Dense correspondence prediction in computer vision

Mikhail Shvets

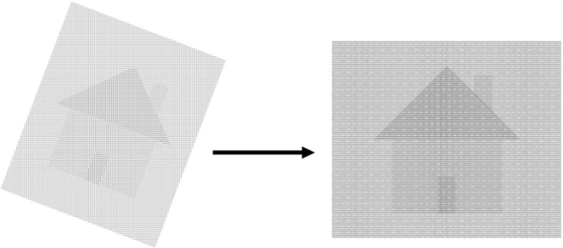
November 6, 2016

Applications

- Structure from motion
- Optical flow, scene flow
- Object detection and tracking
- Scene understanding

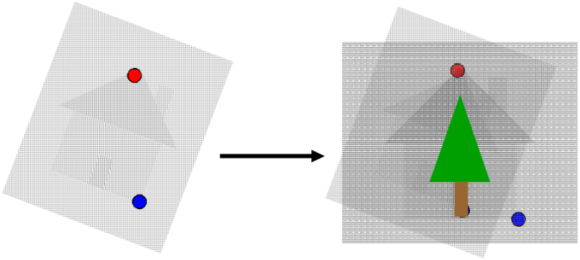
Interest points

First steps



- Brute force:
  - model selection (translation, rotation), parametrization
  - quality function selection (correlation)
- Pyramides (Laplacian, Gaussian)

Interest points



- Find interest points
  - repeatability
  - saliency
  - locality
- Find transformation that matches these points

Harris detector

$$E(u, v) = \sum_{x,y} w(x, y)[I(x + u, y + v) - I(x, y)]^2$$

For small u, v: $E(u, v) \approx \begin{bmatrix} u & v \end{bmatrix} M \begin{bmatrix} u \\ v \end{bmatrix}$, where

$M = \sum_{x,y} w(x, y) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}$ – matrix with characteristic values λ_1, λ_2.

$M = R^{-1} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} R$ as M is symmetric.

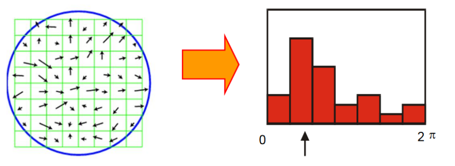
- λ_1, λ_2 are small – monotonic area
- $\lambda_1 \ll \lambda_2$ – horizontal edge
- $\lambda_1 \gg \lambda_2$ – vertical edge
- $\lambda_1 \sim \lambda_2$ – edge

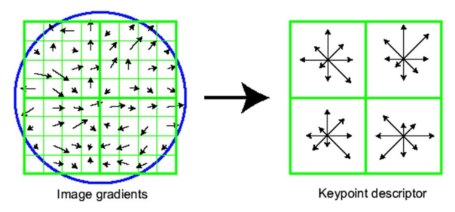
$$F = \det M - k(traceM)^2$$

Descriptors

Build feature vector for each interest point

Scale-Invariant Feature Transform (SIFT)





SIFT Flow[1]

Objective

- Match SIFT descriptors along flow vectors
- Smooth flow field
- Discontinuities agreeing with object boundaries

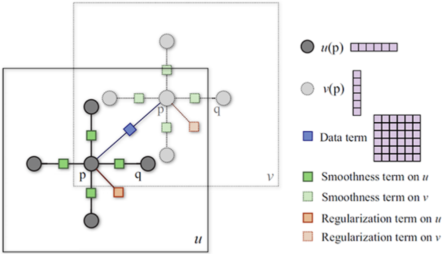
Let $p = (x, y)$ – grid coordinate and $w(p) = (u(p), v(p)) \in \mathbb{Z}^2$ – flow vector.

s_1, s_2 – SIFT images ($h \times w \times 128$).

$$E(w) = \sum_{p} \min\left(\|s_1(p) - s_2(p + w(p))\|_1, t\right) + \sum_{p} \eta\left(\|u(p)\| + \|v(p)\|\right) +$$

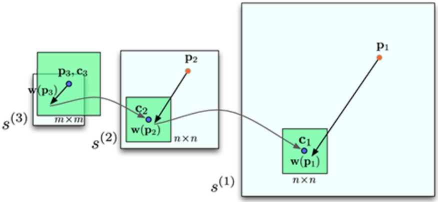
$$+ \sum_{p,q \in \epsilon} \min\left(\alpha\|u(p) - u(q)\|, d\right) + \sum_{p,q \in \epsilon} \min\left(\alpha\|v(p) - v(q)\|, d\right)$$

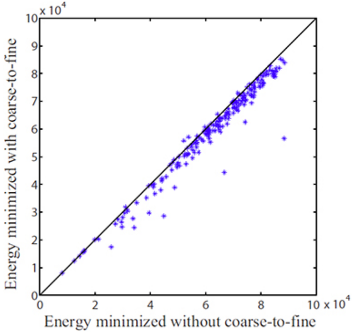
[1] Ce Liu, Jenny Yuen, and Antonio Torralba. "Sift flow: Dense correspondence across scenes and its applications". In: *IEEE transactions on pattern analysis and machine intelligence* 33.5 (2011), pp. 978–994.



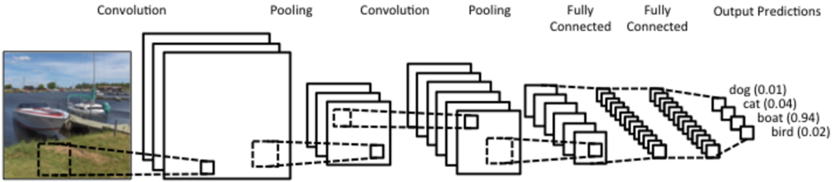
Inference method: loopy belief propagation.
Note: in the objective pairwise terms u and v are decoupled, which enables efficient inference, still $\mathcal{O}((HW)^2)$.

Coarse to fine approach





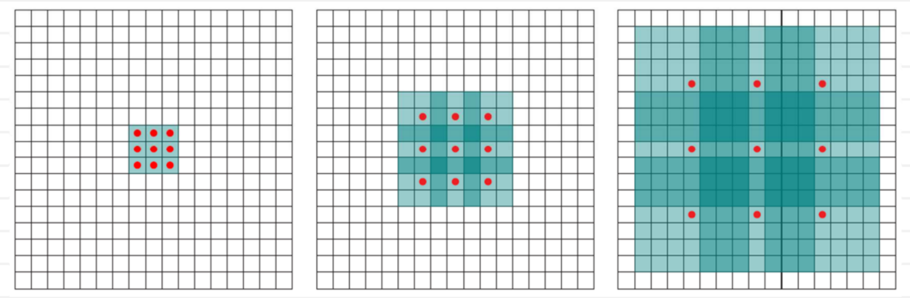
Convolutional neural networks



L-layer CNN: $\langle \mathcal{I}, \mathcal{W}, * \rangle$, where $\mathcal{I} = \{I_l\}_{l=1}^{L}$, $\mathcal{W} = \{W_l\}_{l=1}^{L}$
$W \in \mathbb{R}^{c \times w \times h}$, $I \in \mathbb{R}^{c \times W \times H}$ and $w \ll W$, $h \ll H$.

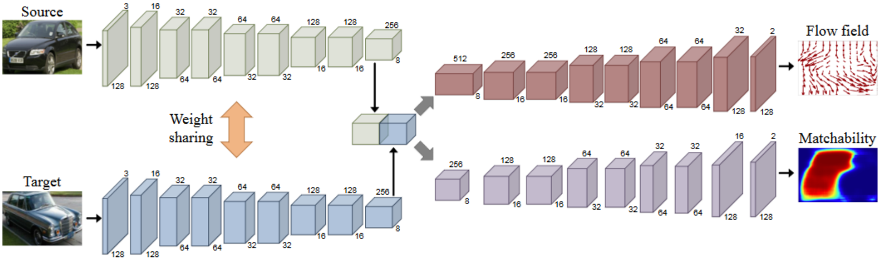
$$V(x, y, t) = \sum_{i=x-\delta}^{x+\delta} \sum_{j=y-\delta}^{y+\delta} \sum_{s=1}^{S} W(i - x + \delta, j - y + \delta, s, t) I(i, j, s)$$

Dilated convolutions[2]



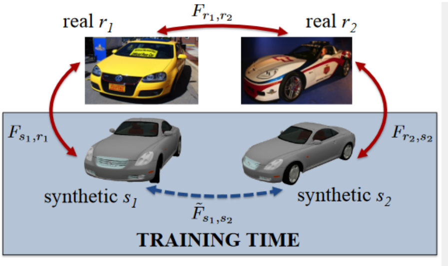
[2] Fisher Yu and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions". In: *arXiv preprint arXiv:1511.07122* (2015).

Predict flow and matchability



$$L_{flow} = \sum_{p:M(p)=1} \min(\|\hat{F}(p) - F(p)\|_2^2, T^2)$$

Cycle consistency[3]



$$L = L(F_{s_1 s_2}, \hat{F}_{s_1 r_1} \circ \hat{F}_{r_1 r_2} \circ \hat{F}_{r_2 s_2})$$

where $\circ$ operation is defined as

$$\hat{F}_{a,b}(p) \circ \hat{F}_{b,c}(p) = \hat{F}_{a,b}(p) + \hat{F}_{b,c}(p + \hat{F}_{a,b}(p))$$

[3] Tinghui Zhou et al. “Learning Dense Correspondence via 3D-guided Cycle Consistency”. In: *arXiv preprint arXiv:1604.05383* (2016).

Consistency

Cycle Consistency results

Cycle Consistency results

Cycle Consistency results

Summary

- Standard pipeline:
  - Detect interest points
  - Extract features
  - Match features
- Efficient dense matching: inference on graphical models
- Neural Networks: straightforward prediction
- Little supervision: cycle consistency