

000  
001  
002  
003  
004  
005  
006  
007  
008  
009  
010  
011  
012  
013  
014  
015  
016  
017  
018  
019  
020  
021  
022  
023  
024  
025  
026  
027  
028  
029  
030  
031  
032  
033  
034  
035  
036  
037  
038  
039  
040  
041  
042  
043  
044  
045  
046  
047  
048  
049  
050  
051  
052  
053

054  
055  
056  
057  
058  
059  
060  
061  
062  
063  
064  
065  
066  
067  
068  
069  
070  
071  
072  
073  
074  
075  
076  
077  
078  
079  
080  
081  
082  
083  
084  
085  
086  
087  
088  
089  
090  
091  
092  
093  
094  
095  
096  
097  
098  
099  
100  
101  
102  
103  
104  
105  
106  
107

# Fisher vector places: learning compact descriptors for place recognition

Anonymous CVPR submission

Paper ID 954

## 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. Secondly, 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. Finally, 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 a 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 a large-scale instance-level retrieval [6, 4, 12, 17, 27, 34, 35], where images are represented using local invariant features [21] aggre-

108 gated into the bag-of-visual-words [5, 32] representation.  
 109 The image database can be further augmented by 3D point  
 110 clouds [16], automatically reconstructed by large-scale  
 111 structure from motion (SfM) [1, 16], which enables accurate  
 112 prediction of query image camera position [20, 25].  
 113 In this work we investigate learning a discriminative  
 114 representation using the compact Fisher vector descriptors [15].  
 115 Fisher vector descriptors have shown excellent place recog-  
 116 nition accuracy [34]. In this work we further improve their  
 117 performance by discriminative learning.  
 118

119  
 120 **Fisher vector image representations** Fisher vector im-  
 121 age representations have recently demonstrated excellent  
 122 performance for a number of visual recognition tasks [3,  
 123 15, 18, 30]. They are specially suited for retrieval applica-  
 124 tions since they are robust to image appearance variations  
 125 and capture richer image statistics than the simple bag-of-  
 126 visual-words (BOW) aggregation. However, the raw ex-  
 127 tracted Fisher vectors are typically high-dimensional, e.g.  
 128 with 32,768 non-sparse dimensions, which is impractical  
 129 for large-scale visual recognition and indexing applications.  
 130 Hence, their dimensionality is often reduced by principal  
 131 component analysis (PCA) and further quantized for effi-  
 132 cient indexing using, e.g. a product quantizer [15]. Other  
 133 recent work has demonstrated improved performance in a face  
 134 recognition application by finding discriminative projection  
 135 using a large amount of training face data [30]. Our work  
 136 is complementary to these methods as it operates on the  
 137 projected low-dimensional descriptor and further learns dis-  
 138 criminative re-weighting of the descriptor specific to each  
 139 image in the database using per-exemplar support vector  
 140 machine [22].  
 141

142  
 143 **Per-exemplar support vector machine** The exemplar  
 144 support vector machine (e-SVM) has been used in a number  
 145 of visual recognition tasks including category-level recog-  
 146 nition [22], cross-domain retrieval [29], scene parsing [33],  
 147 place recognition [12] or as an initialization for more com-  
 148 plex discriminative clustering models [8, 31]. The main  
 149 idea is to train a linear support vector machine (SVM) clas-  
 150 sifier from a single positive example and a large number of  
 151 negatives. The intuition is that the resulting weight vector  
 152 will give a higher weight to the discriminative dimensions  
 153 of the positive training data point and will down weight  
 154 dimensions that are non-discriminative with respect to the  
 155 negative training data. A key advantage is that each per-  
 156 exemplar classifier can be trained independently and hence  
 157 the learning can be heavily parallelized. The per-exemplar  
 158 training brings however also an important drawback. As  
 159 each classifier is trained independently a careful calibration  
 160 of the resulting classifier scores is required [12, 22].  
 161

162 **Contributions** The contributions of this work are three-  
 163 fold. First, we analyze the exemplar support vector machine  
 164 objective and show that the learnt hyperplane can be inter-  
 165 preted as a new descriptor that replaces the original positive  
 166 example and is re-weighted to increase its separation from  
 167 the negative data. Secondly, we demonstrate that after an  
 168 appropriate normalization of the new re-weighted descriptor  
 169 no further calibration is necessary. Finally, we apply  
 170 e-SVM training to compact Fisher vector descriptors for  
 171 large-scale place recognition resulting in a *discriminative*  
 172 yet *compact* representation of each image in the database.  
 173 Place recognition results are shown on a dataset of 25k im-  
 174 ages of Pittsburgh and demonstrate the learnt representation  
 175 consistently improves over the standard Fisher vector de-  
 176 scriptors at different target dimensions.  
 177

### 3. Learning compact place descriptors using per-exemplar SVM

178 Each database image  $j$  is represented by its L2-  
 179 normalized Fisher vector  $\Phi_j$ . The goal is to learn a set  
 180 of new L2-normalized Fisher vectors  $\Psi_j$ , one per each  
 181 database image. At query time, given the Fisher vector  $\Phi_q$   
 182 of an unknown query image, we retrieve the database image  
 183 depicting the same location by finding the image  $j^*$  with the  
 184 highest score measured by a dot product  
 185

$$j^* = \arg \max_j \Phi_q^T \Psi_j. \quad (1)$$

186 In other words, the aim is to replace each original database  
 187 Fisher vector  $\Phi_j$  with a new vector  $\Psi_j$  that is more dis-  
 188 criminative in the sense of separation from descriptors of  
 189 images depicting other places. Inspired by [12], we in-  
 190 vestigate applying the exemplar support vector machine (e-  
 191 SVM) [22] for this task. e-SVM learns a linear classifier  
 192  $w_j^T \Phi + b_j$  given the descriptor  $\Phi_j^+$  of place  $j$  as a single  
 193 positive example (with target label +1) and a large number  
 194 of negative descriptors  $\mathcal{N}_j$  from other places in the database  
 195 (with target labels -1). The intuition of the exemplar SVM  
 196 training [22] is that the learnt weight vector  $w_j$  will give a  
 197 higher weight to the dimensions of the descriptor that are  
 198 discriminative and will down-weight dimensions that are  
 199 non-discriminative with respect to the negative training data  
 200 collected from other far-away places. The optimal  $w_j$  and  
 201  $b_j$  are obtained by minimizing the following objective  
 202

$$\begin{aligned} & \|w_j\|^2 + C_1 \cdot h(w_j^T \Phi_j^+ + b_j) \\ & + C_2 \sum_{\Phi \in \mathcal{N}_j} h(-w_j^T \Phi - b_j), \end{aligned} \quad (2)$$

203 where  $\Phi_j^+$  is the descriptor of the place  $j$  as the positive data  
 204 point,  $\Phi$  are Fisher descriptors from negative training data  
 205  $\mathcal{N}_j$  and  $h$  is the hinge loss,  $h(y) = \max(1 - y, 0)$ . Note  
 206 that the first term in (2) is the regularizer, the second term  
 207

216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
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  $\Psi_j$  that replaces the original positive training descriptor  $\Phi_j^+$ . In particular, we show first that when the weight  $C_2$  of the negative data in objective (2) goes to zero and the learnt  $\Psi_j$  is identical to the original positive training data point  $\Phi_j^+$ . Second, when  $C_2 > 0$ , the learnt  $\Psi_j$  moves away from the positive  $\Phi_j^+$  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 vector of positive training descriptor  $\Phi_j^+$ . 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  $\Phi^+$  (in the following we omit index  $j$  for brevity), i.e.  $w = w^\perp + w^\parallel$ , 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  $w^\parallel$ , i.e. the optimal  $w$  is parallel with  $\Phi^+$ .

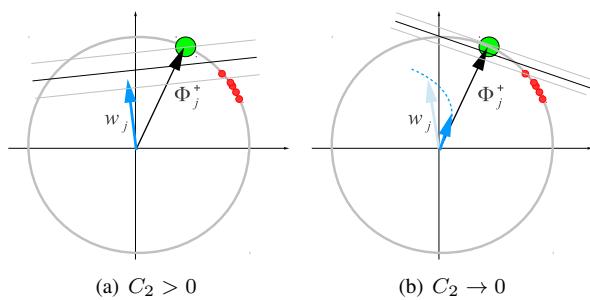
In detail, for  $w = w^\perp + w^\parallel$  the objective (2) can be written as

$$\begin{aligned} & \|w^\perp + w^\parallel\|^2 + C_1 \cdot h((w^\perp + w^\parallel)^T \Phi^+ + b) \\ & + C_2 \sum_{\Phi \in \mathcal{N}} h(-(w^\perp + w^\parallel)^T \Phi - b). \end{aligned} \quad (3)$$

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

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

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



270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
Figure 1: An illustration of the effect of decreasing parameter  $C_2$  in the exemplar support vector machine objective. The positive exemplar  $\Phi_j^+$  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  $\Phi_j^+$  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  $\Phi_j^+$  and its magnitude  $\|w_j\|$  goes to 0.

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  $\|w^\perp + w^\parallel\|^2 \geq \|w^\parallel\|^2$  for any  $w^\perp \neq 0$ .

As a result, in the limit case when  $C_2 \rightarrow 0$  the optimal  $w$  is parallel with  $\Phi^+$ . 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.

**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  $w^\parallel$  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  $\sum_{\Phi \in \mathcal{N}} \max(1 + w^\parallel T \Phi + b, 0)$ , i.e. that reduces the dot product  $w^\parallel T \Phi$  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  $\Phi^+$  from negative data becomes parallel with  $\Phi^+$ , 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  $\Phi^+$  in a manner that reduces the loss on the negative data. This suggests that the learnt  $w$

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

$$\Psi_j = \frac{w_j}{\|w_j\|}. \quad (5)$$

## 335 4. Experimental evaluation

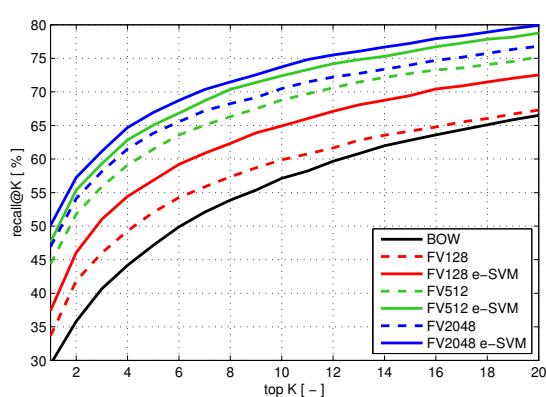
336 In this section we first describe the experimental set-up,  
 337 then we give implementation details, and finally report  
 338 results of the proposed approach on two datasets where we  
 339 compare performance with raw Fisher vector matching and  
 340 several baselines methods.

### 342 4.1. Experimental set-up

343 We perform experiments on a database of Google Street  
 344 View images of Pittsburgh downloaded from the Internet.  
 345 The data contains panoramas covering roughly an area of  
 346  $1.3 \times 1.2 \text{ km}^2$ . Similar to [4], for each panorama we  
 347 generate 12 overlapping perspective views corresponding  
 348 to two different elevation angles to capture both the street-  
 349 level scene and the building façades, resulting in a total of  
 350 24 perspective views each with  $90^\circ$  FOV and resolution of  
 351  $960 \times 720$  pixels. For evaluation we have used two versions  
 352 of this data. The first one was obtained from the authors  
 353 of [12] (25k images). In the second version we download  
 354 additional images to increase the dataset size to 55k images.  
 355 As a query set with known ground truth GPS positions, we  
 356 use the 8999 panoramas from the Google Street View re-  
 357 search dataset. This dataset covers approximately the same  
 358 area, but has been captured at a different time, and depicts  
 359 the same places from different viewpoints and under differ-  
 360 ent illumination conditions. For each test panorama, we  
 361 generate a set of perspective images as described above. Fi-  
 362 nally, we randomly select out of all generated perspective  
 363 views a subset of 4k images, which is used as a test set to  
 364 evaluate the performance of the proposed approach. Since  
 365 all the query images have associated GPS location we can  
 366 compute their spatial distance from the database images re-  
 367 turned by the matching method. We consider a query image  
 368 to be correctly localized if the retrieved database image lies  
 369 within a perimeter of 20m from the location of the query.  
 370

### 371 4.2. Implementation details

372 We first extract rootSIFT descriptors [2] for each im-  
 373 age. Following [15] we project the 128-dimensional SIFT  
 374 descriptors to 64 dimensions using PCA. The projection  
 375 matrix is learnt on a set of descriptors from 5,000 ran-  
 376 domly selected database images. This has also the effect  
 377



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

400 of decorrelating the SIFT descriptor. The 64-dimensional  
 401 SIFT descriptors are then aggregated into Fisher vectors us-  
 402 ing a Gaussian mixture model with  $N = 256$  components,  
 403 which results in a  $2 \times 256 \times 64 = 32,768$ -dimensional  
 404 descriptor for each image. The Gaussian mixture model  
 405 is learnt from descriptors extracted from 5,000 randomly  
 406 sampled database images. The high-dimensional Fisher  
 407 vector descriptors are then projected down to dimension  
 408  $d \in \{128, 512, 2048\}$  using PCA learnt from all avail-  
 409 able images in the database. The resulting low dimensional  
 410 Fisher vectors are then re-normalized to have unit L2-norm,  
 411 which we found to be important in practice.

413 **Learning parameters and training data.** To learn the  
 414 exemplar support vector machine for each database image  
 415  $j$ , the positive and negative training data are constructed as  
 416 follows. The *negative training set*  $\mathcal{N}_j$  is obtained by: (i)  
 417 finding the set of images with geographical distance greater  
 418 than 200m; (ii) sorting the images by decreasing value of  
 419 similarity to image  $j$  measured by the dot product between  
 420 their respective Fisher vectors; (iii) taking the top  $N = 500$   
 421 ranked images as the negative set. In other words, the neg-  
 422 ative training data consists of the hard negative images, i.e.  
 423 those that are similar to image  $j$  but are far away from its  
 424 geographical position, hence, cannot have the same visual  
 425 content. The *positive training set*  $\mathcal{P}_j$  consist of the original  
 426 Fisher vector  $\Phi_j$  of the image  $j$ . For the SVM training we  
 427 use libsvm [10]. We use the same  $C_1$  and  $C_2$  parameters  
 428 for all per-exemplar classifiers, but find the optimal value of  
 429 the parameters for each dimensionality of the Fisher vector  
 430 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  $\Psi_j$  for image  $j$ . At query time we compute the Fisher vector  $\Phi_q$  of the query image and measure its similarity score to the learnt descriptors  $\Psi_j$  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,

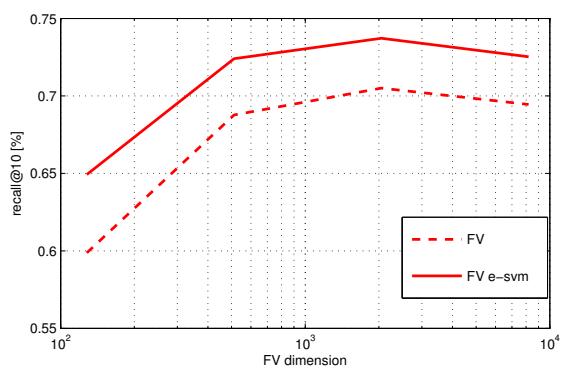


Figure 3: **Memory complexity analysis.** The fraction of correctly localized queries at the top 10 retrieved images (y-axis) for different Fisher vector dimensions (x-axis). The learnt descriptors by our method (FV e-SVM) clearly outperform the raw Fisher vector descriptors (FV) for all dimensions.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

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.

## 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					702
	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	703
BOW raw SVM	6.4	8.1	13.5	17.5	20.5						704
BOW e-SVM	32.2	37.3	46.2	54.4	63.3						705
FV128	33.6	41.8	52.0	59.8	67.7	10.9	14.1	20.2	26.4	33.2	706
FV128 raw SVM	32.4	41.1	52.6	60.4	68.4						707
<b>FV128 e-SVM</b>	<b>37.8</b>	<b>46.1</b>	<b>56.9</b>	<b>64.9</b>	<b>72.6</b>	<b>13.5</b>	<b>17.7</b>	<b>25.0</b>	<b>31.8</b>	<b>39.0</b>	708
FV512	44.3	51.7	61.4	68.7	75.2	17.3	21.1	28.4	34.2	40.3	709
<b>FV512 e-SVM</b>	<b>47.6</b>	<b>55.4</b>	<b>65.1</b>	<b>72.4</b>	<b>78.8</b>	<b>19.8</b>	<b>25.1</b>	<b>32.7</b>	<b>38.7</b>	<b>46.0</b>	710
FV2048	46.9	54.1	63.8	70.5	76.8	19.2	23.5	29.9	35.2	41.9	711
<b>FV2048 e-SVM</b>	<b>50.2</b>	<b>57.3</b>	<b>67.0</b>	<b>73.8</b>	<b>78.0</b>	<b>20.8</b>	<b>25.9</b>	<b>33.1</b>	<b>38.7</b>	<b>45.9</b>	712

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. Avoiding 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

- 756 [29] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros. 810  
757 Data-driven visual similarity for cross-domain image match- 811  
758 ing. In *SIGGRAPH ASIA*, 2011. 2 812  
759 [30] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. 813  
760 Fisher Vector Faces in the Wild. In *British Machine Vision 814  
761 Conference*, 2013. 2 815  
762 [31] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery 816  
763 of mid-level discriminative patches. In *ECCV*, 2012. 2 817  
764 [32] J. Sivic and A. Zisserman. Video Google: A text retrieval 818  
765 approach to object matching in videos. In *ICCV*, 2003. 2 819  
766 [33] J. Tighe and S. Lazebnik. Finding things: Image parsing with 820  
767 regions and per-exemplar detectors. In *CVPR*, 2013. 2 821  
768 [34] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi. Visual place 822  
769 recognition with repetitive structures. In *CVPR*, 2013. 1, 2 823  
770 [35] A. Zamir and M. Shah. Accurate image localization based 824  
771 on google maps street view. In *ECCV*, 2010. 1 825  
772 826  
773 827  
774 828  
775 829  
776 830  
777 831  
778 832  
779 833  
780 834  
781 835  
782 836  
783 837  
784 838  
785 839  
786 840  
787 841  
788 842  
789 843  
790 844  
791 845  
792 846  
793 847  
794 848  
795 849  
796 850  
797 851  
798 852  
799 853  
800 854  
801 855  
802 856  
803 857  
804 858  
805 859  
806 860  
807 861  
808 862  
809 863