Fisher vector places: learning compact image descriptors for place recognition

Anonymous Author(s)
Affiliation
Address
email

Abstract

Exemplar support vector machines (e-SVM) are emerging as a powerful tool to learn a discriminative pre-example representation for visual recognition. However, as individual classifiers are learnt independently for each positive example the classifier scores require careful and tedious calibration on an independent held-out data. The contribution of this work are three fold. First, we analyze the e-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 distance 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 with different 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.* [11] 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) [23], which takes the positive image as a single positive example and other far away images in the database as negative data. 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 [11] 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 [17]. 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 [13, 26] 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 distance 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 of different dimensionality without the need for additional calibration. The same procedure be potentially applied to other descriptors such as HOG [7] or the recently developed convolutional neural network features [9, 20, 24, 30].

2 Related work

Large-scale visual place recognition. The visual localization problem is typically treated as large-scale instance-level retrieval [6, 4, 12, 18, 29, 38, 39], where images are represented using local invariant features [22] aggregated into the bag-of-visual-words [5, 34] representation. The image database can be further augmented by 3D point clouds [17], automatically reconstructed by large-scale structure from motion (SfM) [1, 17], which enables accurate prediction of query image camera position [21, 28]. 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 [38]. 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, 19, 32]. 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. product quantizer [15]. Other recent work has demonstrated improved performance in a face recognition application by finding discriminative projection using large number of training face data [32]. 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 for specific to each image in the database using per-exemplar support vector machine [23].

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 [23], cross-domain retrieval [31], scene parsing [36], place recognition [11] or as an initialization for more complex discriminative clustering models [8, 33]. 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 on is required [11, 23].

Contributions. The contributions of this work are threefold. First, we analyze the 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 distance 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 with different dimensions.

3 Fisher vector overview

(This is a minimalistic overview of what the FV is. This can be omited or shrinked but I think it is worth since many people are not familiar with FV)

The Fisher vector can be thought of as an extension of BOW [35] that is achieved by considering a high-order statistics of the distribution of the feature descriptors. It aggregates a large set of descriptors into a high-dimensional representation of the fixed size. In the following text we briefly overview some basic concepts of Fisher vectors, its normalization and standard dimensionality reduction.

3.1 Computing Fisher vectors

It first starts with learning a generative model of the local image descriptors x of the dimension d (in our case we use SIFT features, see details in Sec. 5). The model is assumed to follow a GMM with N components and diagonal covariances. This model can be understood as a probabilistic visual vocabulary. Having the model trained, the Fisher vector of the sample of the image descriptors is computed as a gradient of the sample's likelihood with respect to the learned parameters of the GMM, which is subsequently scaled by inverse square root of the Fisher information matrix [25]. Considering only the derivatives w.r.t. the mean and covariances, we a obtain representation which captures the average of the first and second order differences between the descriptors and each of the GMM center k. For the image containing T feature vectors it can be expressed as follows:

$$\Phi_k^{(1)} = \frac{1}{N\sqrt{(w_k)}} \sum_{j=1}^{T} \alpha_j(k) \left(\frac{x_j - \mu_k}{\sigma_k} \right) \tag{1}$$

$$\Phi_k^{(2)} = \frac{1}{N\sqrt{(2w_k)}} \sum_{j=1}^{T} \alpha_j(k) \left(\frac{(x_j - \mu_k)^2}{\sigma_k^2} - 1 \right) \tag{2}$$

where w_k, μ_k and σ_k are weight, mean and diagonal of covariance matrix for k-th component of the learned GMM and $\alpha_j(k)$ is a soft assignment of the j-th feature x_j to the Gaussian k. The resulting Fisher vector Φ is then concatenation of these gradients which forms into a $2Nd$ dimensional vector such that $\Phi = [\Phi_1^{(1)T}, \Phi_1^{(2)T}, ..., \Phi_N^{(1)T}, \Phi_N^{(2)T}]^T$.

While some works, e.g.[26], use only a partial derivatives w.r.t. the mean parameters (equation (1)), we use partial derivatives both w.r.t. the mean and variance because this approach have been implemented in a library that we used (details are provided in section 5). As reported in [14] both approaches provide comparable results.

3.2 Normalization

As in the case of standard BOW, Fisher vector is typically $L2$ normalized in order to measure similarity between two FVs by a dot-product. However it has been shown that direct $L2$ normalization

is suboptimal and better performance can be achieved by power-normalization followed by $L2$ normalization. Thus the normalization operator can be written as follows:

$$L(z) = \frac{\text{sign}(z)\,|z|^{\alpha}}{||\text{sign}(z)\,|z|^{\alpha}\,||_2} \tag{3}$$

where $\alpha \in (0, 1)$ is an element-wise power and $z \in R^d$ is an arbitrary real vector. As mentioned in the introduction, the power-normalization has the effect of suppression of the descriptors that occur frequently in the image (bursty features), these can be for instance periodical structures such as skyscraper windows or bricks of the wall. In our setup we use $\alpha = 0.5$ as it has been reported by Jegou et al. [16] that this value is optimal for wide range of GMM components N.

3.3 Dimension reduction

The resulted Fisher vector is a high-dimensional representation of an image, which dimensionality depends only on number of N components of the GMM and dimension d of the local image descriptors which is fixed. In practice, the PCA is being used for decorrelation and projection to a lower dimensional space because it subsequently directly affects a size of the indexed signature (see Sec. **??** and [13]).

We observed that after the PCA decorrelation it is important to $L2$ re-normalize projected FV to achieve better performance. Our interpretation is that after performing the PCA the higher dimensions contain a noise that affects only a magnitude of the projected vectors but not its direction Figure or not?. Furthermore we observed that even slightly better results can be achieved by using the normalization operator (3) instead of $L2$.

Finally, it is worth noting that PCA projection into a 'not-too-low' dimensional space can actually slightly improve the performance of the full FV. What that 'not-too-low' actually means have been partially discussed in [16]. However, in general as the dimension decreases the performance of decorrelated FV decrease which is, indeed, what we observe in our experiments. In the next section we therefore aim on enhancing the performance of projected FVs.

4 Learning compact place descriptors using per-exemplar SVM

Each database image j is represented by its Fisher vector Φ_j. The goal is to learn a set of new 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 can retrieve the the correct database image as the image j^* with the highest score

$$j^* = \arg\max_j \Phi_q^T \Psi_j \tag{4}$$

Hence, we aim on replacing the original database Fisher vector Φ with some new vector Ψ that performs better in the sense of separation from descriptors of other places and that has a unit $L2$-norm in order to measure the similarity by a dot-product (eq. (4)).

Per-exemplar SVM learning. We follow the approach of [12] and [23] and we learn Ψ_j independently for each image in turn using per-exemplar SVM. In the following we analyze a limit behavior of per-exemplar SVM convex objective and show that learned hyperplane normal can be interpreted as a new descriptor that increases the measured distance from negative data.

The objective of per-exemplar SVM that can be written as

$$||w_j||^2 + C_1 \cdot h\left(w_j^T \Phi_j^+ + b_j\right) + C_2 \sum_{\Phi \in \mathcal{N}_j} h\left(-w_j^T \Phi - b_j\right) \tag{5}$$

where Φ_j^+ is a single positive exemplar, Φ are FVs from negative training data, $||w_j||$ is a regularization term, h is a penalty function and C_1, C_2 are penalty weights for positive or negative data $\mathcal{N}_j$.

Analysis of per-exemplar SVM objective. In the following will shown that when C_2 in equation (5) is close to zero the new descriptor Ψ_j is identical to original Φ^+, when C_2 increases Ψ_j declines from Φ^+ such that the measured distance between Ψ_j and negative examples increases.

This objective is being minimized in w_j and b_j. First, let us decompose w_j into parallel and orthogonal part w.r.t. Φ^+ and then let $C_2 \to 0$ and analyze the behavior of the convex objective. After decomposition of w_j such that $w_j = w_j^{\perp} + w_j^{||}$ the convex objective can be re-written as follows

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

$$||w_j^{\perp} + w_j^{||}||^2 + C_1 \cdot h\left(w_j^{||T} \Phi_j^+ + b_j\right) + C_2 \sum_{\Phi \in \mathcal{N}_j} h\left(-(w_j^{\perp} + w_j^{||})^T \Phi - b_j\right) \quad (6)$$

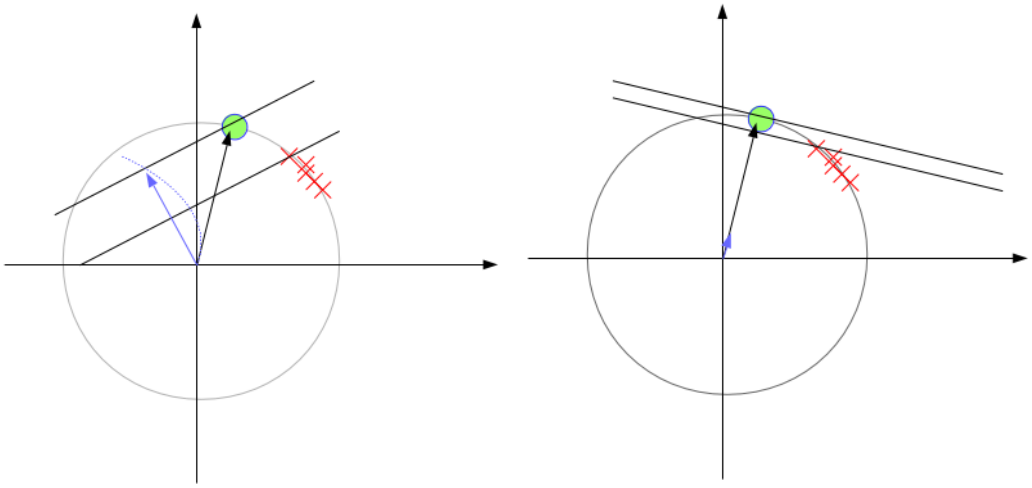
Notice that in the second term the orthogonal part does not change the value of a dot-product since $(w_j^{\perp} + w_j^{||})\Phi^+ = w_j^{||}\Phi^+$. Because $C_2 \to 0$ the first two terms dominate the equation (6), hence in the limit case it follows that

$$||w_j||^2 + C_1 \cdot h\left(w_j^T \Phi_j^+ + b_j\right) = ||w_j^{\perp} + w_j^{||}||^2 + C_1 \cdot h\left(w_j^{||T} \Phi_j^+ + b_j\right) \geq ||w_j^{||}||^2 + C_1 \cdot h\left(w_j^{||T} \Phi_j^+ + b_j\right) \quad (7)$$

Notice that the penalty terms on both sides of inequality (7) are the same since

$$w_j^T \Phi_j^+ + b_j = (w_j^{\perp} + w_j^{||})^T \Phi_j^+ + b_j = w_j^{||T} \Phi_j^+ + b_j$$

.



(a) (b)

Figure 1: An illustration of the effect of decreasing a parameter C_2. As the paramterer C_2 is decreased the resulting vector w is being parallel to Φ^+.

Interpretation, re-normalization. This result shows that as $C2$ decreases the normal of the hyperplane that separates the positive exemplar from negative data tends to be parallel with Φ^+. As C_2 grows, this normal declines such that distance from the negative data increases as shown in figure 1.

Since the goal is to measure the similarity between two descriptors by an angle, equation (4), we interpret the new descriptor Ψ_j as a normalized hypeplane normal such that

$$\Psi_j = \frac{w_j}{||w_j||} \quad (8)$$

.

4.1 Training data

The negative training data set contains hard negative examples of the image j. These hard negatives are database images that are spatially far-away from image j and, at the same time, have a high similarity score with image j measured by a dot-product of its FVs. This approach follows the simple idea of [18] that far-away images do not share the same visual content. A GPS information associated with each database image allows us to construct such a negative set for each image j in turn. Details are provided in experimental section 5.

The positive training data set is represented by the only image j itself. Why is that? It comes from the nature of the dataset. One could find additional positive examples by taking adjacent panorama images and building a graph where each node represents an image and each edge represents a visual overlap. However, due to the panorama sampling density during the capturing process of the Google Street View it is very rare to detect a visual overlap between two adjacent panoramas. Instead, very often the detected visual overlap is between the images that come from the same panorama and differ by a homography. We found that there is no benefit from taking these images into account.

5 Experimental evaluation

(TODO: Show calibrated BOW, why we are not showing calibrated FV, argue for it. Emphasize that the goal was achieved. Highlight its benefits!) In this section we first describe the experimental datasets, then we give implementation details and finally compare performance of the proposed approach on two datasets with several baseline methods.

5.1 Image datasets

(Another image retrieval datasets? INRIA Holliday, Sun, etc., it could be doable, but we have to find a FV competitor)
We performed our experiments on a database of Google Street View images from the Internet. For evaluation we have used two datasets. The first one is identical to [12, 37] ($25k$ of images), the latter one is an extension extension of this dataset to achieve a higher spatial density ($55k$ of images).

We downloaded panoramas from Pittsburgh (U.S.) covering roughly an area of 1.3×1.2 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.

As a query set with known ground truth GPS positions, we use the 8999 panoramas from the Google Street View research dataset, which cover approximately the same area, but were captured at a different time, and typically depict the same places from different viewpoints and under different illumination conditions. For each test panorama, we generate 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.

5.2 Implementation details

SIFT features: For all images in turn SIFT local descriptors [22] are extracted and subsequently these features are converted to a rootSIFT [2]. For extraction we use publicly available library `vlfeat` [**?**].

BOW baseline: A vocabulary of 100k visual words is learned by approximate k-means clustering [27] from a subset of features from 5,000 randomly selected images. A tf-idf [35] vector is computed for each image by assigning each descriptor to the nearest cluster center. Finally, all tf-idf vectors are normalized to have unit L_2 norm.

Fisher vector baseline: We first begin with decorrelation of local descriptors. We train a PCA matrix on a set of descriptors of 5,000 randomly selected database images and we project all root-SIFT descriptors into a $d = 64$ dimensional space. Then we learn a generative GMM that consist of $N = 256$ components. We train this model on a set of decorrelated descriptors from 5,000 randomly selected database images.

recall@K [%]	25k Pittsburgh					55k Pittsburgh				
Method:	1	2	5	10	20	1	2	5	10	20
BOW	22	24	31	34	38					
FV256	37	39	46	52	56					
svmFV256	40	42	55	59						
FV1024										
FV2048										
FV4096	etc....									

Table 1: Fake results! The percentage of correctly localized test queries for which the topK ranked database image is within 20 meters from the ground truth query position. The proposed method (svmFV) outperforms the baseline methods.

It follows by learning a PCA for FV decorrelation. The resulting dimension of full FV is $2Nd = 32,768$. To learn a reasonable non-singular PCA matrix we need at least this amount of data which we do not have for the first dataset which consists of only 25k images. For both databases we use all available data to train the PCA and we underline that for the first $25k$ dataset the projection to $d' > 25k$ is meaningless since the PCA matrix was singular.

Finally, to compute the FV of an image we (i) compute the rootSIFTs and perform the PCA decorrelation to $d = 64$ space, (ii) use the learned GMM and decorrelated descriptors to compute full FV (iii) project the full FV into lower dimensional spaces, namely to 2^β where $\beta = 7 \dots 14$ and (iv) re-normalize projected FVs by operator (3).

Learning vectors: To learn the linear regressor (**??**) for database image j, the positive and negative training data is constructed as follows. The *negative training set* $\mathcal{N}_j$ is obtained by: (i) finding the set of images with geographical distance greater than $200\mathrm{m}$; (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 very 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 only image j itself.

For SVM training we use `libsvm` [10] with $L2$ regularizer and $L2$-loss penaty fucntion $h(x)$ from equation (5). To obtain parameters C_1 and C_2 we perform a grid search and evaluate the performance on held out test set. We observe that for different FV target PCA dimensions the parameter C_1 is quite stable (typically $C_1 = 1$) while the optimal parameter for C_2 differs between 10^{-6} to 10^{-1}.

To learn the new image representation from FVs, for each database image j in turn we (i) learn SVM from $\mathcal{P}_j$ and N_j (ii) $L2$ normalize learned w_j (iii) use this vector as the new image representation. At query time we compute a FV Φ_q of the query image and measure its similarity score to each database image by equation (**??**).

Ground truth: Since all the query images have associated GPS location we can compute the spatial distance from database image. We consider query image to be correctly localized if its retrieved database image lies within a perimeter of $20m$ from the query image.

5.3 Results

For each database we compare our results to two baselines: a standard BOW baseline and a Fisher vector baseline. We measure our performance by recall@K metric which is the portion of query images that have at least one relevant image inside its ranked shortlist of the length K. We perform our experiments on several target FV dimensions as shown in the figure 2.

It is noticeable that proposed method outperforms all baselines and consistently improves the performance of the original FV over all lengths K of the shortlist and all target FV dimensions. This particularly means that learned descriptors are more likely to attract relevant images into top$K$ shortlist. The results are summarized in table 1.

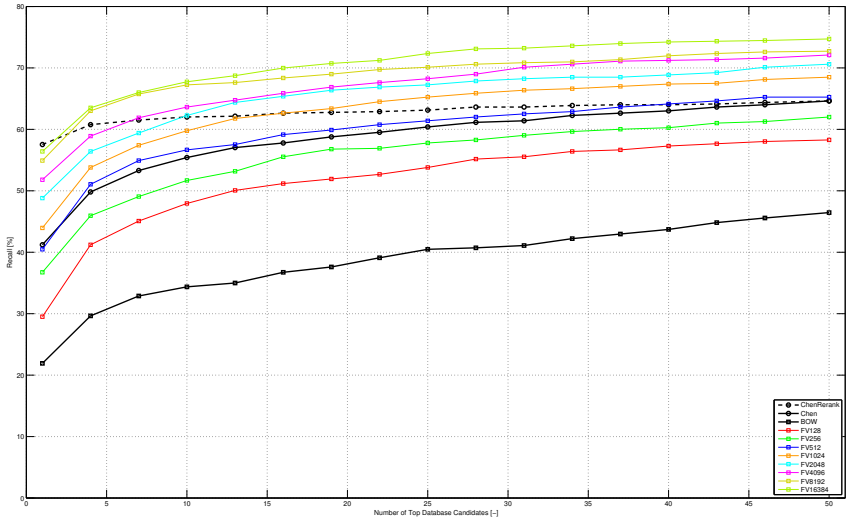


Figure 2: Fake results, actually SF, no embedded FV curves Recall curves for different methods

6 Conclusions

References

[1] Agarwal, S., Snavely, N., Simon, I., Seitz, S., Szeliski, R.: Building Rome in a day. In: ICCV, pp. 72–79 (2009)

[2] Arandjelović, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: IEEE PAMI (2012)

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

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

[5] Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV, pp. 1–22 (2004)

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

[7] Dalal, N., Triggs, B.: Histogram of oriented gradients for human detection. In: CVPR (2005)

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

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

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

[11] Gronat, P., Obozinski, G., Sivic, J., Pajdla, T.: Learning and calibrating per-location classifiers for visual place recognition. In: CVPR (2013)

[12] Gronat, P., Obozinski, G., Sivic, J., Pajdla, T.: Learning and calibrating per-location classifiers for visual place recognition. In: CVPR, pp. 907–914. IEEE (2013). URL `http://dblp.uni-trier.de/db/conf/cvpr/cvpr2013.html#GronatOSP13`

[13] Jégou, H., Douze, M., Schmid, C.: Product Quantization for Nearest Neighbor Search. IEEE PAMI **33**(1), 117–128 (2011). DOI 10.1109/TPAMI.2010.57. URL `http://hal.inria.fr/inria-00514462/en/`

[14] Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: CVPR 2010 - 23rd IEEE Conference on Computer Vision & Pattern Recognition, pp. 3304–3311. IEEE Computer Society, San Francisco, United States (2010). DOI 10.1109/CVPR.2010.5540039. URL http://hal.inria.fr/inria-00548637

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

[16] Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., Schmid, C.: Aggregating local image descriptors into compact codes. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(9), 1704–1716 (2012). DOI 10.1109/TPAMI.2011.235. URL http://hal.inria.fr/inria-00633013. QUAERO

[17] Klingner, B., Martin, D., Roseborough, J.: Street view motion-from-structure-from-motion. In: ICCV (2013)

[18] Knopp, J., Sivic, J., Pajdla, T.: Avoidng confusing features in place recognition. In: ECCV (2010)

[19] Krapac, J., Verbeek, J., Jurie, F.: Modeling Spatial Layout with Fisher Vectors for Image Categorization. In: ICCV 2011 - International Conference on Computer Vision, pp. 1487–1494. IEEE, Barcelona, Spain (2011). DOI 10.1109/ICCV.2011.6126406. URL http://hal.inria.fr/inria-00612277

[20] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)

[21] Li, Y., Snavely, N., Huttenlocher, D., Fua, P.: Worldwide pose estimation using 3d point clouds. In: ECCV (2012)

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

[23] Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-svms for object detection and beyond. In: ICCV (2011)

[24] Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: CVPR (2014)

[25] Perronnin, F., Dance, C.R.: Fisher kernels on visual vocabularies for image categorization. In: CVPR (2007)

[26] Perronnin, F., Liu, Y., Snchez, J., Poirier, H.: Large-scale image retrieval with compressed fisher vectors. In: CVPR, pp. 3384–3391. IEEE (2010). URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2010.html#PerronninLSP10

[27] Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR (2007)

[28] Sattler, T., Leibe, B., Kobbelt, L.: Improving image-based localization by active correspondence search. In: ECCV (2012)

[29] Schindler, G., Brown, M., Szeliski, R.: City-scale location recognition. In: CVPR (2007)

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

[31] Shrivastava, A., Malisiewicz, T., Gupta, A., Efros, A.A.: Data-driven visual similarity for cross-domain image matching. In: SIGGRAPH ASIA (2011)

[32] Simonyan, K., Parkhi, O.M., Vedaldi, A., Zisserman, A.: Fisher Vector Faces in the Wild. In: British Machine Vision Conference (2013)

[33] Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: ECCV (2012)

[34] Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: ICCV (2003). URL http://www.robots.ox.ac.uk/~vgg

[35] Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: ICCV (2003). URL http://www.robots.ox.ac.uk/~vgg

[36] Tighe, J., Lazebnik, S.: Finding things: Image parsing with regions and per-exemplar detectors. In: CVPR (2013)

[37] Torii, A., Sivic, J., Pajdla, T.: Visual localization by linear combination of image descriptors. In: ICCV Workshops, pp. 102–109. IEEE (2011). URL http://dblp.uni-trier.de/db/conf/iccvw/iccvw2011.html#ToriiSP11

[38] Torii, A., Sivic, J., Pajdla, T., Okutomi, M.: Visual place recognition with repetitive structures. In: CVPR (2013)

[39] Zamir, A., Shah, M.: Accurate image localization based on google maps street view. In: ECCV (2010)