distribution over Y_U. The optimization of $J(R, q)$ w.r.t. R and q can be performed by iterating between the following steps.

$$\hat{q}^{(t+1)} = \underset{q}{\operatorname{argmax}}\, J(\hat{R}^{(t)}, q), \quad \text{(E-step)} \quad (9)$$

$$\hat{R}^{(t+1)} = \underset{R}{\operatorname{argmax}}\, J(R, \hat{q}^{(t+1)}). \quad \text{(M-step)} \quad (10)$$

The E-step in Eq.(9) infers the posterior distribution of the unknown labels Y_U using the current model $\hat{R}^{(t)}$. The M-step in Eq.(10) estimates the relational model R based on the updated distribution $\hat{q}^{(t+1)}(Y_U)$. However, solving Eq.(9) and Eq.(10) under our formulation is intractable and we need to resort to variational approximations.

Inferring Unknown Labels (E-STEP): The optimization problem in Eq.(9) can be made tractable using *mean field approximation* [19]. Formally, we restrict

q to be a factorized distribution: $q(Y_U) = \prod_{u=1}^{M} \prod_{i \in U_u} q_u^i(y_u^i)$. Here, U_u correspond to all unlabeled instances in domain $\mathcal{Y}_u$. The approximation results in the following closed form expressions for updating the posteriors:

$$\hat{q}_u^i(k) = \frac{1}{Z_u^i} \exp(\psi_u^i(k)). \tag{11}$$

where, $Z_u^i = \sum_{k'} \exp(\psi_u^i(k'))$ is the normalization constant, and $\psi_u^i(k)$ is:

$$\psi_u^i(k) = \alpha_u \sum_{j=1}^{N_u} w_u(i,j) q_u^j(k) + \sum_{v:(u,v)\in\mathcal{R}} \alpha_{uv} \sum_{j:i\sim j} \sum_{l=1}^{K_v} R_{uv}(k,l) q_v^j(l). \tag{12}$$

Note that despite the factorized form, the parameters of q_u^i for different instances are coupled to each other and effect each other. Further, as observed in Eq.(12), both feature similarity (first term) and cross-domain relations (second term) are utilized in the inference, leading to an estimate of the posterior that considers both within-domain and cross-domain information.

Estimating Relational Model (M-STEP): Given the inferred distribution q, we can estimate the relational model R by solving Eq.(10):

$$R^* = \underset{R}{\operatorname{argmax}}\, \mathrm{E}_q\{\log p(Y_L, Y_U|\mathbf{X}; R)\} - \log Z(\mathbf{X}; R) + \log p(R). \tag{13}$$

Note that the log-partition function $\log Z(\mathbf{X}; R)$ is intractable here. We use *tree-reweighted approximation* [20] to make it tractable. The basic idea is to divide the original model into tractable sub-models, and replace $\log Z(\mathbf{X}; R)$ with a convex combination of the log-partition functions of the sub-models. The substitution results in an upper bound of $\log Z(\mathbf{X}; R)$ [20]. In particular, we divide the joint model into affinity models and cross-domain relations, leading to the following upper bound:

$$\sum_{u=1}^{M} \theta_u A_u + \sum_{u \leftrightarrow v} \theta_{uv} B_{uv}(R_{uv}/\theta_{uv}) \tag{14}$$

Here A_u is the log-partition of the affinity model for $\mathcal{Y}_u$ that is independent of R, and B_{uv} is the log-partition of the cross-domain relation. The coefficients θ_u and θ_{uv} are the weights of the convex combination of the models. Such an approximation simplifies the maximization step and now each relation can be estimated respectively by solving:

$$R_{uv}^* = \underset{R_{uv}}{\operatorname{argmax}}\, E_q\{\Phi_{uv}(Y_u, Y_v; R_{uv})\} - \theta_{uv} B_{uv}(R/\theta_{uv}) + \log p(R_{uv}). \tag{15}$$

For simplicity, we set the weights to be $\theta_{uv} = 1/\#\text{relations}$. The objective is concave with a unique optimum and we use L-BFGS algorithm [21] to solve it.

5 Experiments

There are two publicly available datasets that are commonly used to evaluate research in personal photo tagging, which we call *E-Album* [22] and *G-Album* [23].

Since ground-truth labels are not provided, we estimate ground-truth by manually tagging each detected face. We also manually tag the event and location of each photo. We excluded the photos without any detected faces, and those whose ground-truth event and location labels could not be determined, leaving a subset of each album. In particular, *E-Album* contains 108 photos taken at 21 locations in 19 events, and 19 different people with 145 detected faces. *G-Album* contains 312 photos taken at 117 events, and 13 different people with 441 detected faces. The two albums give rise to different challenges. The sizes of the people classes in the E-Album are more unbalanced, while the G-Album has many more events, each containing only a small number of photos.

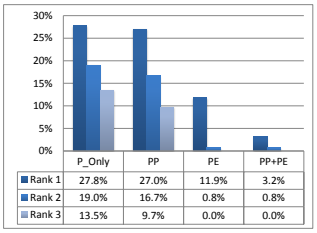
Feature extraction was performed as follows. For the people domain, we used the facial features proposed in [24]. A color histogram was used for the clothing. The location of the clothing relative to the face was determined using a simple geometric rule. For events we used the time-stamps as features. For locations, we used a color histogram of the background scene. For each feature, a distance measure is required. For the face features, we followed the algorithm in [24]. For clothes and location features, we used the Earth-mover's distance [25]. For events, we defined the distance to be 0 if the time-stamps were on the same day, and 1 otherwise. Finally, we need to compute the affinity weights $w_u(i,j)$. We experimented with a number of alternatives, and found that the best approach is to connect each unlabeled instance to just the closest K labeled instances, and set $w_u(i,j) = \exp(-d^2(\mathbf{x}_i, \mathbf{x}_j)/\sigma^2)$. The value of $w_u(i,j)$ for the other instances is set to zero. Here $d(\mathbf{x}_i, \mathbf{x}_j)$ is the distance between the features $\mathbf{x}_i$ and $\mathbf{x}_j$. We determined the optimal values of K and σ by cross validation.

Our algorithm outputs an estimate of the posterior probability of each label for each instance. To compute an error metric for our algorithm, we sort the candidate labels in terms of their posterior probabilities. We then compute rank-k error rates, the proportion of unlabeled instances whose top k candidate labels are all incorrect. To evaluate our algorithm, we generate a pre-labeled subset for each album by random sampling. For the people domain, we randomly chose 19 instances (13%) for the E-Album, and 49 instances (11%) for the G-Album. Here, we require that at least one instance is pre-labeled for each class. However, this requirement can be readily removed using active learning (see section 5.5), by which one can introduce new labels interactively.

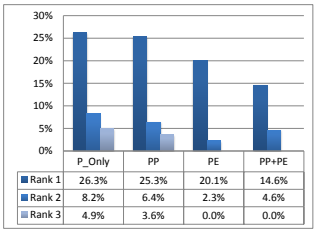
5.1 People Labeling

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 of quantitative evaluation are shown in Figure 2. 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



(a) Error rates on E-Album



(b) Error rates on G-Album

Fig. 2. Comparison of people labeling performance with different configurations.

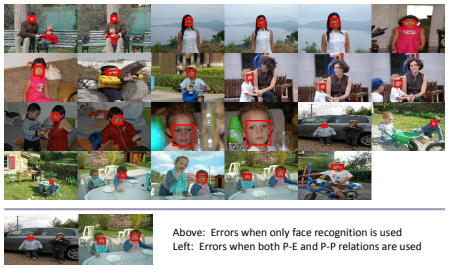


Fig. 3. All rank-1 errors for the E-Album. Above the delimiter: Errors made by our algorithm with no contextual relations (27.8%). Below the delimiter: Errors made by our algorithm with both the people-event and people-people relations (3.2%).

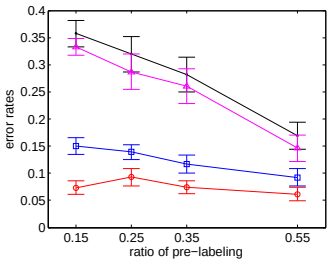


Fig. 4. The results of statistical significance testing obtained on E-Album with different percentages of pre-labeled instances. Curves from top to bottom obtained by: using only face, using people-people, using people-event, and using both relations.

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 3. In the supplemental material, we include a similar figure for the G-Album, together with movies illustrating the results.

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.

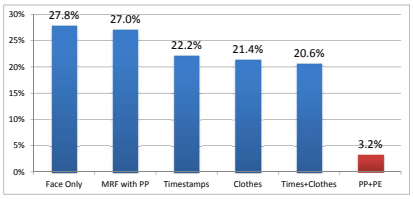


Fig. 5. Comparison between baseline approaches and ours on E-Album.

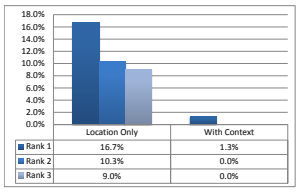


Fig. 6. Location error rates on E-Album.

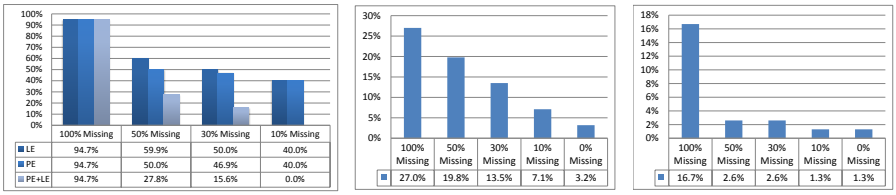
We also evaluated the performance of our framework with clothes features incorporated as conditional features. With both the people-event and people-people relations used, there are only three errors (3.2%) on the E-Album (see Figure 3). The clothing features are unable to correct any of these errors. For the G-Album, the conditional clothing features yield a slight improvement, with the best error rate reduced from 14.6% to 14.3%. The people-event and people-people relations are such powerful contextual cues that clothing adds little.

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 4 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 also performed such testing on G-Album, and the results are provided in supplemental materials. The improvement is significant across the entire range of pre-labeling percentage in both data sets.

5.2 Comparison with Other Approaches

Direct comparison with published methods is difficult due to: (1) lack of a standard testing protocol, e.g. which instances are tagged in advance, and (2) different features were used in different papers, and the features used in prior work are not available. Hence, the most appropriate way to make a fair comparison with other approaches is to implement them and evaluate them using exactly the same data and features that we used. In particular, we compared with a combination of face feature and time-stamp cues (as in [3]), a combination of face feature and clothes feature, and an adaptive combination of face feature and clothes feature conditioned on time stamps (as in [5]). We also note that previous work that used an MRF to capture exclusivity and the group prior (e.g. [8]) is essentially the special case of our framework where only the people-people relation is used. In all cases, we performed cross-validation to ensure that the best possible parameters were set for each particular algorithm.

Figure 5 contains the results on the E-Album. All of the feature-based algorithms yield a reasonable improvement with the rank-1 error rate being reduced from 27.8% to around $20\% - 22\%$. While the MRF model using just the people-people relation (group prior and exclusivity) does not yield a notable reduction of rank-1 errors, it improves the rank-2 and rank-3 performance far more (the error rates are reduced from 19.0% and 13.5% to 16.7% and 9.7% respectively.) How-



(a) Event labeling (b) People labeling (c) Location labeling

Fig. 7. The error rates of event, people and location labeling with varying percentages of missing time-stamps. For (b) the results were obtained with the P-E relation, and for (c) with the L-E relation.

ever, the performance improvement obtained by all of these methods is dwarfed by the improvement obtained by our algorithm when both the people-event and people-people relations are used (the rank-1 error is reduced to 3.2%).

Among the reasons that lead to such an improvement, the effective utilization of cross-domain context is the most important. Consider the people-event relation. When the event of a photo is inferred, the people classes that are not related to this event will be effectively ruled out from label selection (see Equations (5) and (12)), leaving only a very small subset of candidate labels to choose from. This resolves a great deal of ambiguity and makes recognition far easier.

5.3 Location Labeling

Figure 6 shows results for location estimation on the E-Album. We compare the results without any contextual information (location only) with those obtained using the event-location relation. The rank-1 error rate is reduced from 16.7% to 1.3%, and the rank-2 and rank-3 rates to 0%. Note that the event-location relation plays a similar role to the temporal priors used in video clustering [26].

5.4 Event Labeling with Missing Time-Stamps

The feature used for event labeling is the time-stamp of the photo. When present, this feature is very powerful; a temporal clustering of most photo collections breaks it naturally into events. In some cases time-stamps may be missing. For example, social networking sites such as Facebook remove timestamps. Furthermore, when merging two sets of photos collected on different cameras, it may not be wise to trust the time-stamps. In this section, we investigate what happens when time-stamps are missing.

We first investigated if we could estimate the event of a photo without the time-stamp. We randomly discarded $100\%, 50\%, 30\%$, and 10% of the time-stamps. The performance of event labeling under such conditions is shown in Figure 7(a). Note that we only compute the error rates over the photos without time-stamps. If all time-stamps are missing, we can only infer the event labels by random guessing, resulting in nearly 95% errors. If we know some of the time-stamps, both event-location and people-event relations can be used to estimate

a significant fraction of the event labels correctly. These two relations provide very complementary sources of information. The combination of the two is far better than either in isolation.

Next we investigated how the presence of missing time-stamps affects the performance of people and location labeling. In Figure 7(b) we see that the degradation in people-labeling performance with more and more missing time-stamps is very graceful. For location labeling, the removal of up to 50% of the timestamps hardly affects the performance. See Figure 7(c). So long as some photos captured in the same event retain their timestamps, the contextual benefit of the event-location relation is retained.

5.5 Labeling with Active Learning

As our framework estimates the posterior probabilities of the labels, it can be used for active learning [10]. By carefully choosing the order in which instances are pre-labeled, we can reduce the number of instances that need to be labeled to obtain a given recognition rate. We conducted preliminary experiments to illustrate this ability. In each iteration, we determine the unlabeled person instance that would lead to the maximum information gain and add it to the pre-labeled set. On average on the E-Album, it takes 30 iterations to obtain a rank-1 recognition rate of 95% for the people domain. In comparison, it requires 46 iterations with random sampling of the instances to be pre-labeled.

5.6 Timing Results

Our C# implementation runs in less than 2 seconds for both albums on a 2.0GHz Core-Duo laptop.

6 Conclusion

We have proposed the use of cross-domain relations as a mechanism to model context in multi-domain labeling (people, events, locations). Relation estimation and label inference are unified in a optimization algorithm. Our experimental results show that cross-domain relations provide a elegant, powerful, and general method of modeling context in vision-assisted tagging applications.

References

1. Gallagher, A.C., Tsuhan, C.: Using context to recognize people in consumer images. IPSJ Journal **49** (2008) 1234–1245
2. Zhang, L., Chen, L., Li, M., Zhang, H.: Automated annotation of human faces in family albums. In: 11th ACM Conf. on Multimedia. (2003)
3. Davis, M., Smith, M., Canny, J., Good, N., King, S., Janakiraman, R.: Towards context-aware face recognition. In: 13th ACM Conf. on Multimedia. (2005)

4. Davis, M., Smith, M., Stentiford, F., Bamidele, A., Canny, J., Good, N., King, S., Janakiraman, R.: Using context and similarity for face and location identification. In: SPIE'06. (2006)
5. Song, Y., Leung, T.: Context-aided human recognition - clustering. In: ECCV'06. (2006)
6. Naaman, M., Garcia Molina, H., Paepcke, A., Yeh, R.B.: Leveraging context to resolve identity in photo albums. In: ACM/IEEE-CS Joint Conf. on Digi. Lib. (2005)
7. Gallagher, A.C., Tsuhan, C.: Using a markov network to recognize people in consumer images. In: ICIP. (2007)
8. Gallagher, A.C., Chen, T.: Using group prior to identify people in consumer images. In: CVPR Workshop on SLAM'07. (2007)
9. Anguelov, D., Lee, K.c., Gokturk, S.B., Sumengen, B.: Contextual identity recognition in personal photo albums. In: CVPR'07. (2007)
10. Kapoor, A., Hua, G., Akbarzadeh, A., Baker, S.: Which faces to tag: Adding prior constraints into active learning. In: ICCV'09. (2009)
11. Torralba, A., Murphy, K.P., Freeman, W.T., Rubin, M.A.: Context-based vision system for place and object recognition. In: ICCV'03. (2003)
12. Torralba, A.: Contextual priming for object detection. Int'l. J. on Computer Vision **53** (2003) 169–191
13. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. In: ICCV'07. (2007)
14. Galleguillos, C., Rabinovich, A., Belongie, S.: Object categorization using co-occurrence, location and appearance. In: CVPR'08. (2008)
15. Li, L.J., Fei-Fei, L.: What, where and who? classifying events by scene and object recognition. In: CVPR'07. (2007)
16. Li, L.J., Socher, R., Fei-Fei, L.: Towards total scene understanding: Classification, annotation, and segmentation in an automatic framework. In: CVPR'09. (2009)
17. Cao, L., Luo, J., Kautz, H., Huang, T.S.: Annotating collections of photos using hierarchical event and scene models. In: CVPR'08. (2008)
18. Sutton, C., McCallum, A.: An Introduction to Conditional Random Fields for Relational Learning. In: Introduction to Statistical Learning. MIT Press (2007)
19. Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning **1** (2008) 1–305
20. Wainwright, M.J., Jaakkola, T., Willsky, A.: A new class of upper bounds on the log partition function. IEEE Transaction on Information Theory **51** (2005) 2313–2335
21. Byrd, R.H., Lu, P., Nocedal, J.: A limited memory algorithm for bound constrained optimization. SIAM Journal on SSC **16** (1995) 1190–1208
22. Cui, J., Wen, F., Xiao, R., Tian, Y., Tang, X.: Easyalbum: an interactive photo annotation system based on face clustering and re-ranking. In: SIGCHI. (2007) 367–376
23. Gallagher, A.C.: Clothing cosegmentation for recognizing people. In: CVPR'08. (2008)
24. Hua, G., Akbarzadeh, A.: A robust elastic and partial matching metric for face recognition. In: ICCV'09. (2009)
25. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. Int'l. Journal on Computer Vision **40** (2000) 99–121
26. Schroff, F., Zitnick, C., Baker, S.: Clustering videos by location. In: British Machine Vision Conference. (2009)