A Metric Learning Approach for Multi-View Object Recognition and Zero-shot Pose Estimation

Alina Kuznetsova[1], Sung Ju Hwang[2], Bodo Rosenhahn[1], Leonid Sigal[3]
[1]Leibniz University Hannover [2]UNIST [3]Disney Research Pittsburgh

Pose estimation, especially in case of multiple object classes, remains an important and very difficult problem due to extreme pose-dependent appearance variations, as well as challenges associated with obtaining precise ground truth pose for real-world images, necessary for training supervised viewpoint estimation models [5]. As a result, annotated data is often available only for a limited number of classes, and therefore it would be desirable to extend the existing approach for pose estimation to generalize to the new classes. The important property to notice is that there exist common pose-specific similarities across categories.

To exploit those similarities, as well as to address the problems mentioned above, we propose a metric learning approach for the task of joint object categorization and pose estimation, that does not require precise or dense viewpoint labels. Moreover, the learned metric generalizes to new classes, for which the pose labels are not available, and therefore makes it possible to use only partially annotated training sets, relying on the intrinsic similarities in the viewpoint manifolds for information transfer.

We resort to metric learning because modeling the pose variation with similarity constraints is a natural way to express on one side continuity of the appearance variation due to pose changes (unlike classification with discrete labels) and on another side allows to make learned metric independent of the pose label type (unlike regression approaches).

To summarize, our contributions are the following: 1) We explore metric-learning-based approaches for simultaneous pose and category prediction and show how to extend these methods for detection. 2) We propose a novel multi-task metric learning approach, which shares a common metric among the classes to capture shared view-specific components, while still allowing to capture class-specific individual aspects of pose-parametrized appearance. 3) We show that models learned using the multi-task approach are capable of performing zero-shot pose estimation, which, to our knowledge, is a novel task not addressed by any existing models. 4) We obtain state-of-the-art performance on both pose and category recognition in 3DObjects [3] and PASCAL3D+ [5] datasets.[1]

Metric learning formulation: Past works have shown that the instances in different viewpoints form a continuous low-dimensional manifold in the original feature space [6], and our goal is to exploit and magnify such manifold structure with distance constraints; more specifically, we want to learn a Mahalanobis distance matrix Q, such that a sample has a smaller distance to another sample in a similar pose, compared to the distance to a sample that has a very different pose. Given two points $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^D$, the distance between these two points is defined as

$$d_Q(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T Q (\mathbf{x}_i - \mathbf{x}_j) \tag{1}$$

Given the training set $\mathcal{D} = \{(\mathbf{x}_i, y_i, \mathbf{p}_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^D$ is D-dimensional feature vectors, $y_i \in 1 \dots C$ — category labels, and $\mathbf{p}_i \in \mathbb{R}^P$ — pose labels, the problem of learning a joint metric for class and pose estimation can be written as:

$$\min_{Q} \sum_{ijl} (1-\mu)\xi^+_{ijl} + \sum_{ijl} \zeta^+_{ijl}\mu + \lambda \mathrm{tr}(Q), \quad Q \succeq 0 \tag{2}$$

$$d_Q(x_i, x_j) + m_c \leq d_Q(x_i, x_l) + \zeta_{ijl}, \qquad y_i = y_j, y_i \neq y_l \tag{3}$$

$$d_Q(x_i, x_j) + m_v \leq d_Q(x_i, x_l) + \xi_{ijl}, \qquad y_i = y_j = y_l \tag{4}$$

$$d_P(p_i, p_j) \leq t_l, d_P(p_i, p_l) \geq t_u \tag{5}$$

where $\mathrm{tr}(Q)$ is the trace of Q and Q is semidefinit, $\xi^+ = \max(\xi, 0)$, and $d_P(\cdot,\cdot)$ is the distance in the pose space, specific for the annotations provided. Further, t_l, t_u are similarity and dissimilarity thresholds and m_c and m_v are the margins. The relative scale of m_c and m_v is crucial for learning; experimentally we verified, that setting $m_v = m_c/C$ gives good results. Since the view manifold is low-dimensional, it is reasonable to require Q to be low-rank. Minimizing the rank can be in turn approximated by the nuclear norm $\|Q\|_*$, which is equivalent to $\mathrm{tr}(Q)$, that we minimize in (2), for a positive semidefinite matrix Q. However, in general, different classes may not share identical pose metrics. Moreover, classification task differs significantly from the viewpoint estimation task, and therefore the requirements imposed on the metric could differ and even be contradictory, when only a single metric Q is learned. We resolve this issue by introducing a global shared metric Q_0 that discriminates classes as well as preserves common manifold for view estimation. We then enable each class to have its own pose metric Q_c: We propose the following multi-task formulation:

Figure 1: We learn a global metric Q_0 to discriminate classes and preserve global view-specific appearance, as well as class-specific pose estimation metrics Q_{car} and Q_{bus}. This joint learning allows to predict the pose for instances of novel object classes. For example, we can estimate the pose of the class *bus* by utilizing the view labels for class *car*, which is its neighbor in the class space.

$$\min_{Q_0, Q_1, \dots, Q_C} \sum_{ijl} \xi^+_{ijl}(1-\mu) + \sum_{ijl} \zeta^+_{ijl}\mu + \lambda \mathrm{tr}(Q_0) + \sum_c \gamma \mathrm{tr}(Q_c) \tag{6}$$

$$d_{Q_0}(x_i, x_j) + m_c \leq d_{Q_0}(x_i, x_l) + \xi_{ijl}, \qquad y_i = y_j, y_i \neq y_l \tag{7}$$

$$d_{Q_0+Q_c}(x_i, x_j) + m_v \leq d_{Q_0+Q_c}(x_i, x_l) + \zeta_{ijl}, \qquad y_i = y_j = y_l = c \tag{8}$$

$$d_P(p_i, p_j) \leq t_l, d_P(p_i, p_l) \geq t_u, \quad Q_0 \succeq 0, Q_c \succeq 0, c = 1 \dots C$$

The optimization problems formulated above are instances of semidefinite programming. We use a variant of stochastic projected gradient descend which subsamples active constraints for optimization.

Pose estimation and class prediction: Given a set of training triplets $\mathcal{D}$ and a set of learned metrics $Q_0, Q_c, c = 1 \dots C$, for a new sample $\mathbf{x}^*$, the k nearest neighbors $\{\mathbf{x}_i\}_{i \in I_k}$ from the training set are selected using the set of learned metrics, such that distance to the sample $\mathbf{x}_i$ is measured as $d_i = d_{Q_0+Q_{y_i}}(\mathbf{x}^*, \mathbf{x}_i)$. The final pose prediction $\mathbf{p}$ is formed by finding modes of each class of the resulting set, with the confidence defined as $r^{\mathbf{p}} = \sum_{j \in I(\mathbf{p})} d_j^{-1}$, where $I(\mathbf{p}) \subset I_k$ is a subset of the neighbors contributing to the mode.

Class label prediction is done by performing k nearest neighbor search using the learned metric Q_0, and choosing the weighted mode of their class labels as the final prediction; the confidence for the class c is than computed as $r^c = \sum_{j \in I(c)} d_j^{-1}, \quad I(c) = \{j : j \in I_k, y_j = c\}$

Zero-shot pose prediction: The proposed algorithm for pose estimation can be extended for pose prediction for the categories without any pose labels. To do so, we train the model using (6)-(8) (or (2)-(4)) without imposing view-preserving constraints on the categories that do not have viewpoint labels. Then, for zero-shot pose estimation, we only consider the samples that have pose labels as potential nearest neighbors. The main assumption we make here is that the training set will contain some categories that are similar to the category without pose labels.

Since different categories might have different, unaligned, pose labels, the prediction for a sample from a category without a pose label C_z is formed as a set $\tilde{p} = \{\mathbf{p}^c \in \mathbb{R}^P\}_{c \in \mathcal{C}}$, where $\mathbf{p}^c$ is the prediction of the class c and $\mathcal{C}$ denotes different classes among k nearest neighbors. In the experiments we observed, that only a small subset of all classes participate in the prediction formation for all samples of the class C_z.

The set of predictions $\tilde{p}$ can be afterwords transformed to the relative pose prediction between two samples of the class C_z. The relative pose between two samples can be computed as:

$$d(\tilde{p}_i, \tilde{p}_j) = \frac{1}{|\mathcal{C}_{act}|} \sum_{c \in \mathcal{C}_{act}} d_P(\mathbf{p}^c_i, \mathbf{p}^c_j), \tag{9}$$

where $\mathcal{C}_{act} = \mathcal{C}_i \cap \mathcal{C}_j$ is the set of the classes, that formed prediction for both samples i and j; if two samples have a non-intersecting set of predicting

[1]The original publication will appear at AAAI 2016 as "Exploiting View-Specific Appearance Similarities Across Classes for Zero-shot Pose Prediction: A Metric Learning Approach"

	bicycle	car	cell	iron	mouse	shoe	stapler	toaster	mean
KNN-VC	47.0/17.1	**47.3**/**25.0**	45.6/20.7	45.6/19.1	43.2/20.8	**48.5**/22.7	**47.2**/20.8	41.9/19.6	45.7/20.7
J-VC	48.4/20.5	44.6/23.5	**46.3**/**22.5**	45.5/20.6	44.9/23.9	46.1/25.2	46.4/22.8	42.0/19.2	45.5/22.3
MM-VC	47.3/19.7	37.9/19.9	45.5/21.6	44.7/19.2	43.1/21.0	44.6/24.9	45.8/21.8	40.1/18.0	43.6/20.7
MMJ-VC	**49.0**/**20.6**	45.6/24.1	45.7/22.2	**46.3**/**20.8**	**45.1**/**23.2**	48.1/**26.8**	46.5/**22.5**	**43.3**/**20.3**	**46.2**/**22.6**

Table 1: 3DObjects: zero-shot pose estimation accuracy (Acc^{ϕ}/$Acc^{(\phi,\theta)}$).

	aero	bike	boat	bottle	bus	car	chair	table	mbike	sofa	train	tv	mean
KNN-VC	**37.75**	**39.98**	36.55	29.34	31.17	33.14	39.68	48.12	**39.90**	48.37	28.40	47.71	38.34
J-VC	35.65	36.62	35.57	**57.72**	33.71	33.03	37.08	**49.90**	36.77	**55.07**	34.66	55.13	41.74
MM-VC	36.60	40.30	35.34	37.44	**40.71**	**33.97**	39.43	48.07	35.16	51.09	36.22	49.88	40.35
MMJ-VC	34.42	37.79	**36.66**	56.42	36.11	32.47	36.32	49.81	37.81	54.32	**38.49**	**57.40**	**42.33**

Table 2: PASCAL3D+: zero-shot pose estimation accuracy ($Acc_{\pi/6}$) for the whole dataset.

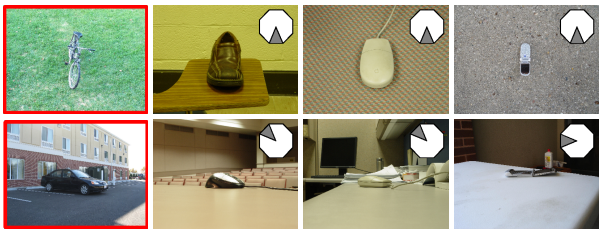

Figure 2: Zero-shot pose estimation examples: the first column shows the input image (denoted by the red boundary) and the remaining columns show samples selected for pose prediction.

classes, $d(\tilde{p}_i, \tilde{p}_j)$ is set equal to the maximal distance in the pose space.

To compute the absolute pose, a small set of the zero-pose samples from the class C_z with zero labels is selected and the relative distance to these samples in pose space allows to determine the absolute pose of a new sample.

Detection: We propose to couple the proposed method with the pre-trained R-CNN detector [2] by re-scoring object proposals using the confidence scores of the model. In the experiments, we show that the combined model allows us to improve the performance of the detector and, in addition, estimate the view point (the detection results are not presented in the abstract, but will be presented at the workshop, due to space limitations).

Experiments: To stress that our approach is independent of the type of labeling provided as pose annotations, we chose one dataset containing discrete labels (3DObjects [3]) and one with continuous labels (PASCAL3D+ [5]) We evaluate performance of our method in two main experiments: 1) the fully supervised case; 2) the zero-shot learning experiment, where we exclude ground truth pose labels from training for one class and evaluate the performance of the model for the same class. We perform this experiment for all classes.

	class	Acc^{ϕ}	Acc^{θ}	$Acc^{(\phi,\theta)}$
[1]	75.7	57.2	59.8	–
[3]	90.53/83.07	80.34/81.86	–	–
KNN-VC	95.17	84.94	85.20	71.68
J-VC	97.35	89.92	91.65	80.84
MM-VC	96.14	89.87	91.69	**82.79**
MMJ-VC	**97.36**	**90.15**	**91.82**	82.00

Table 3: 3DObjects: class recognition and pose estimation accuracy (%)

We compare our method with the state of art methods, as well as provide our baselines to show the advantage of our final formulation in various tasks. We use the following variants of our model for comparison: **KNN-VC** is a simple k nearest neighbors baseline to show the improvement due to learned metric in comparison to the original metric in the feature space, while in **MM-VC** we learn one metric per class for pose estimation, as well as a separate metric for class prediction. To show that joint learning for pose and category prediction can be beneficial, we compare **J-VC** learned using Eq. (2)-(4) and the full multi-metric model **MMJ-VC**, where we learn the multi-metric model, described in Eq. (6)-(8).

3DObjects dataset: This dataset contains 10 object classes, where each class has 10 instances that are presented in different views and scales. The view space is discretized by azimuth angle ϕ into 8 intervals, and by elevation angle θ into 3 intervals. We following the protocol of [3] in our experiments to measure accuracy for azimuth Acc^{ϕ}, elevation Acc^{θ} and total accuracy $Acc^{(\phi,\theta)}$, as well as classification accuracy in Table 3 for the fully supervised case. The learned metrics outperforms a simple KNN-VC baseline both in recognition and in pose estimation, as well as, by far, outperform other approaches.

We evaluate the performance in the zero-shot pose estimation experiment using the relative pose given by Eq. (9). The results are presented in Table 1. Since objects in 3DObject dataset are very distinct, only general features, such as rectangular form, can be transferred between categories (see Fig. 2 — the similarity between the query object and the nearest neighbors is mainly due to rough object form rather than due to details). However, we still are able to predict the pose for the objects from the novel category about 3 times better then random. Our full multi-metric model (MMJ-VC) gives the best performance, since it contains both the joint multi-task learning objective and combination of shared and class-specific metrics. Notably, MM-VC performs slightly worse than simple KNN-VC baseline, which points to the key importance of joint multi-task learning for zero-shot prediction.

	class	$MedError$	$Acc_{\pi/6}$
KNN-VC	61.70/62.72	35.74/37.69	49.76/50.76
J-VC	71.49/82.23	**31.93**/31.54	51.31/55.05
MM-VC	70.35/**85.12**	36.61/38.48	48.55/47.35
MMJ-VC	**71.75**/83.06	32.81/**29.67**	**51.84**/**55.20**

Table 4: PASCAL3D+: class recognition and pose estimation accuracy ($MedError$ is given in degrees) (results on the whole dataset/non-truncated and non-occluded images only).

PASCAL3D+ dataset: The dataset contains images of 12 different categories from PASCAL VOC 2012 training and validation sets. For PASCAL3D+ dataset, we use the distance in the pose space $d_P(\mathbf{p}_1, \mathbf{p}_2)$, as well as two performance metrics, proposed in [4] for evaluation.

The results for the fully supervised case are presented in Table 4. As in the previous experiment, J-VC and MMJ-VC baselines perform better then KNN prediction, however, MM-VC baseline performs poorly this time, that suggests, that sharing appearance between categories, especially if there are multiple categories with similar appearance, can be beneficial.

The results of the zero-shot pose estimation are presented in Table 2. We achieve higher improvement compared to the results we have on the 3DObjects dataset. We attribute this to the fact, that PASCAL3D+ dataset contains many categories that have similar appearance variation due to viewpoint change, such as *bike* and *motorbike* or *car* and *bus*. Furthermore, although KNN-VC still performs slightly better for some classes, for these classes our model performs on par, while for the classes like *train* or *bus*, the performance gain of the proposed approach is significant with respect to KNN-VC.

Conclusion: We have presented a method for simultaneous class prediction and pose estimation using learned metrics, that is able to generalize to the novel classes at almost no cost. We have validated our method on two datasets, and have shown that jointly learned metric outperforms separately learned metrics for the fully supervised pose estimation as well as well generalizes pose estimates for a novel category without pose labels. Furthermore, we showed the multi-task joint formulation further outperforms a single-metric formulation (especially for zero-shot).

[1] Amr Bakry and Ahmed Elgammal. Untangling object-view manifold for multiview recognition and pose estimation. In *ECCV*, pages 434–449, 2014.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[3] Silvio Savarese and Fei-Fei Li. 3d generic object categorization, localization and pose estimation. In *ICCV*, 2007.

[4] Shubham Tulsiani and Jitendra Malik. Viewpoints and keypoints. In *CVPR*, 2015.

[5] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014.

[6] Haopeng Zhang, Tarek El-Gaaly, Ahmed M. Elgammal, and Zhiguo Jiang. Joint object and pose recognition using homeomorphic manifold analysis. In *AAAI*, 2013.