Correlational Gaussian Processes for Cross-domain Visual Recognition

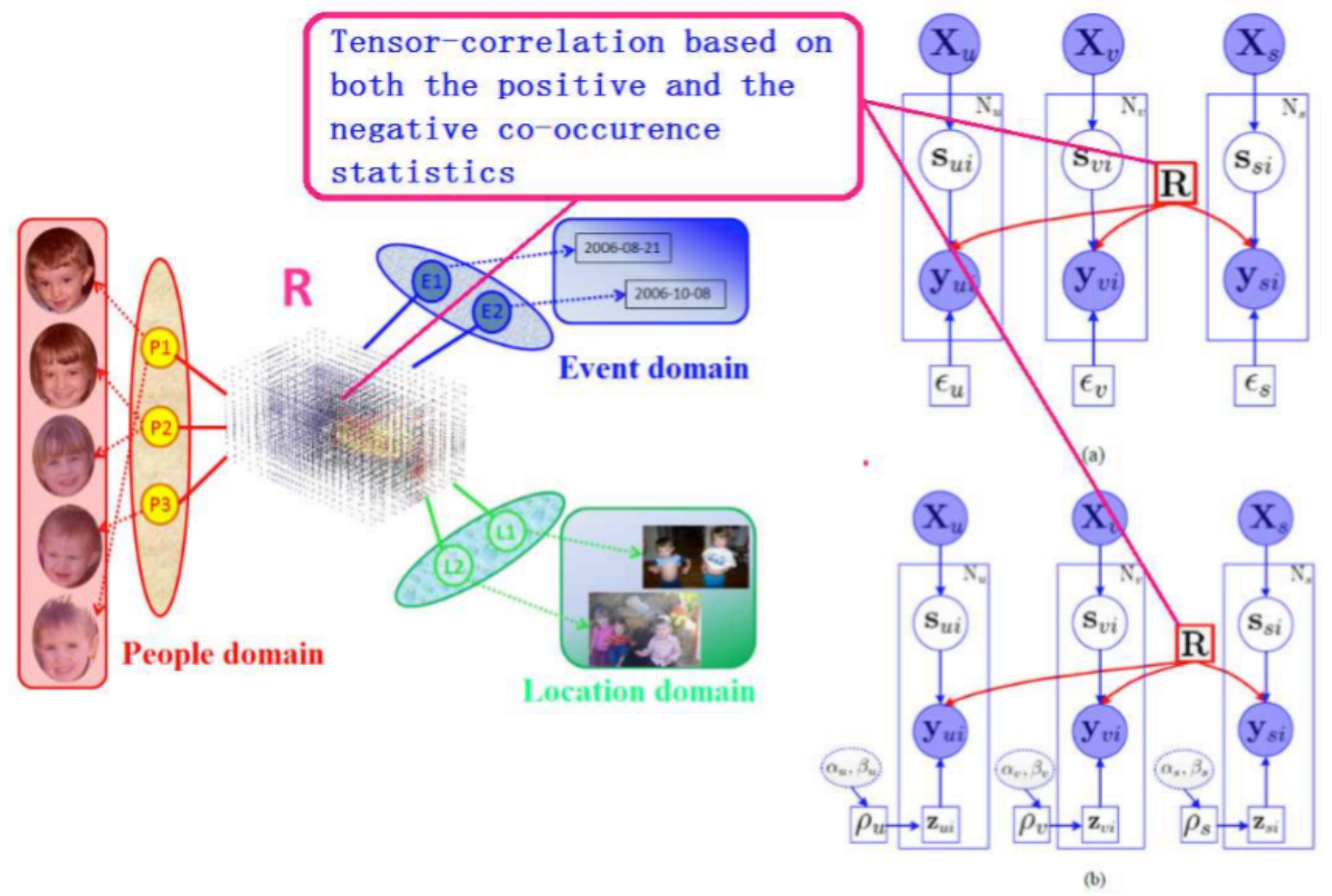
Chengjiang Long
Kitware Inc.
chengjiang.long@kitware.com

Gang Hua
Microsoft Research
ganghua@microsoft.com

July 21-26 HONOLULU 2017

Introduction

- **Observation:** Multiple visual recognition problems in different semantic domains can be simultaneously solved through a joint formulation instead of being handled independently.
- **Intuition:** 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.


Tensor-correlation based on both the positive and the negative co-occurence statistics
R
P1
P2
P3
People domain
E1
E2
2006-08-21
2006-10-08
Event domain
L1
L2
Location domain
(a)
(b)

Competing Algorithms

[Lin] D. Lin et al. Joint people, event, and location recognition in personal photo collections using cross-domain context. In ECCV, 2010.
[Hcontex] M. J. Choi et al. Exploiting hierarchical context on a large database of object categories. In CVPR, 2010.

Sponsors

NSF, Google Research, Microsoft Research, NEC, STEVENS INSTITUTE of TECHNOLOGY

Joint probability: $P(\mathbf{S},\mathbf{Y},\mathbf{R}|\mathbf{X}) \propto p(\mathbf{R})\, p(\mathbf{Y}|\mathbf{R},\mathbf{S},\Theta) \prod_{d\in\Omega} p(\mathbf{S}_d|\mathbf{X}_d)$

$$p(\mathbf{Y}|\mathbf{R},\mathbf{S},\Theta) \approx p(\mathbf{Y}|\mathbf{R})\,p(\mathbf{Y}|\mathbf{S},\Theta)$$

Joint probability: $P(\mathbf{S},\mathbf{Y},\mathbf{R}|\mathbf{X}) \approx p(\mathbf{R})\, p(\mathbf{Y}|\mathbf{R}) \prod_{d\in\Omega} p(\mathbf{S}_d,\mathbf{Y}_d|\mathbf{X}_d,\Theta_d)$

Relational model prior

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

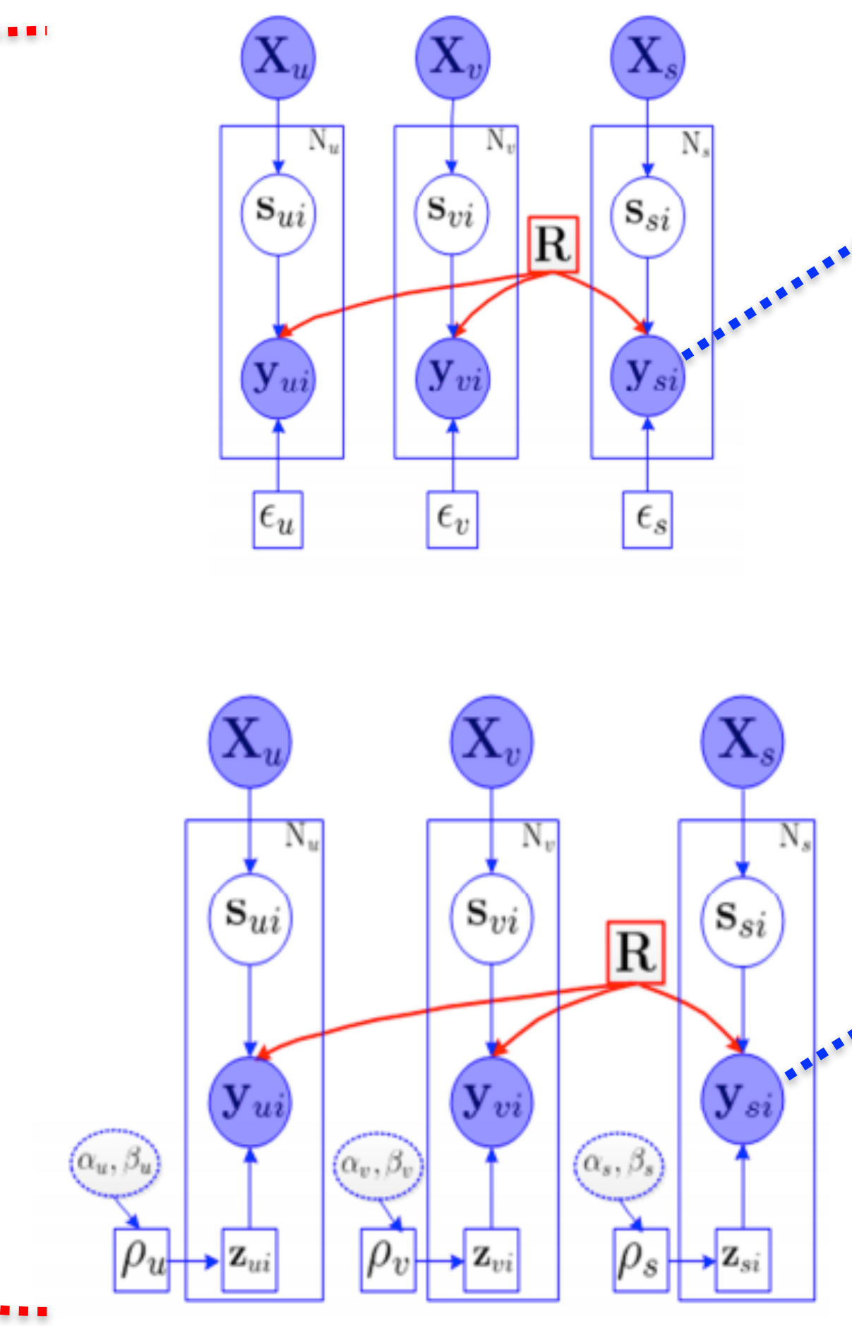
Co-occurrence relational models

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

$$\Phi(\mathcal{Y}^{\mathbf{c}}_{\mathbf{j}}|\mathbf{R}^{\mathbf{c}}) \doteq \sum_{y_{d_1}\sim\ldots\sim y_{d_{|\mathbf{c}|}}} \mathbf{R}^{\mathbf{c}}(y_{d_1},\ldots,y_{d_{|\mathbf{c}|}}) \times I(y_{d_1}=y_{d_1j_1})\ldots I(y_{d_{|\mathbf{c}|}}=y_{d_{|\mathbf{c}|}j_{|\mathbf{c}|}})$$

$$\mathbf{R}^{\mathbf{c}} = w^{\mathbf{c}}_{+}\mathbf{R}^{\mathbf{c}}_{+} + w^{\mathbf{c}}_{-}\mathbf{R}^{\mathbf{c}}_{-} - \sum_{\mathbf{c}_1,\mathbf{c}_2} w^{\mathbf{c}_1\mathbf{c}_2}_{+-}\mathbf{R}^{\mathbf{c}_1\mathbf{c}_2}_{+-}$$

$$w^{\mathbf{c}}_{+} = 1,\ w^{\mathbf{c}}_{-} = \frac{1}{\prod_{d_k\in\mathbf{c}}(l_{d_k}-1)} \text{ and } w^{\mathbf{c}_1\mathbf{c}_2}_{+-} = \frac{1}{\prod_{d_k\in\mathbf{c}_2}(l_{d_k}-1)}$$



SMGPC

$$\prod_k p(\mathbf{S}^k_d|\mathbf{X}_d)\prod_{i=1}^{N_d} p(y^k_{di}|s^k_{di},\varepsilon_d)$$

$$p(y^k_{di}|s^k_{di},\varepsilon_d) = \varepsilon_d H(y^k_{di}s^k_{di}) + (1-\varepsilon_d)H(-y^k_{di}s^k_{di}),$$

RMGPC

$$p(\mathbf{S}_d|\mathbf{X}_d)p(\rho_d)p(\mathbf{z}_d|\rho_d)\prod_{i=1}^{N_d} p(y_{di}|\mathbf{s}_{di},z_{di}),$$

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

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

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

Inference and learning:

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

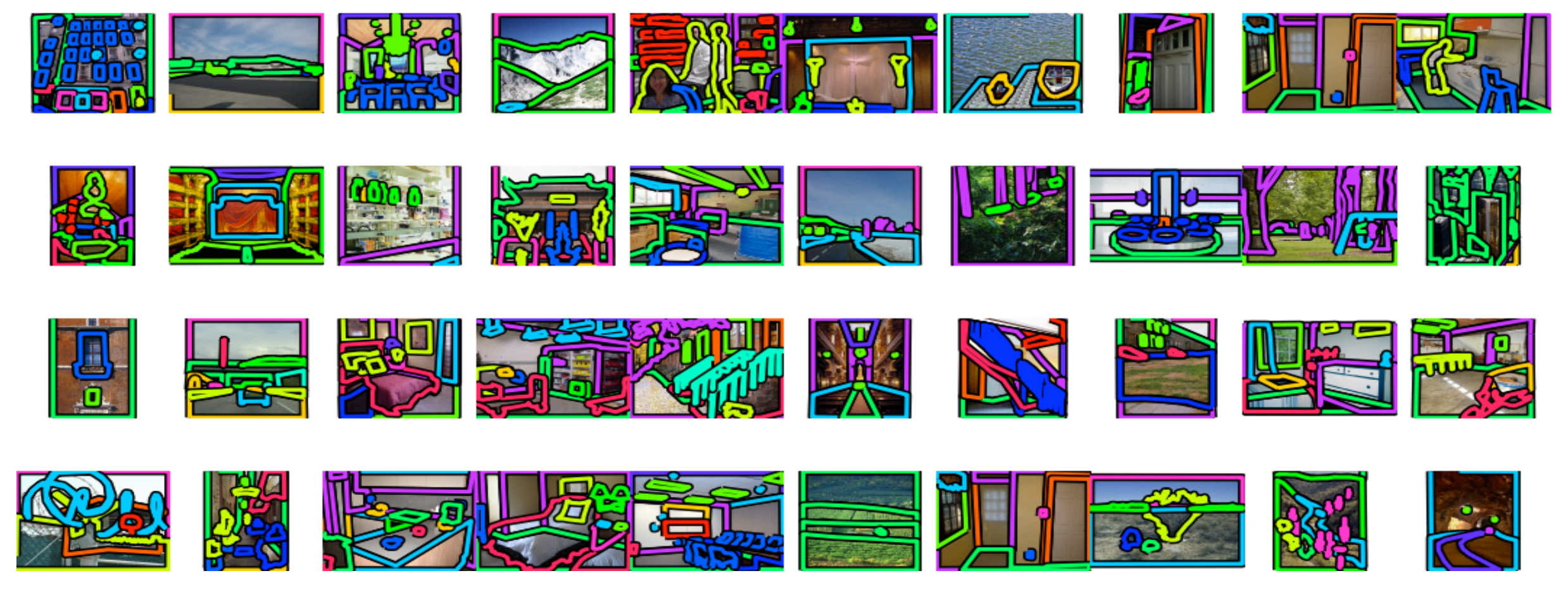
- **E-step**: Infer the distribution of $\mathbf{Y}_U$ based on both the extracted features and the current relational model $\hat{\mathbf{R}}^{(t)}$ by $\hat{q}^{(t+1)} \leftarrow \arg\max_q J(\hat{\mathbf{R}}^{(t)},q)$.
- **M-step**: Estimate and update the relational model using the labels provided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

Dataset	images	domain description
E-Album	108	people(15 in 145 faces), location(21), event(21)
G-Album	312	people(13 in 441 faces), location(141), event(117)
VP	1124	people(8), gesture(64), scene(35)
SUN 09	12,000	107 concepts into 3 domains (35 concepts for each)

E-Album

G-Album

VP Dataset


SUN 09 Dataset

Experiments on E-Alum and G-Album

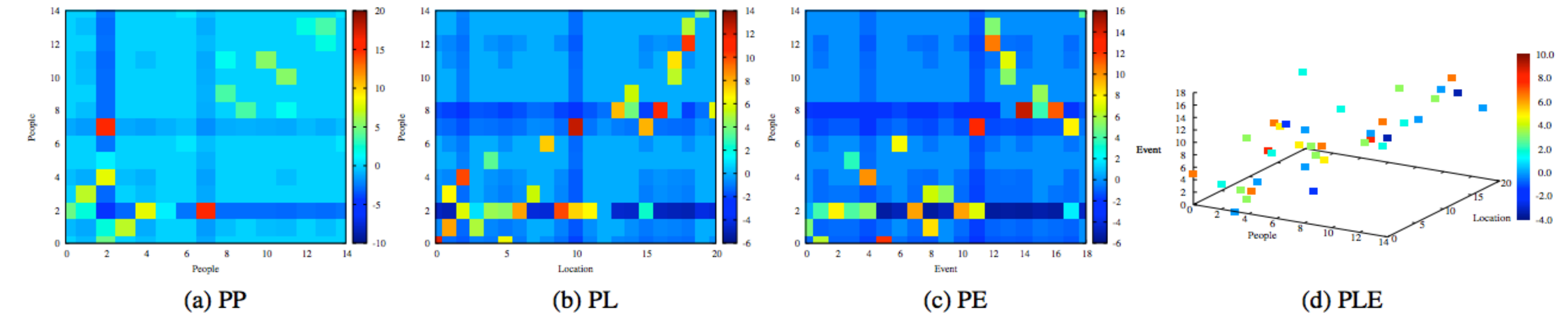
Table 1: Face recognition performance with 4 relational models and 6 kernels on the E-Album.(unit: %)

	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**

Visualization of relational models



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

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**

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**

Experiments on VP Dataset

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**

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**

Experiments on SUN 09 Dataset

We achieve 41.4% correctness for top-3 presence prediction, while that of HContext is 38%.

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. It works well for visual recognition tasks in all individual domains.