Don't Look Back: Post-hoc Category Detection via Sparse Reconstruction

Hyun Oh Song
Mario Fritz
Tim Althoff
Trevor Darrell

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2012-16
http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-16.html

January 24, 2012

Don't Look Back: Post-hoc Category Detection via Sparse Reconstruction

Hyun Oh Song[1] Mario Fritz[2] Tim Althoff[1] Trevor Darrell[1]
[1]UC Berkeley, [2]MPI Informatics
[1]{song, althoff, trevor}@eecs.berkeley.edu, [2]mfritz@mpi-inf.mpg.de

Abstract

We consider optimal representations for representing prototypical categories in the latent deformable part model framework, with a specific emphasis on category-level retrieval tasks defined "on the fly" for a large corpus. In this setting, it is impractical to perform an exhaustive search with a full model; we investigate methods which approximately reconstruct the score function of a novel category from a set of precomputed responses. We propose a novel sparse reconstruction method where part classifiers are decomposed via a shared dictionary of part filters; in turn, our method can efficiently reconstruct approximate part responses on large image corpora using a sparse matrix-vector product based on pre-computed filter responses instead of exhaustive convolutions. We compare our method to baseline schemes using SVD-based or nearest-category approximation and show our method is more effective at detecting novel categories. We additionally demonstrate results towards an end-to-end system for activity detection which trains a protoype category concept model from one dataset (PASCAL), learns post-hoc categories on the fly based on training data from a second dataset where labeled data are available (ImageNet), and sucessfully detects instances in test data from a third dataset (TRECVID MED) via reconstruction with the precomputed prototype models.

1. Introduction

Many perceptual tasks of interest are not known a priori, but are instead defined on the fly when a phenomena of interest is identified. For example, one may decide to search a personal media cache for cases where a certain type of car is present, or look for events of a novel type of sport or dance in online media. Contemporary object or activity category recognition methods largely consider detection of separate categories independently, applying a separate detector for each. We consider here the efficient large-scale detection of "post-hoc" categories, where the desired concept is known only after data has been collected and preprocessed, and there is not enough time to run a detector for that category exhaustively on the entire dataset. We are specifically interested in detection of small-scale object categories that occupy only a subwindow of a larger image or scene, as is customary in contemporary object category detection challenges such as PASCAL VOC.

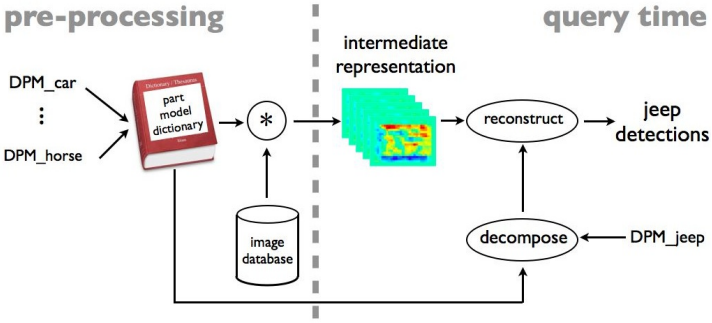
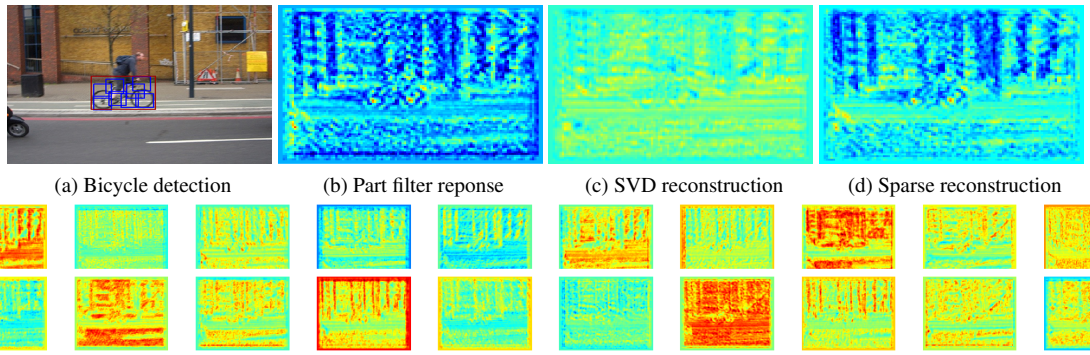


Figure 1: "Post-hoc" detection of novel categories: a part model dictionary is learned from a set of canonical categories, and used to precompute an intermediate representation for a large corpus. Later, a novel category can be efficiently approximated using sparse reconstruction.

The naive application of a windowed object detector trained post-hoc for a specific category is generally impractical on typical large-scale datasets, e.g., TRECVID MED [28]. Existing approaches either forgo any windowed representation and simply compute image-level descriptors (e.g, GIST [23], PHOG [4], SPM-BOW [18, 8]), or rely on the output of specific concept detectors precomputed on an image. The majority of these latter approaches rely also only on image-level descriptors (but see [19], which spatially pools the output of a windowed detector for a fixed number of pre-trained categories.)

While image-level image descriptors provide a successful solution to many retrieval tasks, especially those that are scene-level events (or objects that are highly correlated with scene level events) they are generally inadequate when it comes to retrieval of specific objects that may comprise only a limited region of the image and may occur in a wide range of scene contexts. In this paper, we focus on the challenge of developing an efficient representation for post-hoc windowed detection of novel categories in large scale



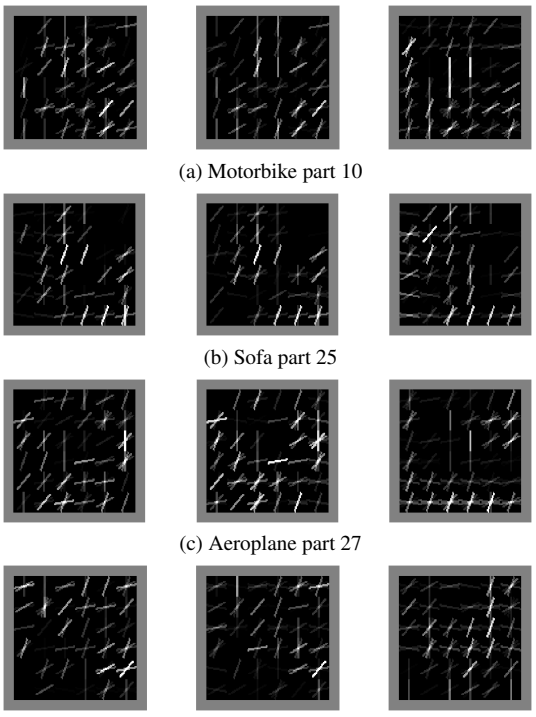
(a) Bicycle detection (b) Part filter reponse (c) SVD reconstruction (d) Sparse reconstruction

(e) Responses from activated dictionary elements

Figure 2: Top row: Bicycle detection and part filter convolution response for the wheel part. Middle row: reconstruction using only 20 bases out of 216 using SVD and sparse reconstructions respectively. Our method still maintains the specificity of the part (high response at the wheel) while SVD reconstruction fails to maintain the sharpness. Bottom row: Collage of responses from activated dictionary elements. Best viewed in color and zoomed in.

datasets. Our model computes and stores the response of a number of prototype categories over a corpus; however in contrast to conventional protoype models, our representation is specifically learned to support a sparse reconstruction of the response of a novel classifier (or category) as illustrated in figure 1. The sparse reconstruction property allows for efficient reconstruction of the response of a novel category detector across the corpus: with a typical sparsity level our method can well-approximate the response of a detector with a low-order matrix-vector product rather than an exhaustive convolution, yielding an order-of-magnitude or more fewer operations in theory, with a speedup in practice of 2X or more comparing unoptimized implementations of our method vs optimized baselines.

Our method is applicable to a wide range of category detectors, including SVM-HOG [9], Live-learning [29], BOW models[8], and Poselets [6, 5]; we have experimented primarily with the DPM model [12], as it is one of the most widely used and well performing methods. We develop sparse prototype models primarily for the part components of the DPM model, since in our experimentation these are what dominate computation time. We evaluate the performance of our method in several ways, including 1) trained and tested on PASCAL directly, 2) on concepts trained and tested on ImageNet (but using the protoype representation from PASCAL), and 3) on an end-to-end application designed to support the TRECVID MED challenge, where object category concepts relevant to post-hoc multimedia events are extracted from the query, trained on the fly from ImageNet training data, and detected in TRECVID MED videos processed using the sparse prototypes learned from the PASCAL corpus.



(a) Motorbike part 10

(b) Sofa part 25

(c) Aeroplane part 27

(d) Cat part 34

Figure 3: Randomly chosen reconstructed part filters. Each row shows original parts, sparse reconstruction with activation cap 20 and 20 SVD bases out of 216 respectively. Only positive weights are shown for clarity.

2. Related Work

Other authors have recently explored schemes for part sharing, including [24]. Our method implicitly also shares part prototypes and can provide a significant (2X) improvement in speed by compressing the parts used in typical DPM models for PASCAL, but our main focus is on the recon-

struction of novel categories. Other authors have similarly explored group regularization of a prototype representation [25], and have explored linear manifold [26] and/or topic models [14] over visual classifier parameter spaces, but similarly have also not addressed the reconstruction of novel categories, nor the advantages of sparsity for large scale detection.

Many works maintain a tree-based hierarchy of classes which can be used to speed up classification by discarding irrelevant classes while descending the tree. We note a few, focusing on recent works with vision results. Some learn a tree in a top-down fashion [16, 2] by spectral clustering on the affinity matrix. Others attempt to optimize against a discriminative objective [3, 17]. [27] builds a taxonomy of object classes based on shared features. Unlike our method, these approaches may preemptively discard a correct detection if it falls on the wrong side of a low depth decision boundary. Attempts have been made to address this using relaxed hierarchies [22, 15].

3. Method

Latent deformable part object models [12] are composed of object sized root filters and collections of smaller part filters. These filters are weight vectors over HOG [9] style features learned from minimizing classification loss functions in latent SVM framework [12]. In general, the main computational bottleneck in running these part based object detectors is in evaluating the part filters. The cascaded detector [13] alleviates this step by learning per part score thresholds to effectively prune the number of parts evaluated. However, this part-to-feature window evaluation still remains a computational bottleneck of method.

Visualizing the learned parts reveals possible directions to tackle this problem. The first column in Figure 3 shows examples of learned part filters from the DPM model; learned structures share some redundancy across parts, suggesting part models may be able to be characterized by a latent manifold-of-prototypes model. In this paper we propose and evaluate compact prototype representations for encoding DPM part filters. We consider both linear manifold models defined using SVD, and a sparse part model dictionary scheme. This latter approach has the advantage of leading to not only accurate object detection but also efficient reconstruction of novel object categories. It is important to note that our sparse manifold scheme learns a dictionary over *classifier weights* of learned part models, not over observed part features, as is the customary application of existing sparse coding methods in the recognition literature.

3.1. Deformable part model and cascaded detection

The DPM [12] defines the score of a feature window at location $\boldsymbol{x}$, $\psi(\boldsymbol{x})$ with learned part filter P_i, bias of the part filter to the root location $\boldsymbol{\Delta}_i$ and deformation cost $\boldsymbol{d}_i$ by maximizing over latent part locations $\boldsymbol{\delta}_i$ as follows,

$$score(\boldsymbol{x}) = m_0(\boldsymbol{x}) + \sum_{i=1}^{N} \max_{\boldsymbol{\delta}_i} m_i(\boldsymbol{x} + \boldsymbol{\Delta}_i + \boldsymbol{\delta}_i) - \boldsymbol{d}_i^T \boldsymbol{\delta}_i \quad (1)$$

where, $m_i(\boldsymbol{z}) = \psi(\boldsymbol{z}) \otimes P_i$

This treats locations of parts as latent variables and defines the score by optimizing over configuration of parts. In practice, evaluating the part to feature window at a location $\boldsymbol{z}$, $m_i(\boldsymbol{z})$ for all the parts are the main bottleneck of the method. The cascaded detector [13] greedily selects a part evaluation sequence and tests the output of the previous part score with learned part specific thresholds to decide whether to prune the search or further evaluate other parts in the sequence.

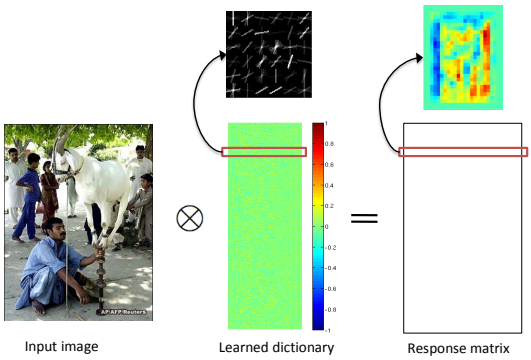


Figure 4: Overview diagram of the precomputation phase of our method: we evaluate the learned generic dictionary of compressed parts on the image.

3.2. Low dimensional representation of shared part models via SVD

A natural tool to extract low dimension representation of the data would be to compute SVD of the data matrix. Once we compute SVD on part filters from training classes and extract singular vectors, we can obtain the reconstruction weight vectors for the previously unseen query category model by taking pseudoinverse of the singular vectors. Explicitly,

$$\begin{aligned} R &= USV^T \\ \min_{\boldsymbol{w}_i} &\frac{1}{2}||\boldsymbol{p}_i - \boldsymbol{w}_i \tilde{V}^T||_2^2 \\ \boldsymbol{w}_i^* &= \boldsymbol{p}_i \tilde{V} \left(\tilde{V}^T \tilde{V}\right)^{-1} \end{aligned} \quad (2)$$

where $R \in \mathbb{R}^{n \times hp^2}$ matrix of part filters from training set stacked row-wise, $\boldsymbol{p}_i \in \mathbb{R}^{1 \times hp^2}$ is vectorized part filter from the query class object model, $V \in \mathbb{R}^{hp^2 \times hp^2}$ is

the matrix of singular vectors learned from the set of training object models by taking SVD of R and $\tilde{V}$ is the first k subset of the singular vectors from V.

3.3. Generic dictionary of compressed parts and sparse reconstruction

A sparse basis has several desirable properties, most significant among them for the purpose of our post-hoc scenario being the relative efficiency of a sparse reconstruction at query time. Our objective is to find a set of generic dictionary of compressed filters $\mathcal{Z} = \{Z_1, Z_2, \cdots, Z_J\}$ that optimally approximates the part filters from the set of training models $\mathcal{P} = \{P_1, P_2, \cdots, P_n\}$ with sparse linear combinations. Explicitly, we formulate the following optimization problem.

$$\begin{aligned} \min_{\alpha_{ij}, Z_j} & \sum_{i=1}^{N} ||P_i - \sum_{j=1}^{J} \alpha_{ij} Z_j||_2^2 \\ \text{subject to} \quad & ||\boldsymbol{\alpha_i}||_0 \leq \epsilon \quad \forall i = 1, ..., n \\ & ||Z_j||_2^2 \leq 1 \quad \forall j = 1, ..., J \end{aligned} \tag{3}$$

where $P_i \in \mathbb{R}^{\,p \times p \times h}$ is a part filter or convolution tensor, Z_j is a dictionary element of the same size. $\boldsymbol{\alpha}_i \in \mathbb{R}^J$ is the activation vector, ϵ imposes a cap on the number of activations, p is filter size, h is feature dimension. Or equivalently, by stacking the columns on top of each other,

$$\begin{aligned} \min_{D, \boldsymbol{\alpha}_i} & \sum_{i=1}^{N} ||\mathbf{y}_i - D\boldsymbol{\alpha_i}||_2^2 \\ \text{subject to} \quad & ||\boldsymbol{\alpha_i}||_0 \leq \epsilon \quad \forall i = 1, ..., n \\ & ||D_j||_2^2 \leq 1 \quad \forall j = 1, ..., J \end{aligned} \tag{4}$$

where $\mathbf{y}_i = \text{vec}\,(P_i)$ and $D = [\text{vec}\,(Z_1), \cdots, \text{vec}\,(Z_J)]$

Although the above optimization is NP-hard in estimating $\boldsymbol{\alpha_i}$, greedy algorithms such as orthogonal matching pursuit algorithm (OMP) [7, 21] can efficiently compute an approximate solution to the problem. OMP iteratively estimates the optimal matching projections of the input signal onto the over complete dictionary D. Interested readers on OMP are referred to [7, 21]. However, the above optimization problem is convex with respect to D if $\boldsymbol{\alpha_i}$ is fixed so we can optimize the objective in coordinate descent fashion iterating between updating $\boldsymbol{\alpha_i}$ while fixing D and vice versa. We used an online dictionary learning algorithm to solve this optimization problem [20]. Figure 3 shows randomly chosen part filters from PASCAL VOC 2007 [11] dataset compared to our sparse reconstruction with $\epsilon = 20$ and reconstruction from 20 singular bases out of feauture size 216.

3.4. Precomputation with learned bases and efficient reconstruction

In a post-hoc setting, we amortize the time required to compute an intermediate representation. By linearity of convolution, we can precompute the convolution response with the dictionary bases. Then, later at the inference stage, we can use sparse reconstruction with the activation vector estimated from the query object model to approximate the convolution response we would have gotten from convolving with the original filter sets.

$$\begin{aligned} \Psi \otimes P_i &= \Psi \otimes \left(\sum_j \alpha_{ij} D_j \right) \\ &= \sum_j \alpha_{ij} \left(\Psi \otimes D_j \right) \end{aligned} \tag{5}$$

This preprocessing of image corpora allows us to not look back at the corpora but rather work with intermediate representation for efficient search. Concretely, we can recover individual part filter responses via sparse matrix multiplication (or lookups) with the activation vector replacing the heavy convolution operation as shown in Eqn (6).

$$\begin{bmatrix} -\Psi \otimes P_1 - \\ -\Psi \otimes P_2 - \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ -\Psi \otimes P_n - \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \boldsymbol{\alpha}_n \end{bmatrix} \begin{bmatrix} -\Psi \otimes D_1 - \\ -\Psi \otimes D_2 - \\ \vdots \\ \vdots \\ -\Psi \otimes D_J - \end{bmatrix} = A\ M \tag{6}$$

Finally, reconstructed approximate score is evaluated as following,

$$\begin{aligned} & recon\ score\,(\boldsymbol{x}) = m_0\,(\boldsymbol{x}) + \\ & \sum_{i=1}^{N} \max_{\boldsymbol{\delta}_i}\ s_i\,(\boldsymbol{x} + \boldsymbol{\Delta}_i + \boldsymbol{\delta}_i) - \boldsymbol{d}_i^T \boldsymbol{\delta}_i \\ & \text{where, } s_i\,(\boldsymbol{z}) = \sum_{\substack{j=1 \\ \forall \alpha_{ij} \neq 0}}^{|D|} \alpha_{ij}\,(\psi\,(\boldsymbol{z}) \otimes D_j) \end{aligned} \tag{7}$$

After the precomputation, the part filter to feature window evaluation term, $s_i\,(\boldsymbol{z})$ simplifies to,

$$s_i\,(\boldsymbol{z}) = \sum_{\substack{j=1 \\ \forall \alpha_{ij} \neq 0}}^{|D|} \alpha_{ij} M\,(\boldsymbol{z}) \tag{8}$$

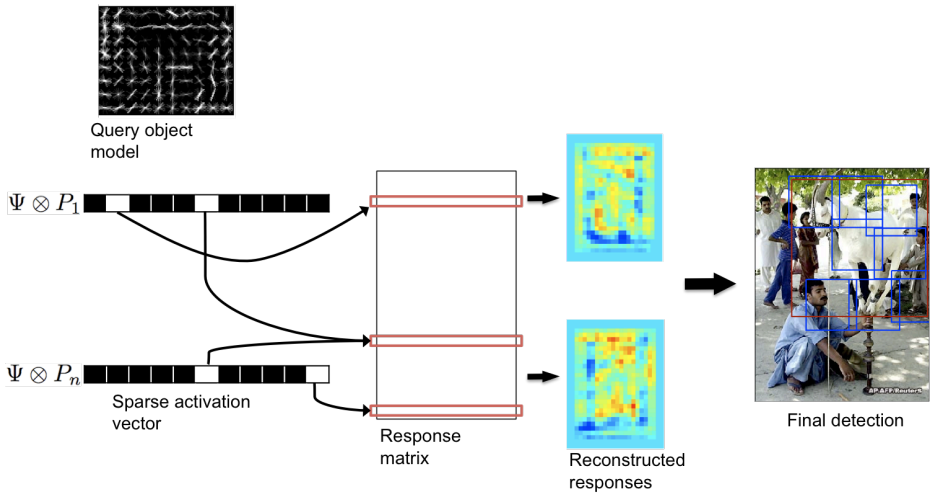


Figure 5: Overview diagram of the inference phase of our method: we estimate the sparse reconstruction vectors from the query object model and efficiently reconstruct the part filter response with the given model.

Note that the summation is only over non-zero elements of the sparse vector $\boldsymbol{\alpha_i}$. Additionally, this could be efficiently implemented as sparse matrix multiplications or lookups. Figure 2 shows a sample reconstruction and figures 4 and 5 summarize our framework.

In the online detection scheme, we convolve the query image with the dictionary elements and do sparse reconstruction on the fly. For dictionary size $|D|$, total number of filters N, filter size p and feature dimension h, an exhaustive convolution based detection scheme requires Nhp^2 operations per pixel in a score pyramid when our scheme requires $|D|hp^2 + N\mathbb{E}[||\boldsymbol{\alpha_i}||_0]$ algebraic operations. The first term is from convolution with dictionary elements and the second term is the average activation level from the sparse reconstruction. As the number of classes to search is much larger than the dictionary size, this precomputation time $|D|hp^2$ is amortized. On the other hand, in a multi-media retrieval scenario where one-time preprocessing of the corpus is allowed, for every pixel in a score pyramid, the factor of speed up in reconstruction is the ratio between the complexity of the convolution kernel and the average activation. Explicity,

$$\frac{hp^2}{\mathbb{E}[\,||\boldsymbol{\alpha_i}||_0\,]}$$

For example, reconstructing response from a 6 by 6 kernel with feature dimension 6 and average activation level of 20, we would get more than an order of magnitude speedup in terms of number of arithmetic operations.

4. Experiments

We performed experiments to analyze how much average precision (AP) we can reconstruct on previously unseen novel categories at given time budget on three datasets: PASCAL VOC2007 dataset [11] dataset, subset of Imagenet [10] data and sample keyframes from TRECVID dataset [28]. To provide a ground truth for AP comparison, we ran cascaded deformable part models[13] trained on the test heldout category model. For comparison, we extracted singular vectors learned from the training models and estimated the reconstruction weight vectors from the previously unseen query category models as discussed in the previous section. We also explored a nearest-neighbor-of-parts baseline where the query object model retreives closest matching part filters (in L2 distance) from the pool of training object models. Global threshold was fixed to -1.1 for all object models throughout the experiment for consistency. This number was roughly the saturation threshold for AP evalutation.

4.1. Reconstruction on heldout categories from PASCAL VOC2007 dataset

Using the original query category object detector as ground truth, in order to test the reconstruction generalization performance against previously unseen category model, we performed leave one class out evaluation where we used dictionaries and the set of singular vectors that are trained on all other classes except the heldout evaluation class.

Figure 6 shows the experimental results in AP and time for all 20 classes from PASCAL VOC 2007 test dataset [11]. Figure 7 shows the plot of class averaged time versus AP for different parameters used in the experiment. We can see from figure 6 that our sparse reconstruction method preserves AP on most categories which DPM model performs well (e.g. bicycle, bus, car, horses, motorbike, person and train) even with only 20 bases while svd reconstruction doesn't quite preserve the AP at the same level of time

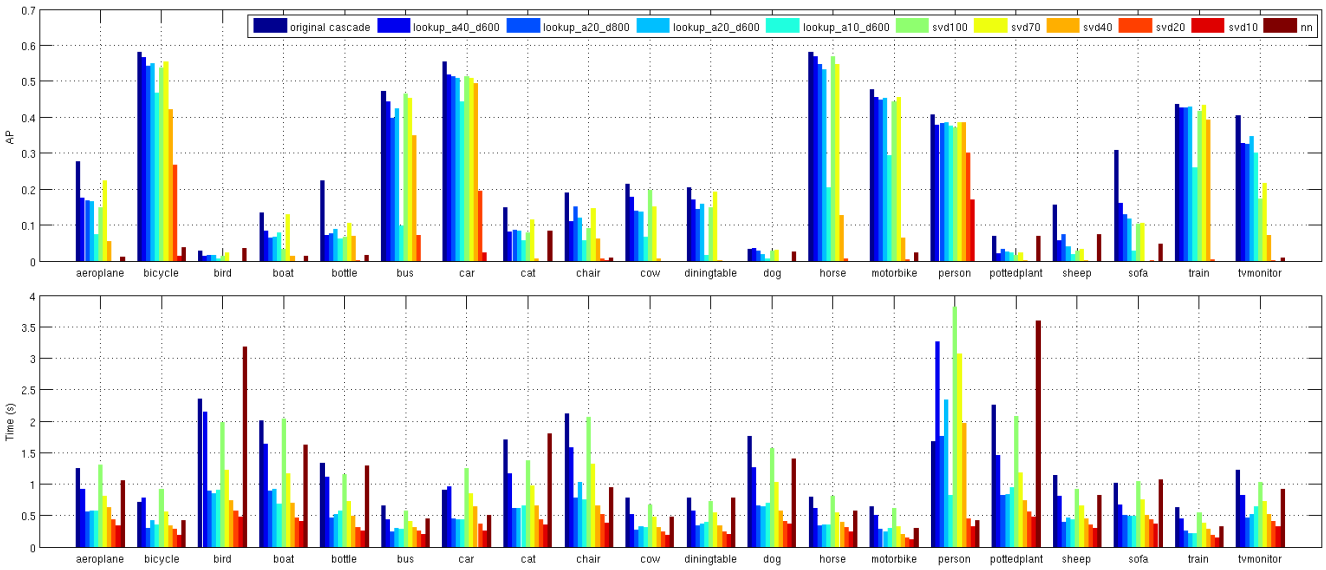


Figure 6: Results on PASCAL VOC2007 dataset [11]. The top row shows AP and the bottom row shows time per class

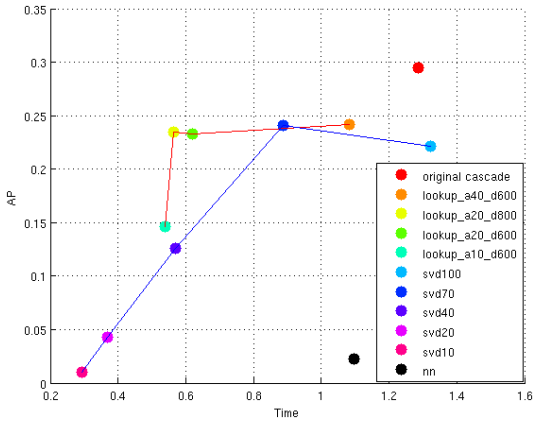


Figure 7: Class averaged time vs AP on PASCAL VOC2007 dataset [11] dataset. lookup_a20_d800 indicates 20 bases with dictionary size 800. svd20 indicates 20 number of bases used for reconstruction. Red line connects results from our method and the blue line connects results from SVD reconstruction

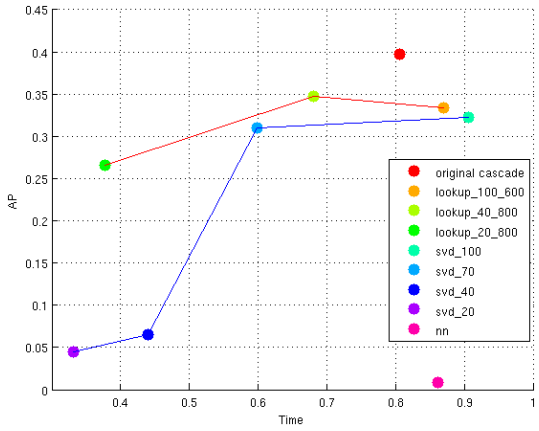


Figure 9: Class averaged time vs AP on ImageNet [10] dataset. lookup_a20_d800 indicates 20 bases with dictionary size 800. svd20 indicates 20 number of bases used for reconstruction. Red line connects results from our method and the blue line connects results from SVD reconstruction

budget (> 2.3X speedup regime) but recovers the AP when more than $\frac{1}{3}$ of total bases are allowed. Although nearest-neighbor-of-parts baseline has poor performance overall, on subset of categories that DPM detector has poor detection performance (e.g. bird, dog and pottedplant), it works as well as or a little bit better than the original query model.

4.2. Transfer from PASCAL to ImageNet models

We selected 9 categories (sailboat, bread, cake, candle, fish, goat, jeep, scissors and tire) from ImageNet [10] that had the potential to be relevant to TRECVID MED event recognition categories (see following section) and which had substantial appearance changes from the 20 existing PASCAL categories; we tested how the dictionary learned from a set of categories learned in the previous experiment performed when approximating previously unseen novel categories trained and tested on ImageNet imagery. Figure 8 shows the experimental results in AP and time for the 9 categories. Figure 9 shows the class averaged time versus AP for different settings and visualizes the reconstruction performance margin between SVD reconstruction and our sparse reconstruction method at limited time budget of 2.5X speedup. This shows the dictionary of parts learned from PASCAL well transfers to novel categories from ImageNet domain.

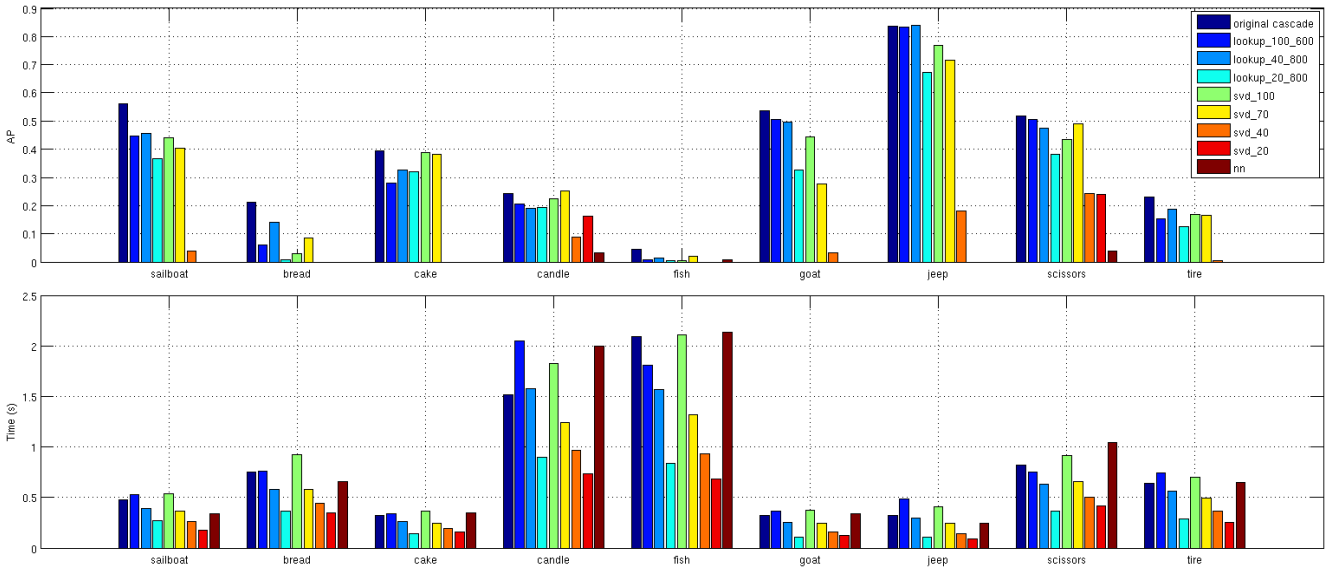


Figure 8: Results on ImageNet [10] test data from dictionaries trained on PASCAL data. The top row shows AP and the bottom row shows time per class

4.3. PASCAL to ImageNet to TREVID dataset

The categories in the previous experiment were selected for possible relevance to categories in the 2011 TRECVID Multimedia Event Detection (MED) Event Kits [28]. The main idea is to propose methods for efficient post-hoc indexing of objects that may co-occur with events of interest, but are not present in a pre-defined concept model. We used the computed object models from the previous experiment (trained on ImageNet examples, approximated using the sparse bases computed on PASCAL), and tested how well the approximate classifiers performed vs. baseline on TRECVID imagery. For this experiment we do not expect all tested words to necessarilly respond to the event imagery, nor do we expect overall high-AP. We are interested in demonstrating whether at least some subset of objects are effective in a high-precision low-recall setting, such that cues based on these detections can be reliably integrated into a larger multi-modal TRECVID MED retrieval system, and most importantly for this paper whether the approximations we need for tractible application of these models preserve accuracy in this high-precision regime.

The TRECVID MED event kit consists of publicly available, user-generated videos. The contents of these videos are highly variant, for example, "wedding ceremony" varies from a traditional catholic mass, to a Hindi ceremony, to home-made music videos. Figure 10 shows some example keyframes from the event kit. From these highly unconstrained videos, we sampled 500 keyframes and had them annotated using Amazon Mechanical Turk [1]. Figure 11 shows category averaged precision for top 50 retrieved examples. We can see that at the same number of bases, sparse reconstruction well approximates the precision as compared

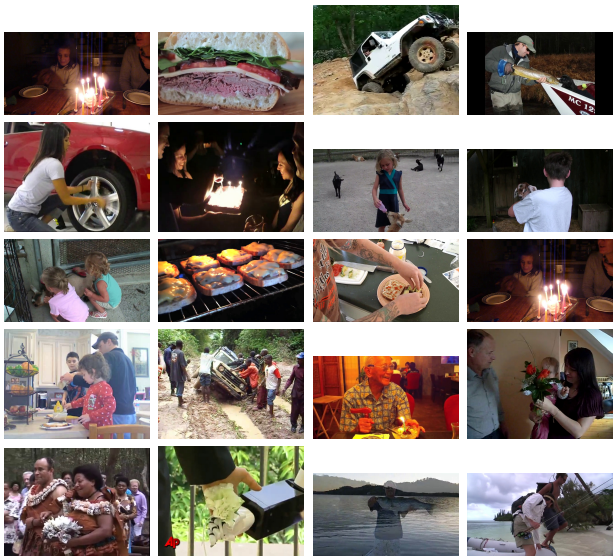


Figure 10: Examples of extracted keyframes from TRECVID [28] dataset

to SVD reconstruction method by a large margin.

5. Conclusion

This paper presents a novel sparse reconstruction method based on precomputing the filter responses with learned dictionary of surrogate part models for efficient reconstruction of post-hoc object categories on three datasets. Our experiments show that the proposed method can well generalize to most of the previously unseen categories not only in the same domain (PASCAL) but also outside the domain (ImageNet, TRECVID) with the benefit of detection

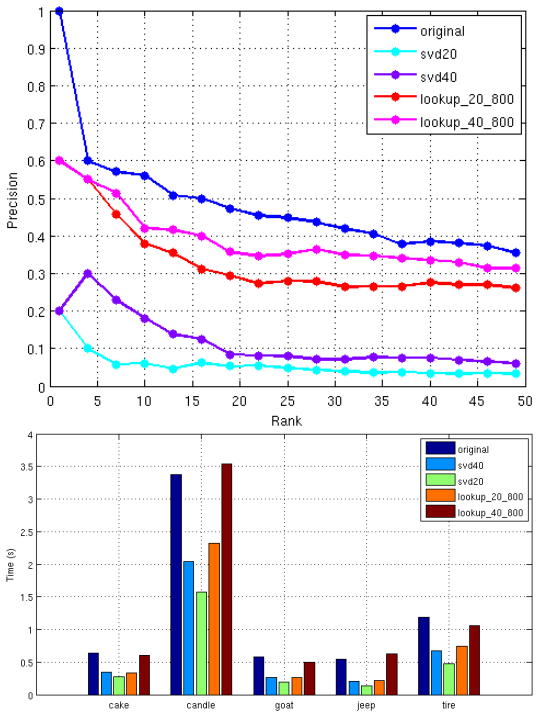


Figure 11: Top row: Ranked retrieved examples vs Precision on TRECVID [28] dataset. Bottom row: Category vs Time

efficiency on top of the speedup from the cascade object detector [13]. We plan to explore optimized implementation of our method to obtain more aggressive efficiency that matches the theoretical speedup of our method.

Acknowledgements

This work has been partially supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11-PC-20066. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government. H.S. is partially supported by Samsung Scholarship Foundation. T.D. is partially supported by DARPA, NSF and Toyota grants.

References

[1] Amazon Mechanical Turk. http://www.mturk.com. 7

[2] S. Bengio, J. Weston, and D. Grangier. Label embedding trees for large multi-class tasks. *Advances in Neural Information Processing Systems*, 23(1):163–171, 2010. 3

[3] A. Binder, K.-R. Müller, and M. Kawanabe. On Taxonomies for Multi-class Image Categorization. *International Journal of Computer Vision*, Jan. 2011. 3

[4] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 401–408. ACM, 2007. 1

[5] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *European Conference on Computer Vision (ECCV)*, 2010. 2

[6] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *International Conference on Computer Vision (ICCV)*, 2009. 2

[7] S. F. Cotter, B. D. Rao, K. Kreutz-Delgado, and J. Adler. Forward sequential algorithms for best basis selection. *IEE Proceedings Vision Image and Signal Processing*, 146(5):235, 1999. 4

[8] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, page 22, 2004. 1, 2

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2, 3

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 5, 6, 7

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 4, 5, 6

[12] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010. 2, 3

[13] P. F. Felzenszwalb, R. B. Girshick, and D. A. McAllester. Cascade object detection with deformable part models. In *CVPR*, 2010. 3, 5, 8

[14] M. Fritz and B. Schiele. Decomposition, discovery and detection of visual categories using topic models. In *CVPR*, 2008. 3

[15] T. Gao and D. Koller. Discriminative Learning of Relaxed Hierarchy for Large-scale Visual Recognition. *ICCV*, 2011. 3

[16] G. Griffin and P. Perona. Learning and using taxonomies for fast visual categorization. *CVPR*, pages 1–8, June 2008. 3

[17] K. Lai, L. Bo, X. Ren, and D. Fox. A Scalable Tree-based Approach for Joint Object and Pose Recognition. *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011. 3

[18] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 1

[19] L. Li, H. Su, E. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. *NIPS*, 2010. 1

[20] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010. 4

[21] S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993. 4

[22] M. Marszalek and C. Schmid. Constructing category hierarchies for visual recognition. *ECCV*, 2008. 3

[23] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. 1

[24] P. Ott and M. Everingham. Shared parts for deformable part-based models. In *CVPR*, pages 1513–1520, 2011. 2

[25] H. Pirsiavash, D. Ramanan, and C. Fowlkes. Bilinear classifiers for visual recognition. In *NIPS*, 2009. 3

[26] A. Quattoni, M. Collins, and T. Darrell. Transfer learning for image classification with sparse prototype representations. In *CVPR*, 2008. 3

[27] N. Razavi, J. Gall, and L. J. V. Gool. Scalable multi-class object detection. In *CVPR*, pages 1505–1512, 2011. 3

[28] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press. 1, 5, 7, 8

[29] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. In *CVPR*, 2011. 2