Pose Estimation from Local Features

David Demirdjian[1] Raquel Urtasun[2] Christopher Wilkens[2]

[1] *Toyota Technical Center, Cambridge, USA*

[2] *Massachusetts Institute of Technology, Cambridge, USA*

Introduction

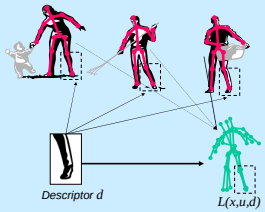
Our goal is to perform pose estimation from single images. Our main contribution is a generalization of the Implicit Shape Model – introduced earlier for object recognition by Leibe & Schiele – to learn the relationship between articulated pose and the appearance of reference patches.

Recovering accurate pose from single images requires a more sophisticated model than the Implicit Shape Model. In our approach, flexibility in the model is obtained by introducing hyper-parameters that are learned from training data and optimally weigh the spatial and pose contribution of patches. In our framework, the appearance of a patch is modeled using multiple descriptors. A task-specific metric in the patch space is learned for best performance on pose estimation.

Pose Estimation from Local Features

Individual local features provide rich and useful information on the global pose.
e.g.

- Location of 'landmarks' in the neighborhood of local features
- Scale, configuration, …



Approach

We employ a technique similar to the Implicit Shape Model introduced

- ***Database construction***. Learn a set of local patches (and associated descriptors) and pose x *relative* to their location in the image.

- ***Vocabulary tree.*** Build a vocabulary tree using k-means clustering using the *Pyramid Match Toolkit* by J. Lee et al.

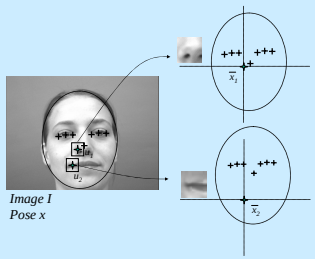
- ***Pose estimation.*** For each local feature of a query image, retrieve its kNNs in the database and compute the density by kernel-based estimation. Pose is found by maximizing the density (voting)

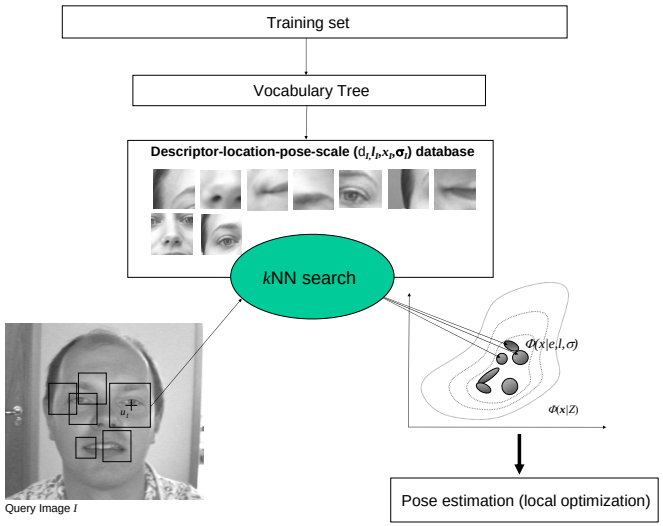
Benefits of our approach:

Robustness to occlusions
Does not require segmentation or normalization
Fast

Relative/absolute pose

In order to incorporate geometric relationships, the relative pose x associated with a local feature I is given with respect to the feature location l_I and scale σ_I. At testing, the absolute pose is *reconstructed* from the feature location l and scale σ.





Density estimation

Given a query image I, descriptors d are estimated at multiple locations l and scales σ and matched against the database. The probability of observing pose x given image Z is:

$$\phi(x) = p(x \mid Z) = \sum_j p(x \mid e_j, l_j, \sigma_j)$$

Density associated with a single local feature

$$p(x \mid e, l, \sigma) = \sum_i p(x \mid I_i, l, \sigma)\, p(I_i \mid e)$$

Implicit Shape Model

The probability of observing pose x given a patch I at location l and scale σ is modeled using a Gaussian model. The model gives more confidence to pose 'parts' m_k closer to patch locations l

$$p(x \mid I, l, \sigma) = \mathrm{N}(x \mid F(x_I, l, \sigma), \Sigma_I)$$

Warping function $F(x_I, l, \sigma) = \frac{\sigma}{\sigma_I}(x_I - L_I) + L \qquad L = (l,...,l)^T \qquad L_I = (l_I,...,l_I)^T$

$$\Sigma_I = \frac{1}{s^2} diag(\Lambda_1,...,\Lambda_N) \qquad \Lambda_k = (m_k - l_k)(m_k - l_k)^T$$

s is the covariance scale. It corresponds to the patch reliability and is learned from training data

Task-Oriented Patch Metric

The patch likelihood $p(I_i|e)$ is defined as a linear combination of kernels

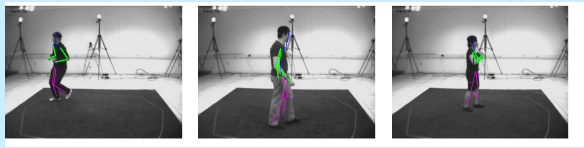
$$p(I_i \mid e) = d^T K(I_i \mid e) \quad d = (d_1,...d_N) \quad with \quad |d| = 1 \qquad K = (K_1,...,K_N)$$

$$\forall e, \sum_j K_j(I_i \mid e)$$

Experiments

Articulated Body Pose

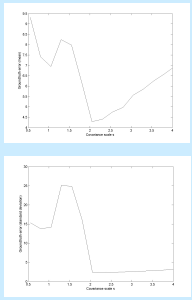
Evaluation on the HumanEva database [L. Sigal and M. J. Black.]. We used a subset of the database containing sequences of *Walking*, *Jogging* and *Boxing*. Sequences were evenly divided into training and testing sets

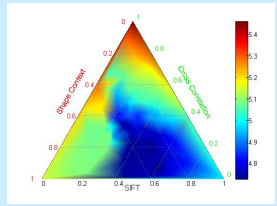


	SIFT	*SC*	*CC*	***Multi-D***
Leibe et al.	11.32 (7.58)	12.36 (5.10)	13.52 (6.29)	
ISM-pose	5.01 (3.58)	5.13 (3.62)	5.47 (4.21)	4.73 (2.58)

	ISM-pose
Walking	4.16 (2.71)
Jogging	4.29 (1.68)
Boxing	7.22 (5.03)

Mean and variance of the errors (in pixels) on the HumanEva testing set.





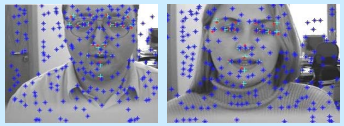
Ternary diagram showing the average error vs. descriptor weights d

Error Mean (top) and variance (bottom) vs. covariance scale s

Fiducial

Estimation of the 2D location of 20 fiducial points (e.g. nose, mouth, eye corners)
BioID database – 1521 images – 23 users

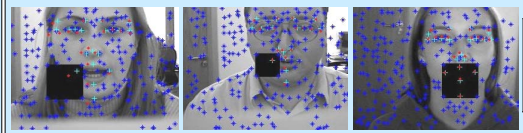
- Database generated from 1000 training images ~ 90000 patches.
- Testing set consists of 100 images (subjects different from training images)



	Error mean and variance (in pix.)
Full database	2.32 (0.52)

Occlusions

Occlusions were synthetically generated (black rectangles) in real images



Occlusion size	Error mean and variance (in pix.)
30x30	2.52 (0.61)
60x60	4.46 (8.41)

Example of results for the *occlusion* experiment.