Correlational Gaussian Processes for Cross-domain Visual Recognition

Chengjiang Long
Kitware Inc.
28 Corporate Dr, Clifton Park, NY, USA 12065
chengjiang.long@kitware.com

Gang Hua
Microsoft Research Asia
Haidian District, Beijing, P.R. China 100080
ganghua@gmail.com

Abstract

We present a probabilistic model that captures higher-order co-occurrence statistics for joint visual recognition in a collection of images and across multiple domains. More importantly, we predict the structured output across multiple domains by correlating outputs from the multi-classes Gaussian process classifiers in each individual domain. A set of correlational tensors is adopted to model the relationship within a single domain as well as across multiple domains. This renders it possible to explore a high-order relational model instead of using just a set of pairwise relational models. Such tensor relations are based on both the positive and negative co-occurrences of different categories of visual instances across multi-domains. This is in contrast to most previous models where only pair-wise relationships are explored. We conduct experiments on four challenging image collections. The experimental results clearly demonstrate the efficacy of our proposed model.

1. Introduction

The cross-domain visual recognition problem was firstly explicitly proposed in [33], although many previous works [29, 3, 40, 4, 36, 37, 38, 7] also implicitly tackled part of such a problem. In such a problem, multiple visual recognition problems in different semantic domains are simultaneously solved through a joint formulation instead of being handled independently. This is based on the intuition that the semantics across different domains are associated with the same visual entity and hence there are intrinsic correlations among them to facilitate the joint inference of all of these visual semantics. For example, we can interpret each photo from people, location and event domain, and then employ the estimated cross-domain correlations to improve the recognition accuracy in each domain, *e.g.*, face recognition in people domain.

To better solve cross-domain visual recognition problems, we propose a probabilistic framework, namely *correlational Gaussian processes classifier* (CGPC), for joint visual recognition across a collection of images, and across multiple domains based on co-occurrence statistics of different instances across multiple domains. This is achieved by correlating the outputs of the Gaussian process classifiers from each individual domain, and formulate the joint visual recognition problem as a structured prediction problem. We choose Gaussian Process because it (1) is a non-parametric model to handle the linear and nonlinear data, (2) has a well-founded framework for learning and model selection, and (3) presents a good interpretation of model predictions.

We explore both the flip-noise model [23], which is good at handling label errors around the decision boundaries, and the robust likelihood model using a back-up mechanism [15], which is expected to be robust when the label errors occur far away from the decision boundaries, in the multi-class Gaussian process classifiers (MGPC) for each individual domain. Hence we obtain two versions of MGPC, dubbed SMGPC and RMGPC in brief with "S" indicating "standard" and "R" referring to "robust", respectively. Since both SMGPC and RMGPC have their own strengths to deal with different labeling noise scenarios, we determine the better one for our CGPC model based on the specific task in each individual domain. And then we integrate outputs from the multi-classes Gaussian process in each individual domain with a set of relational tensors based on both the positive and negative, and both the pair-wise and high-order cross-domain co-occurrence statistics.

Several aspects distinguish our proposed CGPC model from the existing works [43, 45, 24, 44, 6, 41, 32]. First, they often directly integrate the relational information with the input attributes in the Gaussian process prior [6, 43, 45] or encode relations as random variables conditioned on the latent function values of entities involved in pairwise relations [24, 44, 6, 41, 32]. In contrast, our proposed CGPC model integrates outputs from the multi-classes Gaussian process in each individual domain with a set of relational tensors based on the cross-domain co-occurrence statistics. Second, their models focus on predicting the relational tasks and it is not clear how these learned relations can be ex-

ploited to facilitate recognition, which is the focus of our proposed model.

We shall emphasize that unlike [33] which only explored positive pair-wise co-occurrence statistics as most existing works, we make full use of both the *positive* and *negative*, and both the pair-wise and high-order co-occurrence statistics, within a single domain and among the different domains through a unified relational model. The wisdom of using the negative co-occurrence statistics is that knowing what is not in a domain might be a very informative clue about the visual scope of both the current domain and the other related domains. This kind of negative co-occurrence statistics has been explored in image retrieval before [18].

To summarize, our contributions are four-folds: (1) we propose the correlational Gaussian processes classifier for joint visual recognition in a collection of images across multiple semantic domains; (2) we adopt a set of tensor parameters in the proposed model to flexibly capture both pairwise and high-order co-occurrence statistics; (3) we take both the positive and negative co-occurrence statistics among multi-domains into account in our model; and (4) we validate our proposed model on four challenging image collections including the SUN 09 dataset [5, 13], which clearly demonstrate the efficacy of our proposed models.

2. Related work

The related prior work can be roughly split into 2 categories: *Co-occurrence Statistics* and *Gaussian Processes for Statistic Relational Learning*.

Co-occurrence Statistics. As an important contextual cue, co-occurrence statistics have facilitated various computer vision tasks including image segmentation [29, 17, 10], object detection [3, 40], object category recognition [4, 36, 42], image annotation and retrieval [11], image and attributes classification [37, 38, 13], path prediction [46] and video summarization [7]. However, when modeling the co-occurrences of multiple semantics, the vast majority of previous works, if not all of them, only modeled the pairwise co-occurrence relationship. There were limited number of work that have attempted to model higher-order co-occurrence statistics [49, 26, 25, 27, 28], which deserves more exploration.

Gaussian Processes for Statistic Relational Learning. Gaussian processes have also been explored for statistic relational learning in several early works [2, 6, 41, 47, 48, 44, 1]. Boyle *et al.* [2] treated Gaussian processes as white noise sources convolved with smoothing kernels to handle multiple and coupled outputs. Chu *et al.* [6] developed a relational Gaussian process model which uses undirected linkages to incorporate both reciprocal relational information and input attributes. Yu *et al.* [47] proposed a stochastic relational model as a stochastic link-wise process induced by a tensor interaction of multiple Gaussian processes. Yu *et al.* [48] even proposed approximate Gaussian processes for modeling the inter-dependencies of edges in directed, undirected and bipartite networks.

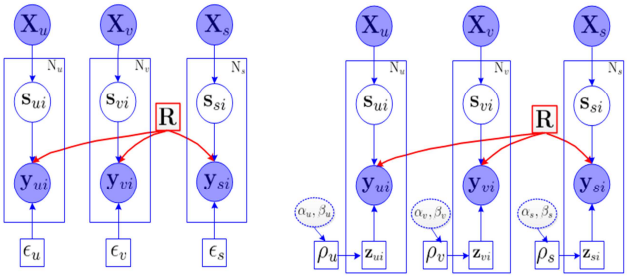


Figure 1: Graphical models of two different forms of our proposed CGPC model on three domains u, v and s. The blue parts indicate the independent MGPC (left: SMGPC, right: RMGPC) in each single domain and the red parts represent the correlation across multiple domains with the relational model $\mathbf{R}$.

Later, Xu *et al.* [43] proposed a multi-relational Gaussian process model to deal with an arbitrary number of relations and entity types in a single domain. Xu *et al.* [45] also generalized the model of Chu *et al.* [6] to a multi-way analysis model, namely *InfTucker*, by capturing nonlinear interactions among different tensor models. Nickel *et al.* [39] proposes a relational learning approach based on the factorization of a three-way tensor for multi-relational data. In contrast to these works which focus on covariance function to model the interactions among multiple Gaussian processes, tensors in our proposed CGPC model are intended to model the correlations among the outputs of MGPC from different domains. Our CGPC model also differs from existing transfer learning methods with deep Gaussian Processes [22], which focuses on the knowledge tranfer from a source deep GP to a target deep GP.

3. Formulation

We assume D semantic domains $\Omega = \{d|d = 1, \ldots, D\}$. For each domain d, there are N_d images. $\mathbf{X}_d$, $\mathbf{S}_d$, $\mathbf{Y}_d$ indicate the set of observed features, the set of random variables as soft labels and the set of true labels, respectively, corresponding to such N_d visual instances. More specifically, for each instance $\mathbf{x}_{di}$ in $\mathbf{X}_d$, its corresponding random variable $\mathbf{s}_{di}$ is used to jointly infer its true label y_{di} with relational models. For notation simplification, we denote $\mathbf{X}_d = \{\mathbf{x}_{di}|i = 1, \ldots, N_d\}$, $\mathbf{S}_d = \{\mathbf{s}_{di}|i = 1, \ldots, N_d\}$, $\mathbf{Y}_d = \{y_{di}|i = 1, \ldots, N_d\}$, $\mathbf{X} = \{\mathbf{X}_d|d \in \Omega\}$, $\mathbf{S} = \{\mathbf{S}_d|d \in \Omega\}$, and $\mathbf{Y} = \{\mathbf{Y}_d|d \in \Omega\}$.

We use a set of interaction tensors $\mathbf{R} = \{\mathbf{R}^{\mathbf{c}}|\mathbf{c} \in \mathcal{C}\}$, where $\mathbf{c}$ is a combination of domains that $\mathbf{R}^{\mathbf{c}}$ correlates and $\mathcal{C}$ is a set of such combinations, to represent the set of cross-domain relational models. For each $\mathbf{R}^{\mathbf{c}}$, depending on how

many domains we are modeling, it can conveniently model co-occurrence relations of arbitrary order. For example, if the relational model couples two domains, then $\mathbf{R^c}$ is a matrix. If the relational model models three domains, then $\mathbf{R^c}$ is a order-3 tensor. Figure 1 is one example to show our proposed CGPC model over a collection of images on three domains u, v and s. As illustrated in Figure 1, the conditional joint probability of the CGPC model is defined as

$$p(\mathbf{S},\mathbf{Y},\mathbf{R}|\mathbf{X}) \propto p(\mathbf{R})p(\mathbf{Y}|\mathbf{R},\mathbf{S},\mathbf{\Theta})\prod_{d\in\Omega} p(\mathbf{S}_d|\mathbf{X}_d), \quad (1)$$

where $\mathbf{\Theta} = \{\Theta_d | d \in \Omega\}$ and Θ_d is the hyperparameter associated with a specific domain d to deal with labeling errors in MGPC, *i.e.*, $\Theta_d = \epsilon_d$ for SMGPC and $\Theta_d = \{\alpha_d, \beta_d, \rho_d, \mathbf{z}_d\}$ for RMGPC (See details in Sec 3.1). $p(\mathbf{Y}|\mathbf{R},\mathbf{S},\mathbf{\Theta})$ is conditioned on both the relational model $\mathbf{R}$ and random variables $\mathbf{S}$, and $p(\mathbf{S}_d|\mathbf{X}_d)$ is dependent on $\mathbf{X}_d$.

In order to ease the learning and inference, we relax the conditional dependence by $p(\mathbf{Y}|\mathbf{R},\mathbf{S},\mathbf{\Theta}) \approx \frac{1}{Z(\mathbf{R},\mathbf{S})} p(\mathbf{Y}|\mathbf{R})p(\mathbf{Y}|\mathbf{S},\mathbf{\Theta})$ (where $Z(\mathbf{R},\mathbf{S})$ is normalized constant) and $p(\mathbf{Y}|\mathbf{S},\mathbf{\Theta}) = \prod_{d\in\Omega} p(\mathbf{Y}_d|\mathbf{S}_d,\Theta_d)$. We take this relaxation to make the inference tractable. Although $p(\mathbf{Y}|\mathbf{S})$ and $p(\mathbf{Y}|\mathbf{R})$ seem conditionally independent, the constant $Z(\mathbf{R},\mathbf{S})$ of the joint probability in this approximation still depends on $\mathbf{R}$ and $\mathbf{S}$, and therefore couples[1] the relational model with the classifier soft score. Hence, this treatment is reasonable and we obtain

$$p(\mathbf{S},\mathbf{Y},\mathbf{R}|\mathbf{X}) \propto p(\mathbf{R})p(\mathbf{Y}|\mathbf{R})\prod_{d\in\Omega} p(\mathbf{Y}_d|\mathbf{S}_d,\Theta_d)p(\mathbf{S}_d|\mathbf{X}_d), \quad (2)$$

where the second term $p(\mathbf{Y}|\mathbf{R})$ defines the prior probability based on the relational statistical tensor $\mathbf{R}$, which encodes both positive and negative co-occurrence relations and correlates the outputs from the different domains in Ω. And the last two terms $p(\mathbf{Y}_d|\mathbf{S}_d,\Theta_d)p(\mathbf{S}_d|\mathbf{X}_d)$ associates with the joint probability from MGPC in the single domain d, which can be further decomposed depending on the specific Gaussian process model adopted in each individual domain. More details about the terms will be described in the subsequent subsections. It worths mentioning that other classifiers can also be leveraged, but that is not we are focusing in this paper.

To clarify, we emphasize multiple domains to fully explore domain specific feature representations for recognition. Analogously, our method can be applied to different classes as different domains, where we regard class specific feature as the domain specific feature representation. Those classes sharing the same feature can be clustered to the same domain.

[1]The "coupling" here means that $Z(\mathbf{R},\mathbf{S})$ is the normalization constant covering both $p(\mathbf{Y}|\mathbf{R})$ and $p(\mathbf{Y}|\mathbf{S},\mathbf{\Theta})$, and it cannot be decomposed.

3.1. $p(\mathbf{Y}_d|\mathbf{S}_d,\Theta_d)p(\mathbf{S}_d|\mathbf{X}_d)$ from the MGPC

For each individual domain d, We select one MGPC from its two versions, *i.e.*, SMGPC and RMGPC, of which the difference is reflected in their treatment of labeling errors. Assuming there are l_d categories in domain d and $\mathbf{S}_d = \{\mathbf{S}_d^k | k = 1,\ldots,l_d\}$, then the Gaussian process prior of the overall function value $\mathbf{S}_d^k$ is defined as $p(\mathbf{S}_d^k|\mathbf{X}_d) = \mathbf{N}(\mathbf{0},\mathbf{K}_d^k)$, where $\mathbf{K}_d^k$ denotes the $N_d \times N_d$ covariance matrix whose ij-th entry is computed with the corresponding covariance function associated with the category k.

We proceed to introduce these two MGPCs.

3.1.1 SMGPC

By reducing the multi-classes classification to binary cases, we can denote $\mathbf{Y}_d = \{\mathbf{Y}_d^k | k = 1,\ldots,l_d\}$, where $\mathbf{Y}_d^k = \{y_{di}^k | i = 1,\ldots,N_d\}$ is a set of binary labels. y_{di}^k is 1 if the instance $\mathbf{x}_i$ belongs to category k and 0 otherwise in domain d. Then $p(\mathbf{Y}_d|\mathbf{S}_d,\Theta_d)p(\mathbf{S}_d|\mathbf{X}_d)$ from SMGPC in each individual domain d associated with each category k can be further decomposed as

$$p(\mathbf{Y}_d|\mathbf{S}_d,\Theta_d)p(\mathbf{S}_d|\mathbf{X}_d) \propto \prod_k p(\mathbf{S}_d^k|\mathbf{X}_d)\prod_{i=1}^{N_d} p(y_{di}^k|s_{di}^k,\varepsilon_d), \quad (3)$$

where Θ_d denotes ε_d and the conditional likelihood model $p(y_{di}^k|\mathbf{s}_{di}^k,\varepsilon_d)$ is defined as a flip noise model [23], *i.e.*,

$$p(y_{di}^k|s_{di}^k,\varepsilon_d) = \varepsilon_d H(y_{di}^k s_{di}^k) + (1-\varepsilon_d)H(-y_{di}^k s_{di}^k), \quad (4)$$

where $H(x) = 1$ if $x > 0$ and $H(x) = 0$ otherwise. In other words, the *a posteriori* estimation of y_{di}^k takes the sign of the predicted soft label s_{di}^k with probability $\varepsilon_d (0 \le \varepsilon_d \le 1)$, and hence ε_d can be used to model the global labeling error rate in domain d, which can be estimated by the EP-EM algorithm [16, 34, 35].

3.1.2 RMGPC

Different from SMGPC, $p(\mathbf{Y}_d|\mathbf{S}_d,\Theta_d)p(\mathbf{S}_d|\mathbf{X}_d)$ from RMGPC for each individual domain d is defined as

$$\begin{aligned} &p(\mathbf{Y}_d|\mathbf{S}_d,\Theta_d)p(\mathbf{S}_d|\mathbf{X}_d) \\ &\propto p(\mathbf{S}_d|\mathbf{X}_d)p(\rho_d)p(\mathbf{z}_d|\rho_d)\prod_{i=1}^{N_d} p(y_{di}|\mathbf{s}_{di},z_{di}), \end{aligned} \quad (5)$$

where ρ_d is the prior fraction of training instances expected to be outliers, and $\mathbf{z}_d = \{z_{d1},\ldots,z_{dN_d}\}$ is a set of binary latent variable for each visual instance to indicate whether $s_{di}^{y_{di}} \ge s_{di}^k$ for any $k \ne y_{di}$ ($z_{di} = 0$) or not ($z_{di} = 1$). Note that $s_{di}^{y_{di}}$ here denotes the latent score of instance $\mathbf{x}_i$ belong to the ground-truth category y_{di} in domain d.

$p(y_{di}|\mathbf{s}_{di},z_{di})$ is a *back-up* mechanism to handle label noises and is defined as

$$p(y_{di}|\mathbf{s}_{di},z_{di}) = \left[\prod_{k\ne y_{di}} H(s_{di}^{y_{di}} - s_{di}^k)\right]^{1-z_{di}} \left[\frac{1}{l_d}\right]^{z_{di}}. \quad (6)$$

Note that the first term directly depends on the accuracy of $\mathbf{s}_{di}^{y_{di}}$. In particular, it takes value 1 when the corresponding instance is correctly classified and 0 otherwise. RMGPC is expected to be robust when the observed data contain labeling errors far from the decision boundaries, because the likelihood function described in Eq. 6 considers only the total number of prediction errors made by $\mathbf{s}_{di}^{y_{di}}$, rather than the distance of these errors to the decision boundary.

$p(\mathbf{z}_d|\rho_d)$ is defined as a factorizing multivariate Bernoulli distribution

$$p(\mathbf{z}_d|\rho_d) = Bern(\mathbf{z}_d|\rho_d) = \prod_{i=1}^{N_d} \rho_d^{z_{di}}(1-\rho_d)^{1-z_{di}}. \quad (7)$$

And the prior for ρ_d is set to be a conjugate beta distribution, *i.e.*,

$$p(\rho_d) = Beta(\rho_d|\alpha_d, \beta_d) = \frac{\rho_d^{\alpha_d-1}(1-\rho_d)^{\beta_d-1}}{B(\alpha_d, \beta_d)}, \quad (8)$$

where $B(\cdot,\cdot)$ is the the beta function, and both α_d and β_d are hyper-parameters as part of Θ_d in Eq. 5. We suggest $\alpha_d = 1$ and $\beta_d = 9$.

Discussion: SMGPC is good at dealing with the additive Gaussian noises near the decision boundaries, while RMGPC is more robust to handle the scenario when the labeling errors are far from the decision boundaries. This suggests that we should determine which one of them to be used for each individual domain based on the specific task.

3.2. $p(\mathbf{Y}|\mathbf{R})$ to Model Co-occurrences

Next, we define $p(\mathbf{Y}|\mathbf{R})$ with the tensor-based relational model $\mathbf{R}$, a complicated combined relational model consisted of a set of different relational models $\mathbf{R}^{\mathbf{c}}$, where $\mathbf{c}$ is a combination of co-occurring domains denoted as $d_1 \sim \ldots \sim d_{|\mathbf{c}|}$ ($|\mathbf{c}|$ is the cardinality of $\mathbf{c}$). We let $\mathbf{j}$ be a set of instance indices of a co-occurrence associated with the corresponding domains $\mathbf{c}$ denoted as $j_1 \sim \ldots \sim j_{|\mathbf{c}|}$, and then $\Phi(\mathcal{Y}_{\mathbf{j}}^{\mathbf{c}}|\mathbf{R}^{\mathbf{c}})$ represents the *relational potential* to measure the co-occurring labels $\mathcal{Y}_{\mathbf{j}}^{\mathbf{c}} = \{y_{d_k j_k} | d_k \in \mathbf{c}, j_k \in \mathbf{j}\}$ with the relational model $\mathbf{R}^{\mathbf{c}}$. Denoting $\mathcal{C}$ to be a set of $\mathbf{c}$ associated with $\mathbf{R}$ and $\mathcal{O}(\mathbf{c})$ to be a set of $\mathbf{j}$ covering all co-occurrences on $\mathbf{c}$, we define

$$p(\mathbf{Y}|\mathbf{R}) \propto \exp\left\{\sum_{\mathbf{c}\in\mathcal{C}}\sum_{\mathbf{j}\in\mathcal{O}(\mathbf{c})} \alpha_{\mathbf{c}}\Phi(\mathcal{Y}_{\mathbf{j}}^{\mathbf{c}}|\mathbf{R}^{\mathbf{c}})\right\}, \quad (9)$$

where $\alpha_{\mathbf{c}}$ is the parameter associated with the relational model $\mathbf{R}^{\mathbf{c}}$ and its value can be determined by 5-fold cross-validation in the range [0.01, 1.0].

To clarify, the relational potential $\Phi(\mathcal{Y}_{\mathbf{j}}^{\mathbf{c}}|\mathbf{R}^{\mathbf{c}})$ can cover the relational models both within a single domain and across multiple domains. In particular, a pairwise relational model associated with any two instances co-occurring within a single domain can be viewed as a particular cross-domain relational that couples the two same domains. In this paper, the relational model $\mathbf{R}^{\mathbf{c}}$ is pairwise or high-order co-occurring coefficients among labels $\mathcal{Y}_{\mathbf{j}}^{\mathbf{c}}$. If $\mathbf{R}^{\mathbf{c}}$ is based on only the positive co-occurrences, then $\Phi(\mathcal{Y}_{\mathbf{j}}^{\mathbf{c}}|\mathbf{R}^{\mathbf{c}})$ can be defined as

$$\Phi(\mathcal{Y}_{\mathbf{j}}^{\mathbf{c}}|\mathbf{R}^{\mathbf{c}}) \doteq \sum_{y_{d_1}\sim\ldots\sim y_{d_{|\mathbf{c}|}}} \mathbf{R}^{\mathbf{c}}(y_{d_1},\ldots,y_{d_{|\mathbf{c}|}}) \times I(y_{d_1} = y_{d_1 j_1})\ldots I(y_{d_{|\mathbf{c}|}} = y_{d_{|\mathbf{c}|} j_{|\mathbf{c}|}}).$$

Intuitively, larger value of the tensor element $\mathbf{R}^{\mathbf{c}}(y_{d_1},\ldots,y_{d_{|\mathbf{c}|}})$ indicates that the combination of labels $y_{d_1} \sim \ldots \sim y_{d_{|\mathbf{c}|}}$ co-occur more frequently, and will encourage $y_{d_k i_k}$ to be assigned y_{d_k} for any $d_k \in \mathbf{c}$. Hence, maximizing $\Phi(\mathcal{Y}_{\mathbf{j}}^{\mathbf{c}}|\mathbf{R}^{\mathbf{c}})$ should lead to the labels that are consistent with the relation.

However, knowing what is not related to the label in one domain might also be a very informative clue about the visual recognition in both the current domain and the other related domains. Hence, besides $\mathbf{R}_+^{\mathbf{c}}$ denoting the relational model based on all the positive co-occurrences, we also consider the other relational models, *i.e.*, (1) $\mathbf{R}_-^{\mathbf{c}}$ based on all the negative co-occurrences (completely $\not\equiv \mathcal{Y}_{\mathbf{j}}^{\mathbf{c}}$), and (2) $\mathbf{R}_{+-}^{\mathbf{c}_1\mathbf{c}_2}$ based on the partly positive and partly negative co-occurrences, where $\mathbf{c}_1$ is the part of domains in which the labels are positive ($\equiv \mathcal{Y}_{\mathbf{j}}^{\mathbf{c}_1}$) and $\mathbf{c}_2$ is the other part of domains where the labels are negative ($\not\equiv \mathcal{Y}_{\mathbf{j}}^{\mathbf{c}_2}$) in the co-occurrences. Note that $\mathcal{Y}_{\mathbf{j}}^{\mathbf{c}} \doteq \mathcal{Y}_{\mathbf{j}}^{\mathbf{c}_1} \sim \mathcal{Y}_{\mathbf{j}}^{\mathbf{c}_2}$ means $\mathcal{Y}_{\mathbf{j}}^{\mathbf{c}_1}$ and $\mathcal{Y}_{\mathbf{j}}^{\mathbf{c}_2}$ occur in cross-domain $\mathbf{c}$ positively. We believe both $\mathbf{R}_-^{\mathbf{c}}$ and $\mathbf{R}_{+-}^{\mathbf{c}_1\mathbf{c}_2}$ are beneficial to estimate the underlying true $\mathbf{R}^{\mathbf{c}}$ in the entire data distribution. Therefore, based on both the positive and negative co-occurrences, the $\mathbf{R}^{\mathbf{c}}$ can be defined as the weighted sum of these relational tensors, *i.e.*,

$$\mathbf{R}^{\mathbf{c}} = w_+^{\mathbf{c}}\mathbf{R}_+^{\mathbf{c}} + w_-^{\mathbf{c}}\mathbf{R}_-^{\mathbf{c}} - \sum_{\mathbf{c}_1,\mathbf{c}_2} w_{+-}^{\mathbf{c}_1\mathbf{c}_2}\mathbf{R}_{+-}^{\mathbf{c}_1\mathbf{c}_2}, \quad (10)$$

where the weights $w_+^{\mathbf{c}}$, $w_-^{\mathbf{c}}$ and $w_{+-}^{\mathbf{c}_1\mathbf{c}_2}$ are in the range $[0,1]$ and can be learned from the observed data or by cross-validation. The reason why we set the positive sign on $\mathbf{R}_+^{\mathbf{c}}$ and $\mathbf{R}_-^{\mathbf{c}}$ in Eq. 10 is that we expect $\mathbf{R}_+^{\mathbf{c}}$ and $\mathbf{R}_-^{\mathbf{c}}$ to be able to reflect the underlying true co-occurring relationship and hence better model the probability of co-occurrences. This has been verified by our experimental observations. For simplification, we even can set them as $w_+^{\mathbf{c}} = 1$, $w_-^{\mathbf{c}} = \frac{1}{\prod_{d_k\in\mathbf{c}}(l_{d_k}-1)}$ and $w_{+-}^{\mathbf{c}_1\mathbf{c}_2} = \frac{1}{\prod_{d_k\in\mathbf{c}_2}(l_{d_k}-1)}$. The weights are set in such a way that we rely a bit more on the positive co-occurrences because they may be more reliably estimated from the data.

3.3. Relational Model Prior $p(\mathbf{R})$

To avoid over-fitting, we need to regularize the relational model $\mathbf{R}$. Ideally, $\mathbf{R}$ is able to capture truly stable relations rather than those only due to the occasional co-occurrences. We impose L1 regularization on $\mathbf{R}$ to enforce sparsity and learn stable relations to rule out the influence of the non-stable co-occurrences. To avoid some elements in $\mathbf{R}$ becoming too large so that instances in small-size classes are

incorrectly classified when the calss sizes are imblanced, we also incorporate L2 regularization on $\mathbf{R}$. Such a regularization with L1 and L2 norm is embedded into the prior probability of $\mathbf{R}$, *i.e.*,

$$p(\mathbf{R}) \propto \exp\{-\beta_1 \| \mathbf{R} \|_1 - \beta_2 \| \mathbf{R} \|_2\}, \quad (11)$$

where β_1 and β_2 are adjusting weights whose values are determined by 5-fold cross-validation in the range [0.01, 1.0].

4. Joint Inference and Learning

Inspired by the ideas in [33], we derive a variational EM algorithm with the goal to jointly infer the labels of instances and estimate the relational model. With a few labels in different domains, denoted as $\mathbf{Y}_L = \{\mathbf{Y}_{dL} | d \in \Omega\}$, provided in advance by user, and any valid distribution $q(\mathbf{Y}_U)$ of the unknown labels $\mathbf{Y}_U = \{\mathbf{Y}_{dU} | d \in \Omega\}$ based on both the output probabilities from the MGPC and the current relational model $\mathbf{R}$ and the observed feature $\mathbf{X} = \{\mathbf{X}_{dL}, \mathbf{X}_{dU} | d \in \Omega\}$, we can obtain a lower bound

$$\begin{aligned} J(\mathbf{R}, q) &= \mathbf{E}_q \{\log p(\mathbf{Y}_U, \mathbf{Y}_L | \mathbf{R})\} \\ &+ \mathbf{E}_q \left\{ \sum_{d \in \Omega} \log p(\mathbf{Y}_{dU} | \mathbf{Y}_{dL}, \mathbf{X}_d, \Theta_d) \right\} \\ &+ \log p(\mathbf{R}) + \mathbf{H}_q(q(\mathbf{Y}_U)), \end{aligned} \quad (12)$$

where $\mathbf{H}_q(q(\mathbf{Y}_U))$ is the entropy of $q(\mathbf{Y}_U)$ and

$$p(\mathbf{Y}_{dU} | \mathbf{Y}_{dL}, \mathbf{X}_d, \Theta_d) = \prod_{\mathbf{x}_{du} \in \mathbf{X}_{dU}} p(\mathbf{y}_{du} | \mathbf{X}_{dL}, \mathbf{Y}_{dL}, \mathbf{x}_{du}, \Theta_d),$$

where $p(\mathbf{y}_{du} | \mathbf{X}_{dL}, \mathbf{Y}_{dL}, \mathbf{x}_{du}, \Theta_d)$ is the output for a unlabeled instance $\mathbf{x}_{du}$ from the MGPC. It is well known that Eq. 12 holds when $q(\mathbf{Y}_U) = p(\mathbf{Y}_U | \mathbf{Y}_L, \mathbf{R}, \mathbf{X})$, *i.e.*, maximizing the lower bound $J(\mathbf{R}, q)$ with respect to both $\mathbf{R}$ and q will not only provide us with an estimate of the relational model $\mathbf{R}$ but also the posterior distribution over $\mathbf{Y}_U$. The EM algorithm can be described by the two iterative steps:

- **E-step**: Infer the distribution of $\mathbf{Y}_U$ based on both the extracted features and the current relational model $\hat{\mathbf{R}}^{(t)}$ by $\hat{q}^{(t+1)} \leftarrow \arg\max_q J(\hat{\mathbf{R}}^{(t)}, q)$.
- **M-step**: Estimate and update the relational model using the labels provided by user and the hidden labels inferred in previous iteration by $\hat{\mathbf{R}}^{(t+1)} \leftarrow \arg\max_{\mathbf{R}} J(\mathbf{R}, \hat{q}^{(t+1)})$.

We consider all the possible combinations using both the known labels and the predicted labels. It is worth mentioning that each component of $\mathbf{R}^{\mathbf{c}}$ in Eq. 10 is estimated in the **M-step**. Once the EM algorithm converges, it outputs an estimate of the posterior probability of each label for each instance.

Discussion: Regarding the relationship between learning MGPC and learning $J(\mathbf{R}, \mathbf{Y})$, the hyper-parameters are fit before learning $J(\mathbf{R}, \mathbf{Y})$. And only $\mathbf{R}$ and $\mathbf{Y}$ are iteratively estimated using a variation EM algorithm.

5. Experiments

Our experiments are first carried out on three image collections, *i.e.*, the E-Album [8] and the G-Album [12], and a newly published VP dataset [21], in which we measure the performance with Rank-1 recognition accuracy. We also extend the experiments to a larger dataset SUN 09 [5, 13], where multiple concepts co-occur in the images and multiple domains can come from a random split of the concepts.

5.1. Experiments on the E-Album and the G-Album

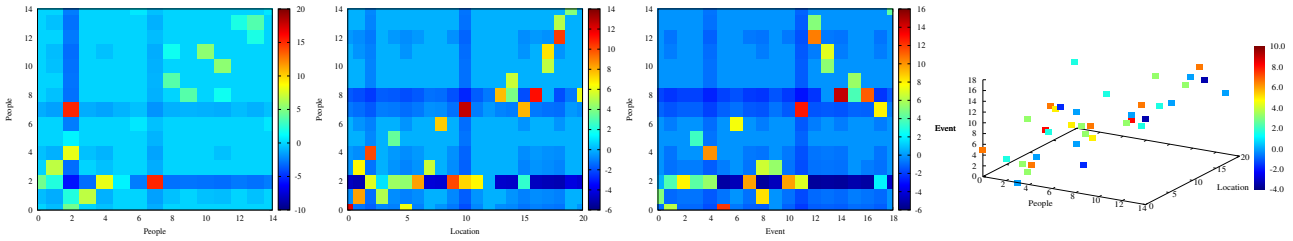
The E-Album is consisted of 108 photos taken at 21 locations in 19 events, and 15 different people in 145 detected faces. The G-Album has 312 photos taken at 141 locations in 117 events, 13 different people in 441 detected faces. We conduct experiments on three domains: people, location and event. The feature we use for people domain is the 100-dimensional feature with a probabilistic elastic part (PEP) representation [30] extracted from detected faces after resizing them to 150 pixels by 150 pixels. For the location domain, we extract 512-dimensional GIST feature from each photo. For the event domain,we use a vector of 374 attributes detector probabilities to be a 374-dimensional attribute feature by adopting the VIREO-374 SVM models provided in [19, 20].

For each domain, we adopt the RBF kernel because it is a squared exponential kernel and in general more flexible than linear or polynomial kernels so that we can model a whole lot more functions with its functional space. As for the similarity or distance measurements, we evaluate the Earth-mover's distance with L1 norm, the Earth-mover's distance with L2 norm, the L1 distance, the L2 distance, and the similarity score from the Joint-Bayesian classifier [31]. We name these 5 different dense RBF kernels as EMDL1-K, EMDL2-K, L1-K, L2-K and JB-K. In addition, we evaluate our proposed algorithm with the original kernel, which we call Lin-Kernel, used in [33] and provided by the authors of [33]. Note that Lin-Kernel is sparse because the nonzero elements in the kernel matrix only occupy a very small percentage (ranging between 2% and 5%).

It worths mentioning here that we use the same setting of pre-labeled subset adopted in [33] for convenience of comparison. We firstly focus on face recognition to show the efficacy of relational models and then evaluate the other recognition tasks on and across all domains.

5.1.1 Visualization of relational models

To better understand the sparsity of relational models in Section 3.3, we visualize the four above-mentioned relational models on the E-Album, as demonstrated in Figure 2. Since PP, PL and PE are pairwise relational models, we adopt colormap to plot the matrixes as in Figure 2a, 2b and 2c. The observations show that only a few number of elements are nonzero. This indicates that pairwise relational



(a) PP (b) PL (c) PE (d) PLE

Figure 2: Visualization of 4 different relational models on the E-Album. Note that 0-14, 0-20 and 0-18 are the label IDs for People, Location and Event domains, respectively. And for (d), the zero-valued elements are not plotted.

Table 1: Face recognition performance with 4 relational models and 6 kernels on the E-Album.(unit: %)

	EMDL1-K	EMDL2-K	L1-K	L2-K	Lin-Kernel	JB-K
P-only	35.71	72.22	67.46	71.43	73.81	**86.51**
PP+	66.67	73.81	71.43	72.22	75.40	**88.89**
PP±	69.84	75.40	73.81	73.02	76.19	**90.48**
PL+	76.19	86.51	85.71	86.51	87.30	**95.24**
PL±	79.37	92.06	90.48	90.48	88.89	**96.83**
PE+	76.19	87.30	85.71	86.51	89.68	**95.24**
PE±	79.37	92.06	90.48	91.27	91.47	**96.83**
PLE+	72.22	86.51	85.71	86.51	87.30	**95.24**
PLE±	76.98	87.30	86.51	87.30	89.68	**96.83**

Table 2: Face recognition performance with 4 relational models and 6 kernels on the G-Album.(unit: %)

	EMDL1-K	EMDL2-K	L1-K	L2-K	Lin-Kernel	JB-K
P-only	53.57	76.28	76.53	75.51	76.02	**82.65**
PP+	70.66	76.53	76.78	77.04	77.30	**82.91**
PP±	72.70	77.81	78.06	78.31	77.81	**84.18**
PL+	68.37	80.10	81.38	80.61	80.61	**83.93**
PL±	69.90	82.14	83.67	83.16	81.63	**84.18**
PE+	70.92	81.63	82.65	81.89	81.89	**86.22**
PE±	72.70	84.43	84.44	84.69	82.91	**88.78**
PLE+	72.70	81.91	84.69	82.40	81.89	**85.46**
PLE±	74.23	82.91	84.95	83.42	82.40	**86.48**

models in our proposed framework have a good property of sparsity as we expected. As for the high-order relational model PLE, we plot the elements whose values are nonzero in Figure 2d. It is easy to observe that the high-order relational model PLE is also sparse.

5.1.2 Efficacy of Positive and Negative Co-occurrences

To evaluate the efficacy of the relational models based on both the positive and negative co-occurrences (marked with the suffix "±"), we conduct experiments with 3 pairwise relational models, *i.e.*, People-People relational model (PP) within the people domain, People-Location relational model (PL) cross the people and location domains, People-Event relational model (PE) cross the people and event domains, and the high-order People-Location-Event relational model (PLE) cross all these 3 domains, on the two albums. For each relational model, we compare its performance with two baselines: the performance without any relational models is called P-only, and the performance with the same relational model but using the positive co-occurrences only (marked with the suffix "+").

We evaluate these 4 relational models on 6 different kernels. For each individual domain with specific kernel on the two albums, the selection of MGPCs for our CGPC model is dependent on the experimental performance of SMGPC and RMGPC.

Here we summarize the results in Table 1 and 2. It can be seen that: (1) with any relational model, the recognition accuracy has been consistently improved when compared to P-only; (2) the relational model based on both the positive and negative co-occurrences always performs better than the corresponding relational model using the positive co-occurrences only. Apparently, exploring the negative co-occurrences together enables the better performance.

As observed, the performance of PLE alone does not necessarily perform better than that of PE. This can be explained by the fact that, in general, compared with pairwise relational model, high-order co-occurrence statistics need more observed examples to obtain a reliable estimation since they correlates more number of domains. However, we will show in the next section that combining PLE with PP and PE together will provide another boost since they are complementary. (Please see Section 5.1.3)

5.1.3 Comparison with the state-of-the-art

We compare our proposed CGPC with Lin's cross-domain learning method [33], denoted as "Lin-CDL", on both the E-Album and the G-Album for face recognition, location recognition and event recognition across 3 domains. To our best acknowledge, [33] is the only work that also explores cross-domain recognition so that we take it as the baseline. To guarantee that the comparison is fair, for each individual domain, we use two kernels, *i.e.*, the sparse kernel Lin-Kernel and the dense kernel JB-K, and distinguish them with the prefixes "K_s-" and "K_d-", respectively. The suffixes "+" and "±" are the same as in Section 5.1.2.

To clarify, the high recognition performance of our proposed CGPC model comes from two parts: (1) the output of MGPC associated with the specific recognition task in the domain d; and (2) the cross relational models with the aids of the other domains. In this paper, we use S-CGPC and R-CGPC to emphasize the selection of MGPC (*i.e.*, SMGPC and RMGPC, respectively) for the domain associated with

Table 3: Performance comparison of face recognition on the E-Album.(unit: %)

	P-only	PP	PE	PLE	PP+PE	PP+PE+PLE
K_s-Lin	72.22	73.02	88.10	–	**96.83**	–
K_d-Lin	38.89	46.03	72.22	–	90.48	–
K_s-S+	73.81	74.60	88.89	86.51	96.83	96.83
K_s-S±	73.81	75.40	89.68	87.30	96.83	**97.62**
K_d-S+	84.92	86.89	94.44	93.65	96.83	**97.62**
K_d-S±	84.92	89.68	95.24	94.44	**97.62**	**97.62**
K_s-R+	73.81	75.40	89.68	87.30	96.83	97.62
K_s-R±	73.81	76.19	91.47	89.68	97.62	97.62
K_d-R+	86.51	88.89	95.24	95.24	97.62	97.62
K_d-R±	86.51	90.48	96.83	96.83	97.62	**98.41**

Table 4: Performance comparison of face recognition on the G-Album. (unit: %)

	P-only	PP	PE	PLE	PP+PE	PP+PE+PLE
K_s-Lin	73.72	74.74	79.85	–	**85.46**	–
K_d-Lin	40.56	41.33	67.09	–	75.26	–
K_s-S+	74.23	75.26	81.12	80.88	86.99	88.27
K_s-S±	74.23	76.78	81.89	82.14	87.76	89.03
K_d-S+	81.89	82.65	84.69	84.44	88.52	89.54
K_d-S±	81.89	83.16	86.73	85.45	89.80	**90.56**
K_s-R+	76.02	77.30	81.89	81.89	87.50	89.03
K_s-R±	76.02	77.81	82.91	82.40	88.78	90.05
K_d-R+	82.65	82.91	86.22	85.46	89.03	90.31
K_d-R±	82.65	84.18	88.78	86.48	90.56	**92.09**

the recognition task. For example, when we evaluate face recognition in the people domain with R-CGPC, the output of people domain is obtained from RMGPC. To shorten names from Table 3 to Table 8, we use "Lin", "S" and "R" to indicate Lin-CDL, S-CGPC and R-CGPC, respectively.

Face recognition Besides the single pairwise relational models mentioned above, we also combine PP and PE together to form a combined relational model PP+PE. What's more, we even include PLE into PP+PE to form a more complex relational model PP+PE+PLE.

The observations in Table 3 and 4 show that for both S-CGPC and R-CGPC, (1) they perform well with both the sparse and dense kernels, while Lin-CDL only works well when using the sparse kernel; (2) they always outperform Lin-CDL based on the same kernel, whether using the relational models based on both the positive and negative co-occurrences, or using the relational models with the positive co-occurrences only ; (3) again, the relational models based on both the positive and negative co-occurrences are always better than those using the positive co-occurrences only, whether with the sparse or dense kernel; (4) they can model the high-order relational model PLE and incorporating it can help improve the final recognition accuracy; (5) the combined relational models are better than the single relational models, which shows that the relational models are complementary to each other; (6) R-CGPC performs better than S-CGPC no matter using the sparse or dense kernel on both the two albums or not; and (7) with the dense kernel, our proposed algorithm R-CGPC can improve the recognition accuracy by about 2% on the E-Album and 7% on the G-Album when compared with the best results in [33].

Also, it is worth paying attention to the comparison between the PP+PE's results and the corresponding PP+PE+PLE's results on the E-Album, and the recognition accuracy 96.83%, 97.62% and 98.41% mean that only 4, 3 and 2 testing examples are misclassified, respectively. Although the gain is small, considering that these are the really difficult visual instances, the improvement from 96.83% to 97.62% and from 97.62% to 98.41% still confirm the power of incorporating the high-order relational model PLE.

Location and event recognition We also run experiments to evaluate the performances of location recognition and event recognition on the other two domains. To make it simple, we only evaluate the relational model LE, the high-order relational model PLE, and the combined relational model LE+PLE. Similarly, "L-only" and "E-only" denote the baseline without any relational models in location and event recognition, respectively.

Table 5: Performance comparison of location recognition on the E-Album (left) and the G-Album (right). (unit: %)

	L-only	LE	PLE	LE+PLE	L-only	LE	PLE	LE+PLE
K_d-Lin	62.82	91.02	–	–	23.92	80.36	–	–
K_d-S+	83.33	92.30	87.17	97.43	27.61	82.21	76.07	85.27
K_d-S±	83.33	96.15	89.74	**98.87**	27.61	85.89	80.98	**87.12**
K_d-R+	84.61	94.87	91.03	98.87	29.45	84.66	79.14	87.73
K_d-R±	84.61	98.71	93.59	**100.00**	29.45	87.12	83.43	**89.57**

Table 6: Performance comparison of event recognition on the E-Album (left) and the G-Album (right). (unit: %)

	E-only	LE	PLE	LE+PLE	E-only	LE	PLE	LE+PLE
K_d-Lin	26.42	60.37	–	–	9.15	41.54	–	–
K_d-S+	43.40	62.26	58.49	67.92	11.27	52.11	48.59	55.63
K_d-S±	43.40	66.04	60.38	**69.81**	11.27	56.33	50.70	**59.15**
K_d-R+	47.17	67.92	64.15	69.81	12.68	54.92	49.30	58.45
K_d-R±	47.17	69.81	66.04	**71.69**	12.68	57.74	51.41	**60.56**

As shown in Table 5 and Table 6, the efficacy of our proposed approach in boosting both the location recognition accuracy and the event recognition accuracy is largely consistent with the claims in the face recognition.

5.2. Experiments on the VP dataset

The VP dataset [21] contains 1124 images (811 images for training and 313 images for testing) of 8 politicians. In this paper, we use 3 domains, *i.e.*, people, gesture and scene, with the task to improve the performance of face recognition, gesture recognition and scene recognition with the aid of the cross-domain relational models. Since the task is different from [21], we only use the shared original images and the labels for the people domain and collect the labels for the other two domains by ourselves. In the gesture domain, we define 64 different kinds of labels, *e.g.*, hand-waving, handshake, finger-pointing, touching-head, hugging, and so

Table 7: Performance comparison of face recognition on the VP dataset.(unit: %)

	P-only	PG	PS	PGS	PG+PS	PG+PS+PGS
K_d-Lin	18.53	24.60	34.50	–	**35.82**	–
K_d-S+	65.18	65.50	65.81	65.50	66.77	68.69
K_d-S±	65.18	65.81	66.45	66.13	67.41	**69.01**
K_d-R+	66.13	66.45	66.77	66.45	67.73	69.33
K_d-R±	66.13	67.09	67.41	67.41	68.37	**70.92**

on. In the scene domain, we define 35 kinds of labels like dark-background and national-flag. We use the same feature for peole domain as in Section 5.1. For the scene domain, we use a 374-dimensional attribute feature by adopting the VIREO-374 SVM models provided in [19, 20]. For gesture domain, we adopt 3-level spatial pyramids with densely sampled SIFT features encoded by the dictionary learned by K-means clustering to obtain a 1024-dimensional feature, which has been proposed for action recognition [9].

We compare our CGPC with Lin-CDL with the dense kernel JB-K (indicated with the prefix "K_d-" as in Section 5.1.3) for face recognition, gesture recognition and scene recognition across 3 domains on the VP dataset.

Face recognition Considering that there are few people-people co-occurrences in this dataset similar to the E-Album and the G-Album, we explore the following cross-domain models: People-Gesture relational model (PG), People-Scene relational model (PS), People-Gesture-Scene relational model (PGS), and the combined relational model PG+PS and PG+PS+PGS.

The results are summarized in Table 7. Not surprisingly, with the combined relational models PG+PS+PGS, both our S-CGPC and R-CPGC obtain the best recognition performance. Note that PG+PS+PGS achieves better recognition accuracy than PG+PS, which demonstrates that PGS is complementary to PG+PS and cannot be replaced with PG+PS directly.

Gesture and scene recognition We evaluate gesture recognition and scene recognition on the VP dataset. We evaluate the performances with the single relational model GS, PGS, and the combined relational model GS+PGS. "G-only" and "S-only" indicate the baselines without any relational models in the gesture recognition task and the scene recognition task, respectively.

We present the result of gesture recognition and scene recognition in Table 8, As expected, both the relational models GS and PGS can boost the gesture recognition accuracy and the scene recognition accuracy, and PGS still shows its complementary advantage over GS so that the combined relational model GS+PGS can reach the better performance for both the gesture and scene recognition task.

5.3. Experiments on the SUN 09 dataset

The SUN 09 dataset [5] is full of contextual information. It contains 12,000 annotated images covering a large number of indoor and outdoor scene categories with >200 object categories and 152,000 annotated object instances.

Table 8: Performance comparison of gesture (left) and scene recognition (right) on the VP dataset.(unit: %)

	G-only	GS	PGS	GS+PGS	S-only	GS	PGS	GS+PGS
K_d-Lin	13.42	30.35	–	–	20.45	46.01	–	–
K_d-S+	25.56	38.34	36.10	42.49	38.02	51.44	49.84	55.59
K_d-S±	25.56	41.21	39.29	**44.72**	38.02	54.31	51.12	**58.15**
K_d-R+	26.84	39.62	38.66	43.13	39.61	53.67	50.16	57.50
K_d-R±	26.84	43.13	41.85	**46.96**	39.61	57.19	53.04	**60.38**

To clarify, we emphasize multiple domains to fully explore domain specific feature representations for recognition. Analogously, our method can be applied to different classes as different domains, where we regard class specific feature as the domain specific feature representation. Those classes sharing the same feature can be clustered to the same domain. To verify such claims, we run experiments on the SUN 09 dataset by randomly dividing it into 3 domains of which each domain covers around 35 concepts (107 concepts used in total). Using the gist features as in HContext [5] for each domain, we achieve 41.4% correctness for top-3 presence prediction, while that of HContext is 38%.

6. Discussion and Conclusion

We propose a correlational Gaussian processes for cross-domain visual recognition with the relational models based on both the positive and negative co-occurrence statistics. Our proposed algorithm flexibly explores both the pairwise and high-order relational models. We evaluate the performance on each individual domain to demonstrate that our learnt relations can indeed improve the performance of each individual domain. It works well for all individual domains. Also, there is a trade-off between the runtime and the recognition performance. *i.e.*, if we incorporate more relations, then we can achieve the better performance with longer runtime, and vice versa.

As verified by the experiments, our proposed method achieves the best recognition accuracy compared to the state-of-the-art. Also, any concepts sharing the same feature representation can be viewed as a domain. Hence our model can be applied to the co-occurring concepts within a single domain and across multiple domains. Our future work includes further developing the inference/learning algorithms to make it more efficient and scalable [14], and extending our CGPC framework to deep learning model with large-scale cross-domain datasets we are still collecting.

Acknowledgement

This work is also partly supported by US NSF Grant IIS 1350763, China NSF Grant 61228303 and 61629301, GHs start-up funds form Stevens Institute of Technology, a Google Research Faculty Award, a gift grant from Microsoft Research, and a gift grant from NEC Labs America.

References

[1] M. A. Álvarez et al. Computationally efficient convolved multiple output gaussian processes. *JMRL*, 2011.

[2] P. Boyle et al. Dependent gaussian processes. In *NIPS*, 2005.

[3] G. Chen et al. Detection evolution with multi-order contextual co-occurrence. In *CVPR*, 2013.

[4] M. J. Choi et al. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010.

[5] M. J. Choi et al. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010.

[6] W. Chu et al. Relational learning with gaussian processes. *NIPS*, 2007.

[7] W.-S. Chu et al. Video co-summarization: Video summarization by visual co-occurrence. In *CVPR*, 2015.

[8] J. Cui et al. Easyalbum: an interactive photo annotation system based on face clustering and re-ranking. In *SIGCHI*, 2007.

[9] V. Delaitre et al. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC*, 2010.

[10] J. Diebold et al. Midrange geometric interactions for semantic segmentation. *IJCV*, 2016.

[11] L. Feng et al. Semantic concept co-occurrence patterns for image annotation and retrieval. *T-PAMI*, 2016.

[12] A. Gallagher. Clothing cosegmentation for recognizing people. In *CVPR*, 2008.

[13] E. Gavves et al. Active transfer learning with zero-shot priors: Reusing past datasets for future tasks. In *ICCV*, 2015.

[14] J. Hensman et al. Gaussian processes for big data. In *UAI*, 2013.

[15] D. Hernández-Lobato et al. Robust multi-class gaussian process classification. In *NIPS*, 2011.

[16] K. Hyun-Chul et al. Bayesian gaussian process classification with the em-ep algorithm. *T-PAMI*, 28(12):1948–1959, 2006.

[17] P. Isola et al. Learning visual groups from co-occurrences in space and time. In *ICLR-Workshop*, 2016.

[18] H. Jégou et al. Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening. In *ECCV*, 2012.

[19] Y. Jiang et al. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *CIVR*, 2007.

[20] Y. Jiang, J. Yang, C.-W. Ngo, and A. G. Hauptmann. Representations of keypoint-based semantic concept detection: A comprehensive study. *T-MM*, 2010.

[21] J. Joo et al. Visual persuasion: Inferring communicative intents of images. In *CVPR*, 2014.

[22] M. Kandemir. Asymmetric transfer learning with deep gaussian processes. In *ICML*, 2015.

[23] A. Kapoor. Learning discriminative models with incomplete data. MIT Thesis, 2006.

[24] K. Kersting et al. Learning preferences with hidden common cause relations. In *ECML PKDD*, 2009.

[25] S. Kim et al. Higher-order correlation clustering for image segmentation. In *NIPS*, 2011.

[26] S. Kim et al. Image segmentation using higher-order correlation clustering. *T-PAMI*, 2014.

[27] T. Kobayashi. Higher-order co-occurrence features based on discriminative co-clusters for image classification. In *BMVC*, 2012.

[28] P. Koniusz et al. Higher-order Occurrence Pooling on Mid- and Low-level Features: Visual Concept Detection. Technical report, 2013.

[29] L. Ladicky et al. Graph cut based inference with co-occurrence statistics. In *ECCV*, 2010.

[30] H. Li et al. Probabilistic elastic part model for unsupervised face detector adaptation. In *ICCV*, 2013.

[31] H. Li et al. Eigen-pep for video face recognition. In *ACCV*, 2014.

[32] W.-J. Li et al. Latent wishart processes for relational kernel learning. In *AISTATS*, 2009.

[33] D. Lin et al. Joint people, event, and location recognition in personal photo collections using cross-domain context. In *ECCV*, 2010.

[34] C. Long et al. Active visual recognition with expertise estimation in crowdsourcing. In *ICCV*, 2013.

[35] C. Long et al. A joint gaussian process model for active visual recognition with expertise estimation in crowdsourcing. *IJCV*, 2016.

[36] T. Malisiewicz et al. Beyond categories: The visual memex model for reasoning about object relationships. In *NIPS*, 2009.

[37] T. Mensink et al. Tree-structured crf models for interactive image labeling. *T-PAMI*, 2013.

[38] T. Mensink et al. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, 2014.

[39] M. Nickel et al. A three-way model for collective learning on multi-relational data. In *ICML*, 2011.

[40] M. A. Sadeghi et al. Recognition using visual phrases. In *CVPR*, 2011.

[41] R. Silva et al. Hidden common cause relations in relational learning. In *NIPS*, 2007.

[42] X. Song et al. Category co-occurrence modeling for large scale scene recognition. *Pattern Recognition*, 2016.

[43] Z. Xu et al. Multi–relational learning with gaussian processes. In *IJCAI*, 2009.

[44] Z. Xu et al. Multi-task learning with task relations. In *ICDM*, 2011.

[45] Z. Xu et al. Infinite tucker decomposition: Nonparametric bayesian models for multiway data analysis. In *ICML*, 2012.

[46] Y. Yoo et al. Visual path prediction in complex scenes with crowded moving objects. In *CVPR*, June 2016.

[47] K. Yu et al. Stochastic relational models for discriminative link prediction. In *NIPS*, 2007.

[48] K. Yu et al. Gaussian process models for link analysis and transfer learning. In *NIPS*, 2008.

[49] L. Zhang et al. High order co-occurrence of visual words for action recognition. In *ICIP*, 2012.