Fisher vector places: learning compact image descriptors for place recognition

Anonymous CVPR submission

Paper ID ****

Abstract

The goal of this work is to localize a query photograph by finding other images depicting the same place in a large geotagged image database. The contributions of this work are three fold. First, we learn a discriminative *yet* compact *descriptor of each image in the database. This is achieved by applying exemplar support vector machine (e-SVM) learning to compact Fisher vector descriptors extracted from database images. Second, we analyze the exemplar support vector machine objective and show that the learnt hyperplane can be interpreted as a new descriptor that replaces the original positive example and is re-weighted to increase its separation from the negative data. Third, based on this analysis we demonstrate that the learnt and re-normalized descriptor could be directly used for matching, thus avoiding the need for expensive and tedious calibration typically needed for exemplar support vector machine methods. Place recognition results are shown on two image datasets of Google street-view images from Pittsburgh and demonstrate the learnt representation consistently improves over the standard Fisher vector descriptors at different target dimensions.*

1. Introduction

The goal of this work is to localize a query image by matching to a large database of geotagged street-level imagery. This is an important problem with practical applications in robotics, augmented reality or navigation. This task is however very difficult. It is hard to distinguish different places, e.g. streets in a city, from each other. The imaged appearance of a place can change drastically due to factors such as viewpoint, illumination or even changes over time. Finally, with the emergence of planet-scale geotagged image collections, such as Google Street-view, the image databases are becoming very large. We estimate a single country like France is covered by more than 60 million street-level panoramas. Hence the fundamental challenge in place recognition lies now in designing robust, discriminative yet compact image representations.

In this work we build on the method of Gronat *et al.* [12] who represent each image in the database by a per-location classifier that is trained to discriminate each place from other places in the database. At query time, the query image is classified by all per-location classifiers and assigned to a place with the highest classification score. The training of each classifier is performed using the per-exemplar support vector machine (e-SVM) [22], which takes the positive image as a single positive example and other far away images in the database as negative examples. The exemplar SVM is well suited for this task as street-level image collections typically contain only one or at most hand-full of images depicting the same place. The intuition is that the exemplar SVM can learn the important features that distinguish the particular place from other similar places in the database. While the results of [12] are promising they suffer from two important drawbacks. First, the learnt place specific representation is not compact, which prohibits its application to planet-scale street-level collections that are now becoming available [16]. Second, the per-exemplar classifiers require careful and time-consuming calibration.

In this work we address both these issues. First, we apply the exemplar SVM training to compact Fisher vector [14, 24] image descriptors, which results in a *discriminative* yet *compact* representation of each image in the database. Second, to avoid the expensive classifier calibration, we analyze the exemplar SVM cost and show that the learnt hyperplane can be interpreted as a new descriptor that replaces the original positive example and is re-weighted to increase its separation from the negative data. As a result of this analysis, we demonstrate that after an appropriate normalization of the new re-weighted descriptor no further calibration is necessary. We show improved results on place recognition using learnt compact Fisher vector descriptors [14] of different dimensionality.

2. Related work

Large-scale visual place recognition The visual localization problem is typically treated as large-scale instance-level retrieval [6, 4, 12, 17, 27, 34, 35], where images are represented using local invariant features [21] aggregated into the bag-of-visual-words [5, 32] representation. The image database can be further augmented by 3D point clouds [16], automatically reconstructed by large-scale structure from motion (SfM) [1, 16], which enables accurate prediction of query image camera position [20, 25]. In this work we investigate learning a discriminative representation using the compact Fisher vector descriptors [15]. Fisher vector descriptors have shown excellent place recognition accuracy [34]. In this work we further improve their performance by discriminative learning.

Fisher vector image representations Fisher vector image representations have recently demonstrated excellent performance for a number of visual recognition tasks [3, 15, 18, 30]. They are specially suited for retrieval applications since they are robust to image appearance variations and capture richer image statistics than the simple bag-of-visual-words (BOW) aggregation. However, the raw extracted Fisher vectors are typically high-dimensional, e.g. with 32,768 non-sparse dimensions, which is impractical for large-scale visual recognition and indexing applications. Hence, their dimensionality is often reduced by principal component analysis (PCA) and further quantized for efficient indexing using, e.g. a product quantizer [15]. Other recent work has demonstrated improved performance in a face recognition application by finding discriminative projection using a large amount of training face data [30]. Our work is complementary to these methods as it operates on the projected low-dimensional descriptor and further learns discriminative re-weighting of the descriptor specific to each image in the database using per-exemplar support vector machine [22].

Per-exemplar support vector machine The exemplar support vector machine (e-SVM) has been used in a number of visual recognition tasks including category-level recognition [22], cross-domain retrieval [29], scene parsing [33], place recognition [12] or as an initialization for more complex discriminative clustering models [8, 31]. The main idea is to train a linear support vector machine (SVM) classifier from a single positive example and a large number of negatives. The intuition is that the resulting weight vector will give a higher weight to the discriminative dimensions of the positive training data point and will down weight dimensions that are non-discriminative with respect to the negative training data. A key advantage is that each per-exemplar classifier can be trained independently and hence the learning can be heavily parallelized. The per-exemplar training brings however also an important drawback. As each classifier is trained independently a careful calibration of the resulting classifier scores is required [12, 22].

Contributions The contributions of this work are three-fold. First, we analyze the exemplar support vector machine objective and show that the learnt hyperplane can be interpreted as a new descriptor that replaces the original positive example and is re-weighted to increase its separation from the negative data. Second, we demonstrate that after an appropriate normalization of the new re-weighted descriptor no further calibration is necessary. Third, we apply e-SVM training to compact Fisher vector descriptors for large-scale place recognition resulting in a *discriminative* yet *compact* representation of each image in the database. Place recognition results are shown on a dataset of 25k images of Pittsburgh and demonstrate the learnt representation consistently improves over the standard Fisher vector descriptors at different target dimensions.

3. Learning compact place descriptors using per-exemplar SVM

Each database image j is represented by its L2-normalized Fisher vector Φ_j. The goal is to learn a set of new L2-normalized Fisher vectors Ψ_j, one per each database image, such that at query time, given the Fisher vector Φ_q of an unknown query image, we retrieve the database image depicting the same location by finding the image j^* with the highest score measured by a dot product

$$j^* = \arg\max_j \Phi_q^T \Psi_j. \tag{1}$$

In other words, the aim is to replace each original database Fisher vector Φ_j with a new vector Ψ_j that is more discriminative in the sense of separation from descriptors of images depicting other places. Inspired by [12], we investigate applying the exemplar support vector machine (e-SVM) [22] for this task. e-SVM learns a linear classifier $w_j^T \Phi + b_j$ given the descriptor Φ_j^+ of place j as a single positive example (with target label +1) and a large number of negative descriptors $\mathcal{N}_j$ from other places in the database (with target labels -1). The intuition of the exemplar SVM training [22] is that the learnt weight vector w_j will give a higher weight to the dimensions of the descriptor that are discriminative and will down-weight dimensions that are non-discriminative with respect to the negative training data collected from other far-away places. The optimal w_j and b_j are obtained by minimizing the following objective

$$||w_j||^2 + C_1 \cdot h\left(w_j^T \Phi_j^+ + b_j\right) + C_2 \sum_{\Phi \in \mathcal{N}_j} h\left(-w_j^T \Phi - b_j\right), \tag{2}$$

where Φ_j^+ is the descriptor of the place j as the positive data point, Φ are Fisher descriptors from negative training data $\mathcal{N}_j$ and h is the hinge loss, $h(y) = \max(1 - y, 0)$. Note that the first term in (2) is the regularizer, the second term is the loss on the positive data weighted by scalar parameter C_1 and the third term is the loss on the negative data weighted by scalar parameter C_2. The objective is convex and can be minimized with respect to w_j and b_j using standard software packages such as [10]. A key advantage is that the per-exemplar classifier for each place can be trained independently and hence the learning can be heavily parallelized. The downside of the independent training for each positive example is that the resulting scores have to be calibrated with respect to each other on additional data [12, 22].

3.1. Analysis of per-exemplar SVM objective

In this section, we analyze the exemplar SVM objective (2) and show the learnt and *re-normalized* weight vector w_j can be interpreted as a new descriptor Ψ_j that replaces the original positive training descriptor Φ_j^+. In particular, we show first that when the weight C_2 of the negative data in objective (2) goes to zero and the learnt Ψ_j is identical to the original positive training data point Φ_j^+. Second, when $C_2 > 0$, the learnt Ψ_j moves away from the positive Φ_j^+ to increase its separation from the negative data. Details are given next.

Case I: $C_2 \rightarrow 0$**.** The goal is to show that when the weight C_2 of the negative data in objective (2) goes towards zero the resulting hyperplane vector w_j is parallel with the positive training descriptor Φ_j^+. When w_j is normalized to have unit L2 norm the two vectors are identical. First, let us decompose w into parallel and orthogonal part with respect to the positive training data point Φ^+ (in the following we omit index j for brevity), i.e. $w = w^\perp + w^{||}$, where $(w^\perp)^T\Phi^+ = 0$. Next, we observe that when the weight of the negative data diminishes ($C_2 \rightarrow 0$), any non-zero component $w^\perp$ will increase the value of the objective. As a result, for $C_2 \rightarrow 0$ the objective is minimized by $w^{||}$, i.e. the optimal w is parallel with Φ^+.

In detail, for $w = w^\perp + w^{||}$ the objective (2) can be written as

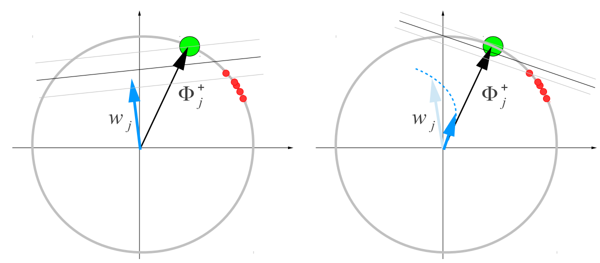
$$||w^\perp + w^{||}||^2 + C_1 \cdot h\left((w^\perp + w^{||})^T\Phi^+ + b\right) + C_2 \sum_{\Phi\in\mathcal{N}} h\left(-(w^\perp + w^{||})^T\Phi - b\right). \quad (3)$$

Note that the orthogonal part $w^\perp$ does not change the value of the second term in (3) because $(w^\perp + w^{||})^T\Phi^+ = (w^{||})^T\Phi^+$, and hence (3) reduces to

$$||w^\perp + w^{||}||^2 + C_1 \cdot h\left(w^{||T}\Phi^+ + b_j\right) + C_2 \sum_{\Phi\in\mathcal{N}} h\left(-(w^\perp + w^{||})^T\Phi - b\right). \quad (4)$$

In the limit case as $C_2 \rightarrow 0$ any non-zero component $w^\perp$ will increase the value of the objective (4). This can be seen by noting that the third term vanishes when $C_2 \rightarrow 0$ and hence the objective is dominated by the first two terms. Further, the second term in (4) is independent of $w^\perp$. Finally, the first term will always increase for any non-zero value of $w^\perp$ as $||w^\perp + w^{||}||^2 \geq ||w^{||}||$ for any $w^\perp \neq 0$.

As a result, in the limit case when $C_2 \rightarrow 0$ the optimal w is parallel with Φ^+. Note also, that when C_2 is exactly equal to zero, $C_2 = 0$, the optimal w vanishes, i.e. the objective (4) is minimized by trivial solution $||w|| = 0$ and $b_j = -1$. The effect of decreasing the parameter C_2 is illustrated in figure 1.



(a) $C_2 > 0$ (b) $C_2 \rightarrow 0$

Figure 1: **An illustration of the effect of decreasing parameter C_2 in the exemplar support vector machine objective.** The positive exemplar Φ_j^+ is shown in green. The negative data points are shown in red. All training data is L2 normalized to lie on a hyper-sphere. (a) For $C_2 > 0$, the normal w_j of the optimal hyper-plane moves away from the direction given by the positive example Φ^+ in a manner that reduces the loss on the negative data. (b) As the parameter C_2 decreases the learnt w_j becomes parallel to the positive training example Φ_j^+ and its magnitude $||w_j||$ goes to 0.

Case II: $C_2 > 0$**.** When the weight C_2 of the negative data in the objective (4) increases the direction of the optimal w will be different from $w^{||}$ and will change to take into account the loss on the negative data points. Explicitly writing the hinge-loss $h(x) = \max(1 - x, 0)$ in the last term of (4), we see that w will move in the direction that reduces $\sum_{\Phi\in\mathcal{N}} \max\left(1 + w^T\Phi + b, 0\right)$, i.e. that reduces the dot product $w^T\Phi$ on the negative examples that are active (support vectors).

3.2. Interpreting normalized w as a new descriptor

The above analysis demonstrates that as C_2 decreases the normal of the optimal hyperplane w that separates the positive exemplar Φ^+ from negative data becomes parallel with Φ^+, as shown in figure 1. As C_2 increases, the normal w of the optimal hyper-plane moves away from the direction given by the positive example Φ^+ in a manner that reduces the loss on the negative data. This suggests that the learnt w

could be interpreted as a modified positive example Φ^+, re-weighted to emphasize directions that separate Φ^+ from the negative data. As discussed above w is not normalized. As we wish to measure the similarity between descriptors by (the cosine of) their angle given by equation (1), additional normalization of the learnt w is necessary. Hence we define the new descriptor Ψ_j as the normalized hyperplane normal w_j

$$\Psi_j = \frac{w_j}{||w_j||}. \tag{5}$$

4. Experimental evaluation

In this section we first describe the experimental set-up, then we give implementation details, and finally report results of the proposed approach on two datasets where we compare performance with raw Fisher vector matching and several baselines methods.

4.1. Experimental set-up

We perform experiments on a database of Google Street View images of Pittsburgh downloaded from the Internet. The data contains panoramas covering roughly an area of $1.3 \times 1.2\ \text{km}^2$. Similar to [4], for each panorama we generate 12 overlapping perspective views corresponding to two different elevation angles to capture both the street-level scene and the building façades, resulting in a total of 24 perspective views each with 90° FOV and resolution of 960×720 pixels. For evaluation we have used two versions of this data. The first one was obtained from the authors of [12] (25k images). In the second version we download additional images to increase the dataset size to 55k images. As a query set with known ground truth GPS positions, we use the 8999 panoramas from the Google Street View research dataset. This dataset covers approximately the same area, but has been captured at a different time, and depicts the same places from different viewpoints and under different illumination conditions. For each test panorama, we generate a set of perspective images as described above. Finally, we randomly select out of all generated perspective views a subset of 4k images, which is used as a test set to evaluate the performance of the proposed approach. Since all the query images have associated GPS location we can compute their spatial distance from the database images returned by the matching method. We consider a query image to be correctly localized if the retrieved database image lies within a perimeter of $20m$ from the location of the query.

4.2. Implementation details

We first extract rootSIFT descriptors [2] for each image. Following [15] we project the 128-dimensional SIFT descriptors to 64 dimensions using PCA. The projection matrix is learnt on a set of descriptors from 5,000 randomly selected database images. This has also the effect of decorrelating the SIFT descriptor. The 64-dimensional SIFT descriptors are then aggregated into Fisher vectors using a Gaussian mixture model with $N = 256$ components, which results in a $2 \times 256 \times 64 = 32,768$-dimensional descriptor for each image. The Gaussian mixture model is learnt from descriptors extracted from 5,000 randomly sampled database images. The high-dimensional Fisher vector descriptors are then projected down to dimension $d \in \{128, 512, 2048\}$ using PCA learnt from all available images in the database. The resulting low dimensional Fisher vectors are then re-normalized to have unit L2-norm, which we found to be important in practice.

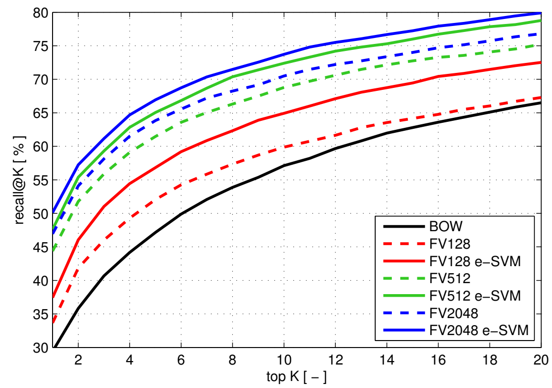


Figure 2: **Evaluation on Pittsburgh 25k [12] dataset.** The fraction of correctly recognized queries (recall@K, y-axis) vs. the number of top K retrieved database images for different Fisher vector dimensions. The learnt descriptors by the proposed method (FV e-SVM) consistently improve over the raw Fisher vector descriptors across the whole range of K and all dimensions.

Learning parameters and training data. To learn the exemplar support vector machine for each database image j, the positive and negative training data are constructed as follows. The *negative training set* $\mathcal{N}_j$ is obtained by: (i) finding the set of images with geographical distance greater than 200m; (ii) sorting the images by decreasing value of similarity to image j measured by the dot product between their respective Fisher vectors; (iii) taking the top $N = 500$ ranked images as the negative set. In other words, the negative training data consists of the hard negative images, i.e. those that are similar to image j but are far away from its geographical position, hence, cannot have the same visual content. The *positive training set* $\mathcal{P}_j$ consist of the original Fisher vector Φ_j of the image j. For the SVM training we use `libsvm` [10]. We use the same C_1 and C_2 parameters for all per-exemplar classifiers, but find the optimal value of the parameters for each dimensionality of the Fisher vector by a grid search evaluating performance on a held out set.

We observe that for different Fisher vector target dimensions the optimal value of parameter C_1 is quite stable (typically $C_1 = 1$) while the optimal parameter for C_2 varies between 10^{-6} to 10^{-1}. To learn the new image representation for each database image j we: (i) learn SVM from $\mathcal{P}_j$ and N_j (see above); (ii) $L2$ normalize the learned w_j using equation (5); and (iii) use this re-normalized vector as the new image descriptor Ψ_j for image j. At query time we compute the Fisher vector Φ_q of the query image and measure its similarity score to the learnt descriptors Ψ_j for each database image by equation (1).

4.3. Results

For each database (Pittsburgh 25k and Pittsburgh 55k) we compare results of our method (FV e-SVM) to two baselines: standard bag-of-visual-words baseline (BOW) and raw Fisher vector matching without learning (FV).

We perform experiments on several target Fisher vector dimensions $d \in \{128, 512, 2048\}$. For each method we measure performance using the percentage of correctly recognized queries (Recall) similarly to, e.g., [4, 17, 26]. The query is correctly localized if at least one of the top K retrieved database images is within 20 meters from the ground truth position of the query. Results are shown for different values of K in table 1. For the Pittsburgh 25k we also show results in the form of a curve in figure 2. The results clearly demonstrate the benefits of the learnt descriptors with respect to the standard Fisher vectors for all target dimensions and lengths of shortlist K. The benefits of discriminative learning are specially prominent for low-dimensional compact descriptors ($d = 128$). The proposed method also significantly outperforms the bag-of-visual-words baseline. Figure 4 shows examples of place recognition results.

Comparison to other methods. On the Pittsburgh 25k database, we compare performance of our learnt discriminative descriptors to the methods of [12] and [17], who report on the same testing data top $K = 1$ recall of 36.5% and 41.9%, respectively (results taken from [12]). Our method outperforms [17] already for dimension $d = 128$ (37.8%) and [12] for dimension $d = 512$ (47.6%). Furthermore, note that [12, 17] are based on a bag-of-visual-words representation, which typically needs to store between 1000-2000 non-zero visual words per image, which is significantly more than our learnt 128 or 512 dimensional descriptor.

Memory complexity analysis. Figure 3 compares the performance of the learnt discriminative descriptor (FV eSVM) to the raw Fisher vectors (FV) for different target dimensions. The results demonstrate that for a given level of accuracy (y-axis) our method learns a more compact (lower-dimensional) representation (x-axis). For example, our learnt 128-dimensional descriptor achieves a similar accuracy (around 65%) to the 256-dimensional raw Fisher descriptor essentially reducing the memory complexity to 50% for the same level of performance. Note that similar to [15], we observe decrease in performance at high-dimensions for both the baseline and our method.

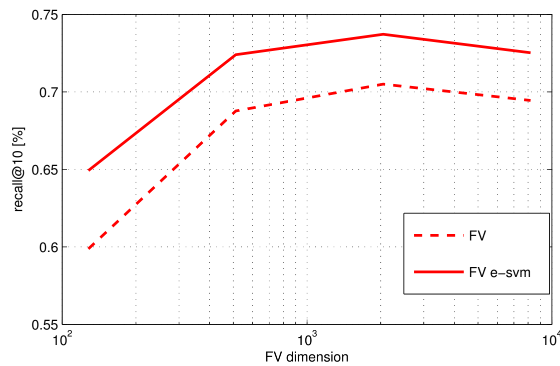


Figure 3: **Memory complexity analysis.** The fraction of correctly localized queries at the top 10 retrieved images (y-axis) for different Fisher vector dimensions (x-axis). The learnt descriptors by our method (FV e-SVM) clearly outperform the raw Fisher vector descriptors (FV) for all dimensions. Note that for a certain level of performance (y-axis) the proposed method learns a more memory efficient (lower dimensional, x-axis) descriptor.

5. Conclusions

We have shown that a discriminative yet compact image representation for place recognition can be learnt using the exemplar support vector machine applied to Fisher vector image descriptors without the need for expensive and tedious calibration typical for other exemplar support vector machine methods. Our results show significant gains in place recognition performance compared to raw Fisher vector matching as well as other baselines. Our work opens-up the possibility of learning a compact and discriminative representation using other descriptors such as HOG [7] or the recently developed convolutional neural network features [9, 19, 23, 28] as well as extending the analysis to other cost functions [11, 13].

References

[1] S. Agarwal, N. Snavely, I. Simon, S. Seitz, and R. Szeliski. Building Rome in a day. In *ICCV*, pages 72–79, 2009. 2

[2] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *IEEE PAMI*, 2012. 4

[3] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *Proc. BMVC*, 2011. 2

Figure 4: **Examples of correctly and incorrectly localized queries for the learnt Fisher vector representation.** Each example shows a query image (left) together with correct (green) and incorrect (red) matches from the database obtained by the learnt Fisher vector representation *w-norm* method (top) and the standard Fisher vector baseline (bottom) for dimension 128. Note that the proposed method is able to recognize the place depicted in the query image despite changes in viewpoint, illumination and partial occlusion by other objects (trees, lamps) and buildings. Note that the baseline methods often finds images depicting the same buildings but in a distance whereas our learnt representation often finds a closer view better matching the content of the query.

Method:	25k Pittsburgh					55k Pittsburgh				
recall@K [%]	1	2	5	10	20	1	2	5	10	20
BOW	29.4	35.7	47.0	57.1	66.5	8.7	11.0	17.3	22.8	25.4
FV128	33.6	41.8	52.0	59.8	67.7	10.9	14.1	20.2	26.4	33.2
FV128 e-SVM	**37.8**	**46.1**	**56.9**	**64.9**	**72.6**	**13.5**	**17.7**	**25.0**	**31.8**	**39.0**
FV512	44.3	51.7	61.4	68.7	75.2	17.3	21.1	28.4	34.2	40.3
FV512 e-SVM	**47.6**	**55.4**	**65.1**	**72.4**	**78.8**	**19.8**	**25.1**	**32.7**	**38.7**	**46.0**
FV2048	46.9	54.1	63.8	70.5	76.8	19.2	23.5	29.9	35.2	41.9
FV2048 e-SVM	**50.2**	**57.3**	**67.0**	**73.8**	**78.0**	**20.8**	**25.9**	**33.1**	**38.7**	**45.9**

Table 1: The fraction of correctly recognized queries (recall@K) vs. the number of top $K \in \{1, 2, 5, 10, 20\}$ retrieved database images for different Fisher vector dimensions $d \in \{128, 512, 2048\}$. The learnt descriptors by the proposed method (FV e-SVM) consistently improve over the raw Fisher vector descriptors across the whole range of K and all dimensions on both the 25k and 55k Pittsburgh image datasets.

[4] D. Chen, G. Baatz, Köser, S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. City-scale landmark identification on mobile devices. In *CVPR*, 2011. 1, 4, 5

[5] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004. 2

[6] M. Cummins and P. Newman. Highly scalable appearance-only SLAM - FAB-MAP 2.0. In *Proceedings of Robotics: Science and Systems*, Seattle, USA, June 2009. 1

[7] N. Dalal and B. Triggs. Histogram of oriented gradients for human detection. In *CVPR*, 2005. 5

[8] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes paris look like paris? *SIGGRAPH*, 31(4), 2012. 2

[9] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv:1310.1531*, 2013. 5

[10] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *J. Machine Learning Research*, 9:1871–1874, 2008. 3, 4

[11] M. Gharbi, T. Malisiewicz, S. Paris, and F. Durand. A Gaussian approximation of feature space for fast image similarity. Technical report, MIT, 2012. 5

[12] P. Gronat, G. Obozinski, J. Sivic, and T. Pajdla. Learning and calibrating per-location classifiers for visual place recognition. In *CVPR*, 2013. 1, 2, 3, 4, 5

[13] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *ECCV*, 2012. 5

[14] H. Jégou, M. Douze, and C. Schmid. Product Quantization for Nearest Neighbor Search. *IEEE PAMI*, 33(1), 2011. 1

[15] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE PAMI*, 34:1704–1716, 2012. 2, 4, 5

[16] B. Klingner, D. Martin, and J. Roseborough. Street view motion-from-structure-from-motion. In *ICCV*, 2013. 1, 2

[17] J. Knopp, J. Sivic, and T. Pajdla. Avoidng confusing features in place recognition. In *ECCV*, 2010. 1, 5

[18] J. Krapac, J. Verbeek, and F. Jurie. Modeling Spatial Layout with Fisher Vectors for Image Categorization. In *ICCV*, 2011. 2

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 5

[20] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua. Worldwide pose estimation using 3d point clouds. In *ECCV*, 2012. 2

[21] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 1

[22] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011. 1, 2, 3

[23] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014. 5

[24] F. Perronnin, Y. Liu, J. Snchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *CVPR*. IEEE, 2010. 1

[25] T. Sattler, B. Leibe, and L. Kobbelt. Improving image-based localization by active correspondence search. In *ECCV*, 2012. 2

[26] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt. Image retrieval for image-based localization revisited. In *Proc. BMVC*, 2012. 5

[27] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *CVPR*, 2007. 1

[28] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv:1312.6229*, 2013. 5

[29] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros. Data-driven visual similarity for cross-domain image matching. In *SIGGRAPH ASIA*, 2011. 2

[30] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher Vector Faces in the Wild. In *British Machine Vision Conference*, 2013. 2

[31] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012. 2

[32] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 2

[33] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013. 2

[34] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi. Visual place recognition with repetitive structures. In *CVPR*, 2013. 1, 2

[35] A. Zamir and M. Shah. Accurate image localization based on google maps street view. In *ECCV*, 2010. 1