designing and learning visual representations

stefano soatto

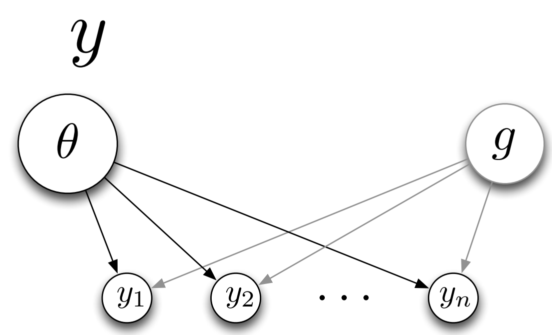
ucla

november 13, 2015

- visual representation: definition
  - tradeoffs
- special (trivial) case: local descriptors
  - the unreasonable effectiveness of sift
- beyond local: modeling intrinsic variability
  - conjectures
- embedding the representation in the scene
  - scene topology, gravity

representation

- a function of the data that is useful for a task ... regardless of nuisance factors affecting (future) data
  - data $x^t = \{x_1, \dots, x_t\}$
    - images
  - task θ
    - decision or control actions on the scene portrayed by the images
  - nuisance factors g
    - viewpoint, illumination, partial occlusion, sensor characteristics
  - useful
    - "informative"



optimal representation

- ‘most informative’ function of the data, for a task: sufficient statistic
- ‘most compressed’: minimal sufficient statistic
- ‘most insensitive’ to nuisance factors affecting future data: minimal sufficient invariant statistic

(fake) bad news

- ‘invariance’ cannot be attained, so settle for ‘approximate invariance’
- invariance trades off information (discriminative power)

a few facts

- The likelihood function $L(\theta) \doteq p_\theta(x)$ is a minimal sufficient statistic, even if θ is infinite-dimensional (Bahadur, 1954).
- Nuisance variability can be marginalized: $p_\theta(x|G) \doteq \int_G p_\theta(x|g)dP(g)$

  but invariant only if marginalization is wrt the base measure and in general *not maximal.*
- Profiling (max-out): $p_{\theta,G}(x) \doteq \sup_{g \in G} p_{\theta,g}(x)$

  yields *a maximal invariant*, but (non-convex) search at test time.
- (Down)-sampling the profile likelihood introduces aliasing phenomena (extrema that do not exist before downsampling/reconstruction)
- Anti-aliasing = pooling = local marginalization

$$\hat{p}_{\theta,G}(y) = \max_i \hat{p}_{\theta,g_i}(y) = \max_i \int_G p_{\theta,g_i}(gy)w(g)d\mu(g)$$

the sample-orbit antialiased likelihood is a minimal sufficient invariant statistic: optimal representation

simple example

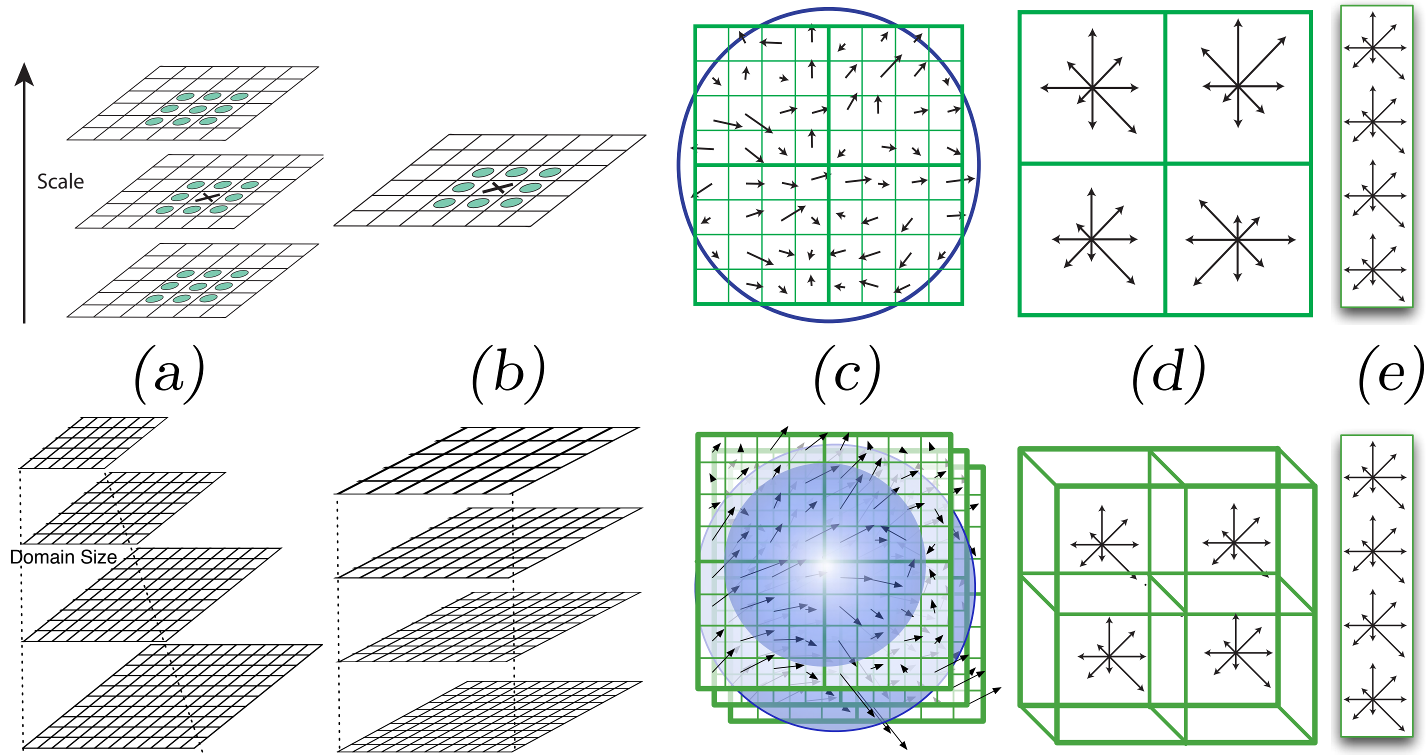
- training set: a single image
- task: binary classification (correspondence)
- nuisances: planar similarities
- optimal representation (closed form):

$$\boxed{p_x(y|\mathcal{H}) \doteq p(\angle\nabla y|\nabla x) = \frac{1}{\sqrt{2\pi\epsilon^2}} \exp\left(-\frac{1}{2\epsilon^2}\sin^2(\angle\nabla y - \angle\nabla x)\|\nabla x\|^2\right) M}$$

$$M = \frac{\epsilon e^{-\frac{(m)^2}{2\epsilon^2}}}{\sqrt{2\pi}} + m - m\Psi\left(-\frac{m}{\epsilon}\right)$$

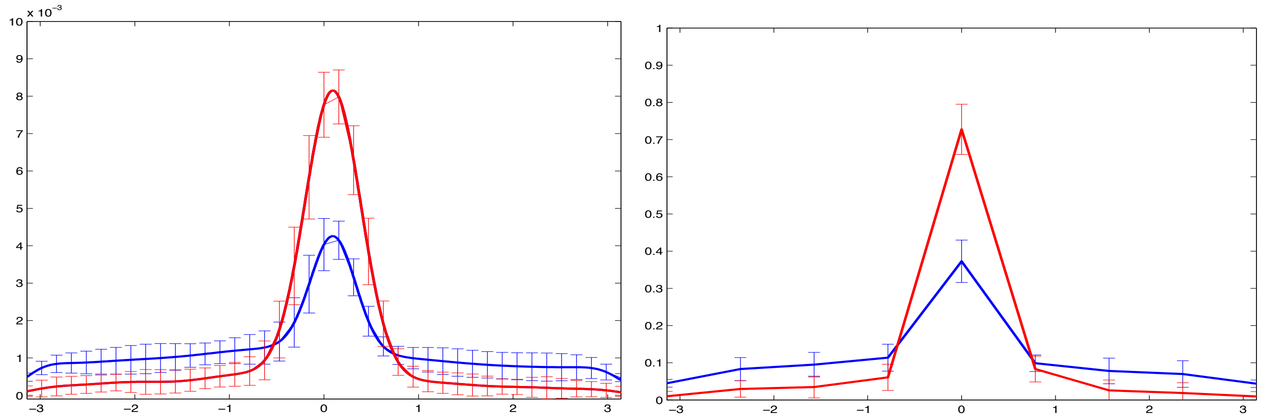
SIFT revisited

$$h_{\mathrm{SIFT}}(\theta|I) = \int \kappa_\epsilon\left(\theta - \angle\nabla I(y)\right)\kappa_\sigma(y-x)\|\nabla I(y)\|dy$$

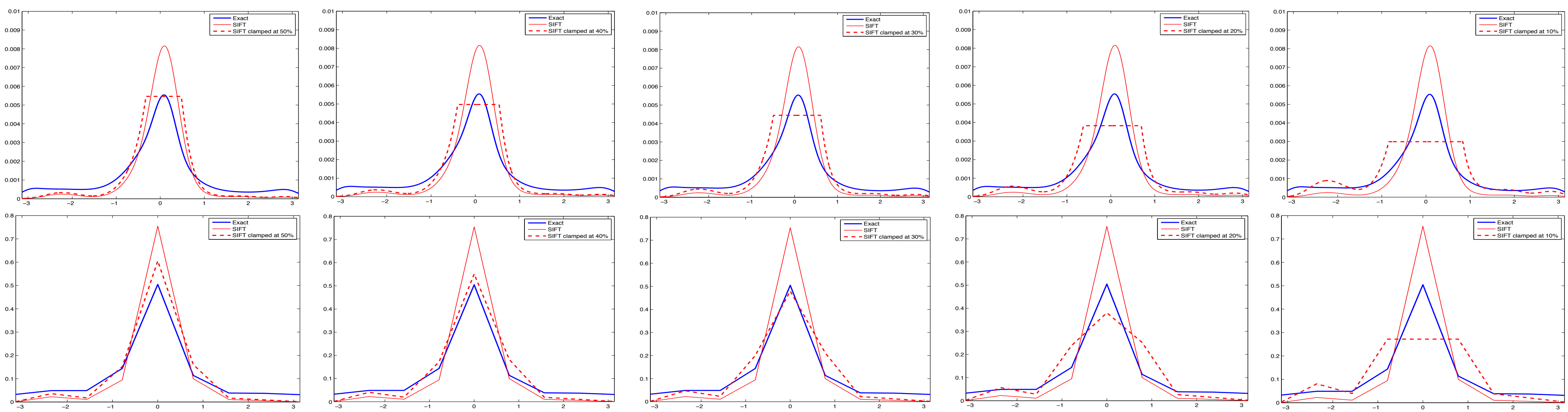


$$h_{DSP}(\theta|I) = \iint \kappa_\epsilon\left(\theta - \angle\nabla I(y)\right)\kappa_\sigma(y-x)\|\nabla I(y)\|dP(\sigma)dy$$

w/ j. dong



CLAMPING


Exact
SIFT
SIFT clamped at 50%
SIFT clamped at 40%
SIFT clamped at 30%
SIFT clamped at 20%
SIFT clamped at 10%

multi-view descriptors

- MV-HOG $\phi^t_{\mathbf{z}}(\theta|\mathbf{y}^t) \doteq \frac{1}{\tau}\sum_{\tau=1}^{t}\int \mathcal{N}_{\mathbb{S}^1}(\theta - \angle g\mathbf{y}(\tau))\|\nabla \mathbf{y}\|dP(g)$
- R-HOG $\phi^t_{\hat{\mathbf{z}}}(\theta) \doteq \int_{SO(3)\times\mathbb{R}} \mathcal{N}_{\mathbb{S}^1}(\theta - \angle g\hat{\mathbf{z}})dP_{SO(3)}(g)\|\nabla \hat{\mathbf{z}}\|d\mu$

w/ j. dong, j. hernandez, d. davis, j. balzer

where are we?

- structure in the representation?
- topology? geometry?

more realistically

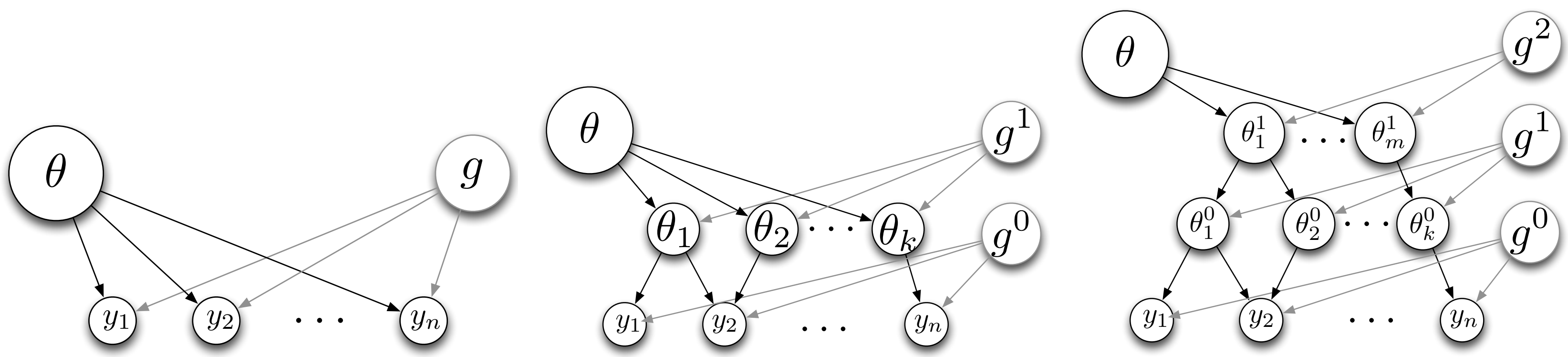
- intra-class variability (separation principle)

$$p_\theta(y|k) = \int p_\theta(y|x_k, g_k)dP(g_k|k) \doteq \int p_{\theta_k, g_k}(y)dP(g_k|k)$$

- occlusion (combinatorics)

$$\begin{aligned}
\hat{p}_{\theta,G,\hat{V}}(y|k) &= \max_{i,V\in\mathcal{P}(D)} \int_{\mathrm{diff}(D)} \prod_{j\in V} \hat{p}_{\theta_k,g_i}(y_j|x_k,g_k)dP(g_k|k) \\
&\simeq \max_{i,V} \int_{G^M} \prod_{j\in V} \hat{p}_{\theta_k,g_k g_i}(y_{|_{g_j\mathcal{B}_0}})dP(g_1^{-1}g_k,\ldots,g_M^{-1}g_k|k) \\
&= \max_{i,V} \int_{G^M} \prod_{j\in V} \hat{p}_{\theta_k,g_k g_i}(g_j y_{|\mathcal{B}_0})dP(g_1^{-1}g_k,\ldots,g_M^{-1}g_k|k) \\
&= \max_{V} \int_{G^M} \prod_{j\in V} \hat{p}_{\theta_k,g_{k_j}}(y)dP_G(g_{k_1},\ldots,g_{k_M}|k)
\end{aligned}$$

- architecture?



still missing

- scene topology (detachable objects)
- global referencing (gravity)
- extension to tasks other than detection

summary

- definition, analytical characterization of an ideal visual representation:
- simple case where inference is tractable
- conjectures on extensions to include intra-class variability: relation to cnn's
- representation of the scene, not the image
- extension to more general (control, decision) tasks
- support "query system" on the scene