Annotation Propagation in Large Image Databases via Dense Image Correspondence

Michael Rubinstein   Ce Liu   William T. Freeman

Problem and Contributions

Input
Sparsely tagged and (even not) labeled image database

Output
Fully tagged and labeled image database

* actual results

Prior Work	This Work
➤ **Large training set** of densely labeled images	➤ **Weakly supervised** setup: use both image tags (cheaper to obtain) and labeled pixels (expensive to obtain)
➤ Each image classified **independently**	➤ Solve **jointly** for all pixels and classes
	➤ Use **dense image correspondences** to resolve visual ambiguities

Harder to disambiguate *sea* and *sky*

Easier to disambiguate *sea* and *sky*

Appearance only | With spatial regularization | With spatial regularization and correspondences

Formulation

Input:
Images $I = \{I_1, ..., I_N\}$, vocabulary $V = \{l_1, ..., l_L\}$
Partial annotations: I_t, T_t - tagged images + their tags, I_l, C_l - labeled images + their labels

Output:
Labeling $C = \{c_1, ..., c_N\}$, s.t. for pixel $p = (x, y) \in \lambda_i$ $c_i(p) \in \{1, ..., L, \emptyset\}$
Tags $T = \{t_1, ..., t_N, t_i \in \{1, ..., L\}\}$ defined directly as the set union of pixel labels

Assumptions:
a) Each word has distinct visual appearance
b) Neighboring pixels in an image tend to have the same annotation
c) Similar visual patterns across different images should have similar annotation
d) Tags and tag co-occurrences which are more frequent in the database are also more probable

$$E(C) = \sum_{i=1}^{N} \sum_{p \in \Lambda_i} \left[\Phi_p(c_i(p)) + \sum_{q \in N_p} \Psi_{int}(c_i(p), c_i(q)) + \sum_{j \in N_i} \Psi_{ext}\left(c_i(p), c_j\left(p + w_{ij}(p)\right)\right) \right]$$

Local evidence | Intra-image regularization | Inter-image regularization

w_{ij} - dense correspondence $I_i \rightarrow I_j$

Image graph for LabelMe Outdoors (LMO)

Algorithm

Text-to-image Correspondence (Learning)

Appearance model:
$$P(D_i(p); \Theta) = \sum_{l=1}^{L} \left(\rho_l \sum_{k=1}^{M} \pi_{l,k} N(D_i(p); \mu_{l,k}, \Sigma_{l,k}) \right) + \rho_\epsilon N(D_i(p); \mu_\epsilon, \Sigma_\epsilon)$$

Posterior:
$$P_a(c_i(p) = l; D_i(p), \Theta) = \frac{\rho_l \sum_{k=1}^{M} \pi_{l,k} N(D_i(p); \mu_{l,k}, \Sigma_{l,k})}{\sum_{j=1}^{L}(\rho_j \sum_{k=1}^{M} \pi_{j,k} N(D_i(p); \mu_{j,k}, \Sigma_{j,k})) + \rho_\epsilon N(D_i(p); \mu_\epsilon, \Sigma_\epsilon)}$$

grass | mountain | sky | tree | MAP appearance

LMO dataset – 2688 images, 33 words, **50% images tagged, 1% labeled**

Propagation (Inference)

Local evidence (data term):
$$\Phi_p(c_i(p) = l) = -\log P_a(c_i(p); D_i(p), \Theta) - \log P_t^i(l) - \lambda_s \log P_s(c_i(p)) - \lambda_c \log P_c^i(c_i(p))$$

Appearance | *Bias towards frequent tags and tag co-occurences* | *Spatial prior* | *Color prior*

$$P_t^i(l) = \frac{\beta}{|N_i|} \delta[l \in t_i] + \frac{1-\beta}{Z} \sum_{j \in N_i} \sum_{m \in t_j} h^o(l, m)$$

$\max_{c_i} P_a(c_i(p))$ | $\max_{c_i}\{P_a(c_i(p) = l) + P_t^i(l)\}$

Intra-image regularization: $p, q \in \Lambda_i,$ $q \in N_p$
$$\Psi_{int}(c_i(p) = l_p, c_j(q) = l_q) = \delta[l_p \neq l_q] \lambda_{int} \exp(-\|I_i(p) - I_i(q)\|)$$

Inter-image regularization: $I_j \in N_i,$ $p \in \Lambda_i,$ $r = p + w_{ij}(p)$
$$\Psi_{ext}(c_i(p) = l_p, c_j(r) = l_r) = \delta[l_p \neq l_r] \lambda_{ext} \exp(-|S_i(p) - S_j(r)|)$$

Input image | Flow color coding | Local evidence | MAP appearance

Neighbors | Correspondence | Neighbors warped | Neighbors local evidence warped | MAP appearance + intra-image + inter-image reg. (final result)

MAP appearance + intra-image reg.

Results

SUN (9556 images, 522 words) and LMO (2688 images, 33 words) - 50% images tagged, 5% images labeled

	Labeling		Tagging			
	r	$\bar{r}$	P	$\bar{P}$	R	$\bar{R}$
Makadia [1]	—	—	53.87	30.23	60.82	25.51
STF [4]	52.83	24.9	34.21	30.56	**73.83**	**58.67**
LT [5]	53.1	24.6	41.07	35.5	44.3	19.33
AP (ours)	**63.29**	**29.52**	**55.26**	**38.8**	59.09	22.62
AP-RF	56.17	26.1	48.9	36.43	60.22	24.34
AP-NN	57.62	26.45	47.5	35.34	59.83	24.01

Results on LMO

	Labeling		Tagging			
	r	$\bar{r}$	P	$\bar{P}$	R	$\bar{R}$
Makadia [1]	—	—	26.67	11.33	39.5	14.32
STF [4]	20.52	9.18	11.2	5.81	**62.04**	**16.13**
AP (ours)	**33.29**	**19.21**	**32.25**	**14.1**	47	13.74

Results on SUN

Comparison with Semantic Texton Forests [Shotton et al. 2008]

Input | Ours | STF | Input | Ours | STF | Input | Ours | STF

ESP (21,846 images, 269 words) and IAPR (19,805 images, 291 words) - 90% images tagged, 0% labeled

	ESP				IAPR			
	P	$\bar{P}$	R	$\bar{R}$	P	$\bar{P}$	R	$\bar{R}$
Makadia [1]	22	**25**	—	—	**28**	**29**	—	—
AP (Ours)	**24.17**	23.64	20.28	13.78	27.89	25.63	19.89	12.23

Failure cases