Patch-level Spatial Layout for Classification and Weakly Supervised Localization

Valentina Zadrija[1], Josip Krapac[1], Jakob Verbeek[2], and Siniša Šegvić[1]

[1] University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia
[2] LEAR, INRIA Grenoble, France

Abstract. We propose a discriminative patch-level model which combines appearance and spatial layout cues. We start from a block-sparse model of patch appearance based on the normalized Fisher vector representation. The appearance model is responsible for i) selecting a discriminative subset of visual words, and ii) identifying distinctive patches assigned to the selected subset. These patches are further filtered by a sparse spatial model operating on a novel representation of pairwise patch layout. We have evaluated the proposed pipeline in image classification and weakly supervised localization experiments on a public traffic sign dataset. The results show significant advantage of the combined model over state of the art appearance models.

1 Introduction

Detecting the presence and precise locations of objects in images is a fundamental problem in computer vision. Best results are achieved with strongly supervised training [6, 10, 16, 24] where object locations have to be annotated with bounding boxes. However, the annotation process is difficult, time-consuming and error-prone, especially when the objects are small. These problems are alleviated in weakly supervised localization which learns from image-wide labels only.

Most previous work on weakly supervised learning for object localization follows the multiple instance learning [2, 27] approach (MIL) in order to account for the missing ground truth locations [7, 8, 12, 31, 36]. MIL iteratively trains an instance classifier on bags of instances. A positive bag contains at least one positive instance, while negative bags contain only negative instances. Bags correspond to images, while instances in the bags are tentative object locations.

However, the localization problem can also be expressed as a search for patches which contribute most to the image classification score. We have previously shown [22] that traffic signs can be localized by a sparse linear model trained on non-normalized Fisher vectors (FV) of entire images. In this paper we present two contributions which further improve these results. First, we propose to approximate the patch contribution to the normalized FV score with the first-order Taylor expansion. This allows to improve the patch appearance model by training it on the normalized FVs. Second, we propose a novel spatial representation of the pairwise patch layout. This representation captures distinctive spatial configurations between the visual words selected by the appearance model. The interplay between the two models is illustrated in Fig. 1.

Fig. 1. The appearance model identifies distinctive patches (enclosed in yellow rectangles) assigned to selected visual words (shown in colors). The spatial model learns consistent spatial configurations between pairs of selected visual words, e.g. between a_1 (purple) and a_2 (cyan), a_1 (purple) and a_3 (green) and so on.

2 Related Work

Most of the existing weakly-supervised localization approaches mitigate the computational complexity by relying on bottom-up location proposals. Unfortunately, this risks to overlook true object patches, which is especially pertinent in traffic scenes with small objects and rich backgrounds. In our preliminary experiments, a popular objectness algorithm [1] failed to produce accurate traffic sign locations in top 2000 proposals. Due to recent success of cascaded classifiers [28, 39], strongly supervised traffic sign localization is considered solved today. However, due to greedy training, these approaches have a limited feature sharing potential, and none of them is able to detect all kinds of traffic signs at once. Thus, current research and commercial products typically disregard important classes such as the stop sign, priority road, no entry etc.

Modeling co-occurrence of visual words has been of interest ever since the introduction of the bag-of-words (BoW) image classification paradigm [9]. Most previous research considered unordered co-occurrence patterns of particular visual words. The discovery of such patterns can be cast as a frequent pattern mining problem, where BoW histograms are viewed as transactions while co-occurring tuples of visual words correspond to frequent patterns [40] or itemsets [17, 42]. Many approaches attempt to discover co-occurrence patterns in an unsupervised setting, and to use these patterns to augment the BoW representation [17, 42] or to supply weak classifiers for boosting [40]. Recent work suggests that better performance can be obtained in a supervised discriminative context, by employing so called jumping emerging patterns [38]. This relation between frequent and discriminative patterns in data mining is similar to the generative-discriminative dichotomy in computer vision classification models.

The second line of research goes beyond simple co-occurrence and attempts to model spatial constellations of visual words. The approach by Lin et al. [25] uses histograms to represent the spatial layout of pairs of visual words. A major problem with this approach is stability. Many pairs of features may be needed to represent a given trait of an object class, since several visual words typically fire in any discriminative image region. Due to use of histograms, this approach

may require large training datasets in order to properly model discretization issues. The approach by Yang et al. [41] deals with these problems by choosing a small dictionary of 100 visual words, and by considering crude spatial predicates of proximity and orientation. Singh et al. [35] present an interesting iterative approach for selection of discriminative visual words. In each round of learning and for each visual word, a discriminative classifier is trained on the first fold of training data. The classifiers are subsequently applied to the second fold, and the positive responses are clustered to define the visual words for the next round (this procedure is similar to the multi-fold multiple-instance learning for weakly supervised localization [7]). Pairs of spatially correlated visual words (doublets) are greedily discovered in the postprocessing phase, which provides a slight increase in classification performance.

In this work, we present an approach for learning the spatial layout of visual word pairs, which is suitable for classification and weakly supervised localization. In contrast with [25, 41, 35], we perform a globally optimal selection of visual words from a large dictionary. The selection procedure is optimal in the sense of image classification performance over the Fisher vector representation. Our approach does not rely on bottom-up location proposals such as segmentation [18, 6, 7] or objectness [36, 12]. Due to generative front-end, we have a better sharing potential than pure discriminative approaches used in [28]. In contrast with [22], we use block-sparsity [21], the normalized score gradient and the spatial model of the pairwise layout.

Our appearance model is able to provide two-fold filtering of patches from the test image. The filtering procedure discards (i) patches which are not assigned to the selected set of visual words, and (ii) patches with a negative contribution to the classification score. The filtered patches are further tested by the spatial model based on [23], which improves the performance by considering pairwise spatial relations in a local neighbourhood. Our approach is non-iterative and therefore provides potential for combining with other approaches [17, 35].

3 Selecting Discriminative Visual Words

We regard images as bags of visual words and represent them with a normalized FV embedding built atop the generative Gaussian mixture model (GMM) of patch appearance. Two types of FV normalizations are widely used to improve the performance in this setup [32]. The power normalization is applied to each dimension X_d of the FV as $s(X_d) = \mathrm{sign}(X_d)|X_d|^\rho$, with $0 < \rho < 1$. This "unsparsifies" the vector $\mathbf{X}$ and makes it more suitable for comparison with the dot product. The metric normalization projects the FV onto the unit hyper-sphere by dividing it by its ℓ_2 norm. This accounts for the fact that different images contain different amounts of background information. The ℓ_2 normalization is applied by dividing the power-normalized FV $s(\mathbf{X})$ with $\sqrt{n(\mathbf{X})}$ where $n(\mathbf{X}) = \sum_d s(X_d)^2$.

In our work, we use the intra-component normalization [3] where the ℓ_2 normalization is separately applied to the components of the FV corresponding to different visual words. This accounts for the effect of "burstiness" [19] where a

few large components of the FV can dominate the similarity computed towards another FV. In order to formally define the intra-normalized FV of the image, we use $\mathbf{X_k}$ to denote the part of the FV corresponding to the k-th visual word and write the corresponding ℓ_2 norm as $n(\mathbf{X_k})$.

We train our appearance model from image-wide training labels y_i as a linear classifier $\mathbf{w}$ which minimizes the following regularized logistic loss function:

$$\ell(\mathbf{w}, \mathbf{X}, \mathbf{y}) = \sum_{i=1}^{N} \log\left(1 + \exp(y_i \cdot \mathbf{w}^\top \mathbf{X}_i)\right) + \lambda \cdot \mathcal{R}(\mathbf{w}) \; . \tag{1}$$

In the above equation, N denotes the number of the training examples, $\mathcal{R}$ denotes the regularizer, while the parameter λ represents a trade-off between the loss and the regularization. We prefer a sparse regularizer because it (i) alleviates the high dimensionality of the FV and (ii) performs a globally optimal feature selection within the learning algorithm. The most commonly used choice for this purpose is the ℓ_1 norm [30]. However this would ignore the specific FV structure induced by the blocks that correspond to different visual words. In order to provide better regularization, we capture this structure by using the $\ell_{2,1}$ norm [20, 43]: $\mathcal{R}(w) = \lambda \sum_k \|\mathbf{w}_k\|$, where k denotes visual words. This acts like the lasso at the group level: depending on the choice of λ, all coefficients corresponding to the particular visual word are set to zero. Note that block sparsity favours the selection of discriminative visual words, which is especially helpful in weakly supervised localization and fine-grained classification [21]. The main benefits include faster execution (many patches can be discarded without applying the model) and better performance due to reduced overfitting.

4 Gradient of the Classification Score

For the purpose of image classification, we denote the score of the full-image FV descriptor as $f(\mathbf{X})$, and the contribution of the patch $\mathbf{x}$ as $f(\mathbf{X}) - f(\mathbf{X} - \mathbf{x})$. In the case of un-normalized FV representation, the contribution of local features to the final classification score can be easily derived. The linearity of the classifier and the sum-pooling of the encoding of the local features makes that the scoring and pooling can be reversed, i.e. $\mathbf{w}^T \cdot \mathbf{X} = \sum_i \mathbf{w}^T \cdot \mathbf{x}_i$. As a result, we obtain the patch contribution using a simple dot product with the model [22].

On the other hand, the score of the normalized image FV corresponds to[3] $f(\mathbf{X}) = \mathbf{w}^\top \cdot s(\mathbf{X})/\sqrt{n(\mathbf{X})}$. Due to the non-linear normalizations, the above linear decomposition of the image score into patch scores is no longer possible. The patch contribution could be computed directly as $f(\mathbf{X}) - f(\mathbf{X} - \mathbf{x})$. However, that would require for each patch $\mathbf{x}$ to subtract it from $\mathbf{X}$, apply power and ℓ_2 normalizations to the $\mathbf{X} - \mathbf{x}$, and finally to score it with the classifier and subtract it from $f(\mathbf{X})$. A computationally more efficient approach is to approximate the contribution to the score by using the gradient $\nabla_{\mathbf{x}} f(\mathbf{X})$ of the score

[3] For the sake of simplicity, we assume the global ℓ_2 normalization $n(X)$. We later show the proposed reasoning also holds in the case of the intra- ℓ_2 normalization.

w.r.t. the unnormalized FV $\mathbf{x}$. The dot-product of the local FV with this gradient $\langle \mathbf{x}, \nabla_{\mathbf{x}} f(\mathbf{X}) \rangle$ then approximates the impact of a local descriptor on the final classification score. Let us now derive the gradient of the classification score w.r.t. the non-normalized FV. The partial derivative of the score w.r.t. an element of the non-normalized patch x_d is given by $\partial f(\mathbf{X})/\partial x_d = \partial f(\mathbf{X})/\partial \mathbf{X} \cdot \partial \mathbf{X}/\partial x_d$. The derivative of the non-normalized image FV w.r.t. the d-th element of the patch FV corresponds to the vector with all zero elements except the d-th, which is equal to one. Hence, the gradient w.r.t. the patch element x_d is equal to the gradient w.r.t. an image element X_d:

$$\frac{\partial f(\mathbf{X})}{\partial x_d} = \frac{\partial f(\mathbf{X})}{\partial X_d} = \frac{\rho |X_d|^{\rho-1}}{\sqrt{n(\mathbf{X})}} \left(w_d - \frac{s(X_d) f(X)}{\sqrt{n(\mathbf{X})}} \right) . \tag{2}$$

Please note that this derivative is undefined for $X_d = 0$. In practice, we set the derivative to zero in this case, to ignore the impact of such dimensions.

In the case of per-component intra-normalization, the classification score is a sum of per-component classification scores: $f(\mathbf{X}) = \sum_k f_k(\mathbf{X}_k)$. Since the $f_k(\mathbf{X}_k)$ have precisely the same form as $f(\mathbf{X})$ above, we can compute the gradients in the same manner, per block. Note that the gradient of the intra-normalized FV preserves the sparsity (i.e. the zero elements) of the original model (1). This is not the case for the global ℓ_2 normalization, where the gradient sparsity depends on the difference between the model $\mathbf{w}$ and the normalized FV multiplied with the score. A fixed set of visual words makes the gradient of the intra-normalized FV more suitable for the construction of the spatial layout.

5 Spatial Layout Model

The proposed patch appearance model reduces the number of possible object locations by an order of magnitude (e.g. from 100000 to 7000). Still, some of the difficult background patches are scored positively and as such generate false alarms (see Fig. 1). One way to address this problem is to learn a distinctive spatial layout between the patches corresponding to different visual words. We assume that the *soft-assign* distribution is sharply peaked, i.e. each local feature is assigned to a single GMM component (see Fig. 2) [33, 34]. The appearance model identifies K_w discriminative components $\{a_i\}_{i=1}^{K_w}$ from the GMM vocabulary of the size K, where $K_w \ll K$. For each positively scored patch p assigned to some visual word a_i, we consider a square neighbourhood upon which we construct the spatial descriptors. The spatial features are based on displacement vectors $\mathbf{d}(p, q)$ between the central patch p and neighbouring patches q. We aggregate the spatial descriptors over the whole image and train an ℓ_1 regularized model using image-wide labels. In the evaluation stage, the spatial model is applied only to patches which are positively scored by the appearance model.

We experiment with two types of descriptors: (i) spatial histograms (SH) [25], and (ii) spatial Fisher vectors (SFV) [23]. The SHs are constructed as follows. For each pair of the visual words, we construct a 2D histogram by discretizing

the local neighbourhood into b bins over both axes. The displacement vectors $\mathbf{d}(p,q)$ are assigned to the appropriate bins, to which they contribute with the appearance score of the patch q. The dimensionality of the 2D histogram is b^2, and since there are K_w^2 possible pairs, the size of the final SH is $K_w^2 \cdot b^2$.

We construct the SFVs as follows (cf. Fig. 2). For each visual word pair (a_i, a_j) we assume a distinct spatial GMM with K_s components and diagonal covariance. For each patch p assigned to a_i we aggregate the weighted contributions $\boldsymbol{\phi}(p,q)$ of all neighbouring patches q assigned to a_j into the SFV component $\boldsymbol{\Phi}_{a_i,a_j}$. We incorporate the appearance information by weighting the contributions $\boldsymbol{\phi}(p,q)$ with the appearance score $f(q)$. This is similar to [6] where the segmentation masks are used to weight the Fisher vectors of the candidate windows. The final SFV is obtained by concatenating the $\boldsymbol{\Phi}_{a_i,a_j}$ for all (a_i, a_j). Each of the $\boldsymbol{\Phi}_{a_i,a_j}$ is of dimension $5K_s$, since for each of the K_s spatial Gaussians it concatenates 2D gradients for its mean and variance, as well as one dimension for its mixing weight. The SFV dimensionality is $K_w^2 \cdot K_s \cdot (2D+1) = K_w^2 \cdot K_s \cdot 5$.

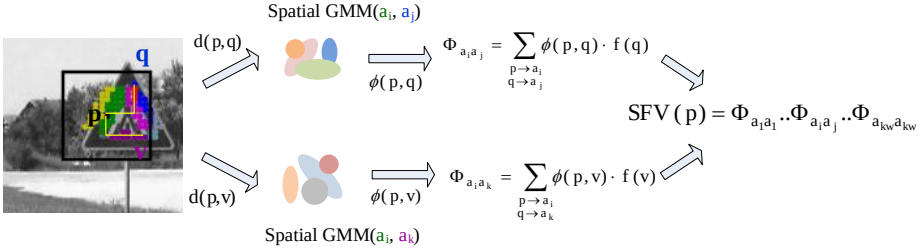

Fig. 2. SFV derivation for the local neighbourhood around the patch p (black rectangle) with $K_s = 4$. Patches assigned to different visual words are shown in different colors. The patches p, q and v are assigned to visual words a_i, a_j, a_k. The SFV contribution $\boldsymbol{\phi}(p,q)$ is determined as the gradient of the log likelihood of the displacement $\mathbf{d}(p,q)$ with respect to the spatial GMM component (a_i, a_j).

6 Experiments

Dataset. We evaluate the proposed approach on a public traffic sign dataset [5]. The dataset contains 3296 images acquired from the driver's perspective along local countryside roads. We focus on triangular warning signs (30 different types). We train our classifiers using image-wide labels on the training split with 453 positive and 1252 negative images. The train and test splits are disjoint: images containing the same physical traffic sign are assigned to the same split. In general, the dataset contains very small objects taking approximately 1% of the image area making the classification and weakly supervised localization difficult. We perform the bounding box evaluation as proposed in [15] and use the average precision (AP) as the performance measure.

Implementation details. We extract dense 128-dimensional SIFT descriptors over square patches being 16, 24, 32 and 40 pixels wide, with the stride of 1/8 patch width. The descriptors are ℓ_2 normalized and projected onto a 80-dimensional PCA subspace. We train a GMM vocabulary with $K = 1024$ components and diagonal covariance with EM, as implemented in Yael [14]. The resulting appearance FV is 164864-dimensional. We train our classifiers by optimizing the logistic loss with block sparse regularization by proximal gradient descent (FISTA), as implemented in SPAMS [26]. The regularization parameter λ is determined using 10-fold cross-validation for all presented experiments.

We build local spatial layout descriptors by considering a neighbourhood 4 times larger than the corresponding reference patch. We construct the spatial 2D histograms by discretizing the patch neighbourhood into 8 bins per each axis. As a result, for each pair of visual words, we obtain a 64 dimensional descriptor. We construct the spatial Fisher vectors over a fixed GMM with $K_s = 4$ components shared across all visual word pairs. The mean and the variances of the components match the first and second order moments of the uniform distribution over the four quadrants of the unit square [23]. The dimensionality of the SFV descriptor is $K_s \cdot (2 \cdot 2 + 1) = 20$ per each pair of visual words.

Classification. We apply the proposed spatial layout model through the following stages: (1) extract the dense SIFT descriptors and determine their Fisher vectors, (2) apply the power normalization and ℓ_2 intra-component normalization, (3) identify positive patches by employing the gradient of the appearance-based classification score, and (4) aggregate the spatial layout descriptor and score it with the spatial model. We present the obtained results in Table 1.

In the first set of experiments (rows 1-3) we consider Fisher vectors without non-linear normalizations and evaluate models trained with different regularizers. The results show that the group-sparse model outperforms the ℓ_2 regularized model for 7 percentage points (pp). In comparison with the ℓ_1 regularized model [22], the group-sparse model is 17 times more sparse and achieves comparable AP. This implies substantial performance advantage in terms of execution time.

In the next set of experiments (rows 4-5), we evaluate the effect of non-linear normalizations to the performance of the group-sparse model (note that here the power normalization is always on). The ℓ_2 global and ℓ_2-intra normalizations produce comparable results and improve the performance for approximately 6 pp w.r.t model without normalizations (row 3). We further observe that intra-component normalization obtains a sparser model (7 vs. 10 components) without any performance hit. The next two experiments (rows 6-7) explore the gradient approximation presented in Section 4. Here we (i) compute the gradient of the normalized classification score w.r.t. the raw Fisher vector, and (ii) score that gradient with the raw Fisher vector of the image. We observe almost no penalty of the approximation. However, we note that the global ℓ_2 normalization (row 7) does not preserve the number of non-zero visual words in the gradient of the classification score. As a consequence, it is not suitable for constructing a spatial layout model where we require a fixed set of selected model components.

Table 1. Classification performance with different configurations (M: appearance model, G: gradient of the appearance model, SH: spatial histogram, SFV: spatial Fisher vector), FV normalizations (p: power, ℓ_2 intra: metric per component, ℓ_2 global: metric across the entire vector) and regularizations (ℓ_1, ℓ_2, group: ℓ_2 inside component, ℓ_1 between components). K_w denotes the number of non-zero components of the appearance model (out of 1024 total), where each component corresponds to a visual word.

Nr.	Configuration	FV normalization	Penalty	K_w	AP train	AP test
1	M	-	ℓ_2	1024	100	64
2	M [22]	-	ℓ_1	185	98	71.9
3	M	-	group	11	80	71.1
4	M	p, ℓ_2 global	group	10	83	76.9
5	M	p, ℓ_2 intra	group	7	81	76.8
6	G	p, ℓ_2 global	group	*	83	76.9
7	G	p, ℓ_2 intra	group	7	81	76.7
8	G + SH	p, ℓ_2 intra	group	7	92	**81.8**
9	G + SFV	p, ℓ_2 intra	group	7	94	81.2

Finally, we evaluate the spatial layout model (rows 8-9). Here we require the classification score gradient in order to be able to identify the positive patches. We observe that the combination of the group-sparsity and spatial layout model achieves the best classification AP (around 81%), which is 4 pp better than the appearance-based counterpart (row 7) and more than 9 pp better than [22] (row 2). The group-sparse model identifies only 7 visual words, so there are only 49 possible pairs to consider in the spatial model. The spatial histograms (SH) and spatial Fisher vectors (SFV) achieve comparable AP. However, the SFV descriptor is more than 3 times smaller than the SH (20 vs 64 dimensions per visual word pair), which makes it more efficient in terms of execution time.

Localization. The localization results are shown in Table 2. We first provide a strongly supervised baseline [11] which employs HOG features in the sliding window[4]. In comparison to our best weakly supervised result (row 7), the supervised HOG obtains a higher AP by 7 pp. However, the HOG implementation scans the image at 64 scales, while we only extract the SIFT descriptors at 4 scales. The second set of experiments (rows 2-5) concerns the weakly supervised appearance-only models. We identify the bounding boxes by looking at T=100 top scored patches. We construct the spatial connectivity graph according to the patch overlap and identify the connected components. The results stress out the importance of non-linear normalizations (rows 4 and 5) as they increase the AP by 5 pp w.r.t. the weakly supervised baseline [22] (row 2). Further, by using the gradient approximation (row 5) instead of the direct patch contribution $f(\mathbf{X}) - f(\mathbf{X} - \mathbf{x})$ (row 4), we obtain a comparable AP but increase the p_{miss} for 5 pp. However, we shall show that the gradient requires less execution time.

[4] These results are worse than [22] since here we do not use additional negative images for training, i.e. the training dataset is the same as in other experiments.

Table 2. Localization performance. T denotes the number of patches used to compute the object bounding box. K_w denotes the number of non-zero model components. p_{miss} denotes the miss frequency at the rightmost data point of the PR curve.

Nr.	Configuration	FV Normalization	Penalty	K_w	T	AP test	p_{miss}
1	S HOG [11]	-	l_2	-	-	**88**	**0.05**
2	M [22]	-	l_1	64	100	72	0.13
3	M	-	group	11	100	74	0.25
4	M	p, ℓ_2 intra	group	7	100	77	0.11
5	G	p, ℓ_2 intra	group	7	100	77	0.16
6	G + SH	p, ℓ_2 intra	group	7	all	75	0.14
7	G + SFV	p, ℓ_2 intra	group	7	all	**81**	**0.11**

In the third set of experiments (rows 6-7), we evaluate the localization performance of the spatial layout models. We construct the bounding boxes by taking a union of *all* patches which are positively scored by the spatial layout model. The SH achieves somewhat worse results w.r.t. appearance-only counterpart (row 5), but reduces the number of parameters (we do not have to choose T). The best performance is achieved with the SFV (row 7), where we increase the AP by 9 pp in comparison to the baseline (row 2) and by 4 pp in comparison to the appearance model (row 5). Thus the SFV model outperforms the SH model which is unable to take into account intra-bin distribution.

Fig. 3 shows some localization examples. We are quite successful in detecting very small (distant) objects. Most of our false alarms are caused by multiple detections on objects which are very close to the camera.

Execution speed. All experiments have been performed on a 3.4 GHz Intel Core i7-3770 CPU. Our Python + numpy implementation takes on average 7.4 s per image for G + SFV (7.2 s for G and 0.2 s for SFV), which is 3 s faster than the HOG baseline. We are currently unable to match the cascaded approaches [37, 13], but we think that our approach might scale better in the multi-class case due to feature sharing. Significant speed-ups could be achieved by approximate soft assign [21] or by using random decision forests as a generative model [4].

Fig. 3. Localization results: first two images depict the successful operation of our approach. The positively scored patches corresponding to different visual words are shown in different colors. The second two images show examples of false alarms.

Preliminary experiments have shown that additional speed-up could be achieved by an optimized soft-assign implementation in C. We further discuss our two main contributions in terms of execution speed.

The gradient-based evaluation of a single patch is almost twice as fast than the direct computation $f(\mathbf{X}) - f(\mathbf{X} - \mathbf{x})$ (70 μs vs. 160 μs). The effects of this speed-up are especially important when the appearance model is not so sparse and the number of patches assigned to the selected components is large. An interesting application area is the traffic sign localization in the multi-class case. To the best of our knowledge, prominent commercial real-time systems still detect only a single type of sign at a time [29] (typically the prohibition signs).

As for the choice of the spatial model, both SFV and SH take approximately 0.2 s per image. In comparison to SH, the SFV includes the computation of the gradient w.r.t. spatial GMM. However, the SFVs can be precomputed due to (i) the known size of the local neighbourhood, and (ii) fixed GMM shared across all component pairs. Thus, by using SFV we improve the classification performance while retaining the execution speed.

7 Conclusion

We have presented an approach to learn discriminative spatial relations between pairs of visual words selected by a block-sparse appearance classifier trained on the FV of entire images. The local spatial layout between the visual word pairs is represented by a suitable spatial descriptor and aggregated across the image. The recovered spatial descriptors are used to train a spatial classification model suitable for image classification and weakly supervised object localization.

Our first contribution concerns the applicability of power and metric normalizations in patch-level appearance models. Although these normalizations invalidate the additivity of Fisher vectors, we show that excellent results can be achieved by considering the gradient of the normalized Fisher vector score instead of the raw linear model. Our second contribution enriches the sparse patch-level classification model with spatial information. We show that the second-level descriptors can be formulated as spatial Fisher vectors corresponding to the pairs of selected visual words. We have evaluated the presented contributions on a public traffic sign dataset. The experimental results clearly show advantages of the normalized FV score gradient and the proposed pairwise spatial layout model in image classification and weakly supervised localization.

The obtained results suggest that sparse patch-level models may be strong enough to support weakly supervised learning of rich visual representations. Our future work shall explore further applications of the proposed spatial layout representation in the multi-class scenario.

Acknowledgement

This work has been fully supported by Croatian Science Foundation under the project I-2433-2014.

References

1. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. IEEE Trans. Pattern Anal. Mach. Intell. 34(11), 2189–2202 (2012)
2. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: NIPS (2003)
3. Arandjelović, R., Zisserman, A.: All about VLAD. In: CVPR (2013)
4. Baecchi, C., Turchini, F., Seidenari, L., Bagdanov, A.D., Bimbo, A.D.: Fisher Vectors over Random Density Forests for Object Recognition. In: ICPR (2014)
5. Brkić, K., Pinz, A., Šegvić, S., Kalafatić, Z.: Histogram-based description of local space-time appearance. In: SCIA. pp. 206–217 (2011)
6. Cinbis, R., Verbeek, J., Schmid, C.: Segmentation driven object detection with Fisher vectors. In: ICCV (2013)
7. Cinbis, R., Verbeek, J., Schmid, C.: Multi-fold MIL training for weakly supervised object localization. In: CVPR (2014)
8. Crowley, E.J., Zisserman, A.: Of gods and goats: Weakly supervised learning of figurative art. In: BMVC (2013)
9. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. Workshop on Statistical Learning in Computer Vision, ECCV (2004)
10. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
11. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
12. Deselaers, T., Alexe, B., Ferrari, V.: Weakly supervised localization and learning with generic knowledge. International Journal of Computer Vision 100(3), 275–293 (2012)
13. Dollár, P., Belongie, S., Perona, P.: The fastest pedestrian detector in the west. In: BMVC (2010)
14. Douze, M., Jégou, H.: The Yael library. In: Proceedings of the ACM International Conference on Multimedia (2014)
15. Everingham, M., Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal visual object classes (VOC) challenge. Int. J. Comput. Vision 88(2), 303–338 (Jun 2010)
16. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1627–1645 (2010)
17. Fernando, B., Fromont, E., Tuytelaars, T.: Mining mid-level features for image classification. International Journal of Computer Vision 108(3), 186–203 (2014)
18. Galleguillos, C., Babenko, B., Rabinovich, A., Belongie, S.J.: Weakly supervised object localization with stable segmentations. In: ECCV. pp. 193–207 (2008)
19. Jégou, H., Douze, M., Schmid, C.: On the burstiness of visual elements. In: CVPR (2009)
20. Jenatton, R., Mairal, J., Obozinski, G., Bach, F.R.: Proximal methods for hierarchical sparse coding. Journal of Machine Learning Research 12, 2297–2334 (2011)
21. Krapac, J., Šegvić, S.: Fast Approximate GMM Soft-Assign for Fine-Grained Image Classification with Large Fisher Vectors. In: GCPR (2015)
22. Krapac, J., Šegvić, S.: Weakly supervised object localization with large Fisher vectors. In: VISAPP (2015)
23. Krapac, J., Verbeek, J., Jurie, F.: Modeling spatial layout with Fisher vectors for image categorization. In: ICCV (2011)

24. Lampert, C.H., Blaschko, M.B., Hofmann, T.: Efficient subwindow search: A branch and bound framework for object localization. IEEE Transactions on Pattern Analysis and Machine Intelligence (2009)
25. Liu, D., Hua, G., Viola, P.A., Chen, T.: Integrated feature selection and higher-order spatial feature extraction for object categorization. In: CVPR (2008)
26. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. J. Mach. Learn. Res. 11, 19–60 (Mar 2010)
27. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. In: NIPS. pp. 570–576 (1997)
28. Mathias, M., Timofte, R., Benenson, R., Gool, L.J.V.: Traffic sign recognition - how far are we from the solution? In: IJCNN. pp. 1–8 (2013)
29. Mobileye: Traffic Sign Detection, [Online] Available: http://www.mobileye.com, Accessed: 2015-07-22
30. Murphy, K.: Machine learning a probabilistic perspective. MIT Press, Cambridge, Mass (2012)
31. Nguyen, M.H., Torresani, L., De la Torre, F., Rother, C.: Learning Discriminative Localization from Weakly Labeled Data. Pattern Recognition 47(3), 1523–1534 (2014)
32. Perronnin, F., Sánchez, J., Mensink, T.: Improving the Fisher kernel for large-scale image classification. In: ECCV. pp. 143–156 (2010)
33. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image Classification with the Fisher Vector: Theory and Practice. International Journal of Computer Vision 105(3), 222–245 (2013)
34. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep fisher networks for large-scale image classification. In: NIPS. pp. 163–171 (2013)
35. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: ECCV. pp. 73–86 (2012)
36. Siva, P., Xiang, T.: Weakly supervised object detector learning with model drift detection. In: ICCV (2011)
37. Viola, P.A., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004), http://dx.doi.org/10.1023/B:VISI.0000013087.49260.fb
38. Voravuthikunchai, W., Cremilleux, B., Jurie, F.: Histograms of pattern sets for image classification and object recognition. In: CVPR (2014)
39. Šegvić, S., Brkic, K., Kalafatic, Z., Pinz, A.: Exploiting temporal and spatial constraints in traffic sign detection from a moving vehicle. Mach. Vis. Appl. 25(3), 649–665 (2014)
40. Weng, C., Yuan, J.: Efficient mining of optimal AND/OR patterns for visual recognition. IEEE Transactions on Multimedia 17(5), 626–635 (2015)
41. Yang, Y., Newsam, S.: Spatial pyramid co-occurrence for image classification. In: ICCV (2011)
42. Yuan, J., Wu, Y., Yang, M.: Discovery of collocation patterns: from visual words to visual phrases. In: CVPR (2007)
43. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68(1), 49–67 (2006)