Embedding Geometry in Generative Models for Pose Estimation of Object Categories

Michele Fenzi
http://www.tnt.uni-hannover.de/staff/fenzi
Jörn Ostermann
http://www.tnt.uni-hannover.de/staff/ostermann

Institut für Informationsverarbeitung (TNT)
Leibniz Universität Hannover
Hannover, Germany

Pose estimation for object classes is central in many Computer Vision tasks. Many approaches have been proposed to estimate the pose of an unknown object from a given category, and those based on local features have shown to be very effective. While some use 3D information obtained through CAD models [4] or 3D reconstructions [2], others have shown that coupling feature regression and view labeling efficiently solves this task [1, 5]. However, they rely solely on the discriminative power of local features, and this is problematic if objects have similar appearance in different views, as Figure 1 shows. To handle these situations they need to resort to external coarse-grained pose estimators for disambiguation.

We propose a method that solves this problem by integrating feature regression and graph matching in a unified probabilistic framework. The former predicts the descriptor of each patch in a query pose, while the latter evaluates the geometrical consistency between pairs of matches. As a consequence, our approach does not resort to external pose pre-processing and in addition experimentally shows to be more accurate in comparison. This permits to avoid any initial hard decision, postponing it to a later stage when more data is available.

Feature regression allows to treat pose estimation as a continuous problem, unlike most methods that provide only discrete values for the pose [3, 4]. Graph matching permits to *softly* align the unknown object to the class model, bringing additional consistency and precision to the solution. In a nutshell, our method retains the benefits of regression-based methods, like continuity and generality, while favoring geometrically consistent results through graph matching.

Our feature regression method leverages [1]. Regression functions model feature descriptors as a function of the pose. Given a patch i, $t^i = \{(f^i_1,\alpha^i_1),(f^i_2,\alpha^i_2),\dots,(f^i_n,\alpha^i_n)\}$, *i.e.*, t^i is a set of feature descriptors f^i_j labelled by their corresponding viewing angle α^i_j. For each t^i, a generative feature model F^i is defined as a linear combination of Gaussian kernels centered at the training poses,

$$F^i(\alpha) = \sum_{j=1}^{n} G(\alpha,\alpha^i_j)\mathbf{w}^i_j,$$

where $\mathbf{w}^i_j$ are estimated from t^i, and G measures the distance between two viewing angles. The class model is built by grouping all tracks from all class instances on the basis of their similarity in descriptor and pose space through spectral clustering.

At run time, query features are matched against a set of model representatives, which are the cluster centers in descriptor space. The nearest neighbor matching in [1] is prone to ambiguities occurring with similar views. Graph matching permits to favor geometrically consistent poses by exploiting the inherent spatial ordering of the features.

According to the graph matching paradigm, each feature set is interpreted as an attributed graph defined by $G = (V,E,A)$, where V is the set of vertices, E is the set of edges and A is an attribute matrix. We consider all test features as nodes of the test graph G and a subset of the model features as nodes of the model graph G'. Each entry A_{mn} represents some *relationship* between vertices $m,n \in V$. We defined $A_{mm} = f_m$, where f_m is the feature descriptor, and $A_{mn} = (\alpha_{mn}, r_{mn})$, where α_{mn} is the angle between the x-axis and the directed segment P_{mn} connecting the locations of features f_m and f_n, r_{mn} is the length of P_{mn}. $A'_{m'n'}$ is similarly defined.

We search for a mapping $M = \{(m,m')|m \in V, m' \in V'\}$ of the vertices that best respects the original attributes by maximizing the score

$$S = \sum_{(m,m')\in M,(n,n')\in M} g(A_{mn},A'_{m'n'}),$$

where g evaluates the attribute similarity. If M is expressed as a binary vector $\mathbf{x}$, such that $x_{mm'} = 1$ if $(m,m') \in M$, the problem solution is

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} S = \arg\max_{\mathbf{x}} \mathbf{x}^T\mathbf{W}\mathbf{x}, \qquad \text{s.t.}\ x_{mn} \in \{0,1\} \text{ and } \mathbf{Cx} = \mathbf{b},$$

where $\mathbf{W}$ is a matrix such that $W_{mm',nn'} = g(A_{mn},A'_{m'n'})$. $\mathbf{Cx} = \mathbf{b}$ is a set of linear constraints that may be imposed on the solution. Diagonal entries in $\mathbf{W}$ are defined in terms of the descriptor distance, so that a high entry is assigned to feature pairs close in descriptor space; off-diagonal entries are defined in terms of the absolute angular distance and the Euclidean distance ratio of the corresponding segments. Therefore, a high entry is assigned to feature pairs whose locations are geometrically consistent in orientation and distance.

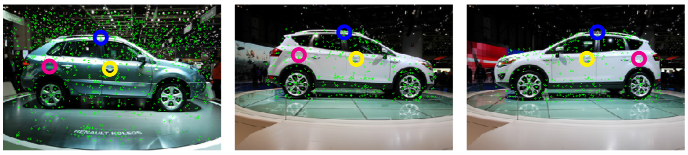

Figure 1: Query features (left) are matched to the class model represented by the two rightmost images. If matching is based on descriptor distance and absolute spatial distance between features, ambiguity still remains. If oriented distances are considered, the correct configuration is favored.

By relaxing the integer quadratic problem, the solution $\mathbf{x}^*$ is the principal eigenvector of $\mathbf{W}$. As $\mathbf{W}$ has only non-negative entries, all entries in $\mathbf{x}^*$ are in $[0,1]$, and the solution can be interpreted in probabilistic terms.

If $p(\alpha,c|f) = p(\alpha|f,c)p(c|f)$ expresses the likelihood of observing feature f from viewpoint α and c being the correct match ($f \sim c$), then the best pose and matching for the query feature set $\mathcal{F} = \{f\}_{q=1}^{Q}$ and model set $\mathcal{C} = \{c\}_{r=1}^{N}$ is

$$(\alpha^*,c^*) = \arg\max_{(\alpha,c)} \sum_{(q,r):f_q\sim c_r} p(\alpha|f_q,c_r)p(c_r|f_q),$$

where $p(\alpha|f,c)$ is expressed in terms of the generative feature model and $p(c|f)$ in terms of the graph matching results. As $\|\mathbf{x}\| = 1$ and $x^*_{fc} \in [0,1]$, the square of each score can be interpreted as a probability.

Method	MAE [°] (90th percentile)	MAE [°]
Ozuysal et al. [3]	-	46.48
Torki et al. [5]	19.40	33.98
Fenzi et al. [1]	14.51	31.27
Ours	**12.67**	**23.38**

Table 1: EPFL dataset [3].

Experiments on two car datasets show that our approach outperforms state-of-the-art algorithms by 25%, as Table 1 shows. Even when the pose classifier is almost perfect, our method not only recovers the correct orientation over the whole pose range, instead of the smaller correct interval given by the classifier, but it is also more accurate.

The increase in performance is due to the higher capability of our algorithm to solve view-problematic situations as well as to an overall additional accuracy given by the introduction of geometric context in the process.

[1] M. Fenzi, L. Leal-Taixé, B. Rosenhahn, and J. Ostermann. Class Generative Models based on Feature Regression for Pose Estimation of Object Categories. In *CVPR*, 2013.

[2] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich. Viewpoint-Aware Object Detection and Pose Estimation. In *ICCV*, 2011.

[3] M. Özuysal, V. Lepetit, and P. Fua. Pose Estimation for Category Specific Multiview Object Localization. In *CVPR*, 2009.

[4] B. Pepik, P. Gehler, M. Stark, and B. Schiele. 3D^{2}PM - 3D Deformable Part Models. In *ECCV*, 2012.

[5] M. Torki and A. M. Elgammal. Regression from Local Features for Viewpoint and Pose Estimation. In *ICCV*, 2011.