Visualizing Object Detection Features

Xinyi Gu, Zhuoyuan Xu, Yupei Zhou
New York University, New York, NY, 10012
{xg2085, zx1137, yz7005}@nyu.edu

Abstract

The paper "Visualizing Object Detection Features" (Vondrick et al, 2013) introduces the paired dictionary algorithm to visualize feature spaces, especially the histogram of oriented gradients (HOGs), in object detection problems. The algorithm inverts high-dimensional features to natural images for researchers to directly inspect the potential reasons in detection failures. In this project, our team first aims to re-implement the algorithm in the paper. We explore the effects of two major parameters that affect model performance and speed. Then, we compare the algorithm with a baseline and a more current neural-network-based method to gain a thorough understanding of its pros and cons.

1. Introduction

Computer vision researchers often find it challenging to explain the failures in their object detectors, as they may result from the training sets, the learning algorithms, or the features. Yet, it is crucial to understand the failures to build the next generation detectors. The visualization of the high-dimensional feature spaces with natural images provides a straightforward method for researchers to check the behaviors of their object detectors' features. After visualization, some falsely identified objects have deceptive looks to true positives in the chosen feature spaces, introducing confusion to the detection models.

The major goal of this project is to implement the feature visualization tool, the paired dictionary method, described in "Visualizing Object Detection Features" (Vondrick et al, 2013). This method focuses on inverting the HOGs back to natural images using dictionaries trained on a massive database. It works on convolutional neural network (CNN) features as well, demonstrating its feature independence capability. Meanwhile, Vondrick states it is fast to train on a single laptop and complete a single test in seconds. Our team replicates the method with selected mislabeled images and explores the effects of the regularization parameter on the results. Furthermore, we compare the method with the exemplar linear discriminant analysis (ELDA) baseline and a more recent neural-network based algorithm for a thorough understanding of the proposed method and its contribution to the progress of object detection algorithms.

2. Related Works

The paired dictionary algorithm is part of the growing amount of studies in feature inversion. In earlier works, Oliva and Torralba (2001) iteratively recovered an image given GIST descriptors. Weinzaepfel et al (2011) designed a nearest-neighbor-based approach given DSIFT descriptors (Lowe, 1999), and Hariharan et al (2012) took an exemplar-LDA-based method, both requiring a massive dataset. d'Angelo et al (2012) analytically solved the problem given only LBP features (Calonder et al, 2010; Alahi et al, 2012). We also note the image reconstruction with Bag-of-Visual-Words model by Kato et al (2014). Recently, works have been done to understand the learned representations in and with deep neural networks (NNs). Zeiler and Fergus (2013) visualized activations from a CNN (DeCovNet), and Mahendran and Vedaldi (2014) proposed a general visual feature inversion method from CNNs using natural image priors. Dosovitskiy and Brox (2016) inverted visual representations of various kinds with CNNs.

The Vondrick paper claimed that its algorithm is able to generalize to different features fast and optimizes for multiple inversions. Still, it is worth noting that the algorithm received critiques in some later studies for its instablities to adversarial noises and inconsistencies on non-shallow representations (Mahendran and Vedaldi, 2014; Dosovitskiy and Brox, 2016). While many more feature inversion tools with superior results are available nowadays, the thoughts and creativity behind this milestone method are still worth analyzing as the ground of shallow representations in object detection problems and contemporary feature inversion and visualization methods.

3. Paired Dictionary Algorithm

Vondrick formulates feature inversion as a process reconstructing the image that generates a feature vector. Let $x \in R^D$ be an image and $y = \phi(x) \in R^d$ be its HOG

descriptor. Since the descriptor is a many-to-one function, it has no analytic inverse. Thus, the problem turns to find an image x which minimizes the distance between its actual HOG descriptor ϕ and the original descriptor y.

$$\phi^{-1}(y) = arg \min_{x \in R^D} \|\phi(x) - y\|_2^2 \tag{1}$$

Since Eqn.1 is not convex and has frequent local minima, it can be challenging to optimize. The paper adds two modifications to overcome the difficulty. First, it rewrites x and y respectively in terms of basis $U \in R^{D \times K}$ for the image and basis $V \in R^{d \times K}$ for the HOG features, with shared coefficient $\alpha \in R^K$. The U and V are the "paired dictionary" for the algorithm.

$$x = U\alpha \text{ and } y = V\alpha \tag{2}$$

The core observation from Eqn.2 is that inversion can be obtained by first projecting the HOG features y onto the HOG basis V and then projecting α onto the image basis U to get the image recovered from the feature.

$$\begin{aligned} &\phi^{-1}(y) = U\alpha^* \\ &\text{where } \alpha^* = arg \min_{\alpha \in R^K} \|V\alpha - y\|_2^2 \text{ s.t. } \|\alpha\|_1 \leq \lambda \end{aligned} \tag{3}$$

There is a sparsity prior on α parameterized by $\lambda \in R$ that penalizes pairs of images too similar in the image space.

To hold Eqn.2 and Eqn.3, the paired dictionary algorithm requires to learn the optimal dictionaries U and V. The paper solves this learning process by a sparse-coding inspired algorithm (Yang et al, 2010; Wang et al, 2012):

$$\begin{aligned} &arg \min_{U,V,\alpha} \sum_{i=1}^{N} (\|x_i - U\alpha_i\|_2^2 + \|\phi(x_i) - V\alpha_i\|_2^2) \\ &\text{s.t. } \|\alpha_i\|_1 \leq \lambda, \|U\|_2^2 \leq \gamma_1, \|V\|_2^2 \leq \gamma_2 \end{aligned} \tag{4}$$

γ_1 and γ_2 are hyperparameters. The second modification on Eqn.1 revealed in Eqn.4 is that it takes a sub-optimal greedy approach by supposing the first $i-1$ inversions have already been computed. The overall algorithm is then optimized by SPAM (Mairal et al, 2009). The exact steps taken during the implementation is illurated in Fig.1

4. Experiment Results

4.1. Data

We mainly use the PASCAL VOC dataset (Everingham et al, 2010) to conduct the following experiments, similar to the Vondrick paper due to its representativeness and rigorous procedure in sample collection. We also introduced a new set of mislabeled images from the study done by Hoiem et al (2012) to test the efficiency of the visualization method. These images from human judgement can be deceptive to machine after HOG transformation.

4.2. Evaluation of Single Inversion

The purpose of the inversion is to convert HOG features to a format which introduces human the perception of machines to analyze errors. We present some demonstrations in gray scale, while aware that paired dictionary can recover color information. Since feature mapping using HOGs will include background noises, object detection for certain images can be distracted. The car image in Fig.2 is a typical case: the cloud in the background interrupts the classification from our human perspective. With a single inversion, the cloud shows a aeroplane shape which lead to a false machine labeling. Other presented cases in Fig.2 and Fig.3 also reveal deceptive looks to false labels after the HOG transformation. In these examples, feature-related elements would probably be root of object detection failures.

4.3. Parameter Exploration

This section focuses on the sparsity regularization parameter λ in Eqn.3 and 4 and the size of the paired dictionary which is indicated to affect the performance and speed of sparse coding (Yang et al, 2010). The training process uses a subset of ∼15K random images from the PASCAL VOC dataset. We only change one of the two variables each time and keep others under the default of the paper on a MacBook Pro with 16GB memory and M1 chip.

As shown in Fig.4, the details and noises in the inversion results increase as the magnitude of λ decrease from 1 to 0.005. The paper determines 0.02 to be the best value considering the clarity and time based on human judgement. The total training time also increases as λ values decrease in Fig.5. As the dictionary size decreases from 2000 to 50, both the image quality and learning time decrease in Fig.6 and Fig.7. The image gradually becomes more granular and blurry under smaller dictionary sizes. In practice, we should choose appropriate values that balances between the reconstruction quality and computation resources.

4.4. Comparison with Other Inversion Methods

We compare the paired dictionary algorithm in Sec.3 with the ELDA algorithm by Hariharan et al (2012) and the CNN based algorithm by Mahendran and Vedaldi (2014). The ELDA model is the average of M images in a large image database whose HOG features have the highest scores determined by a LDA classifier trained with this given HOG feature ϕ on the same image database. Mahendran and Vedaldi (2014) implements HOG feature extraction as CNNs and minimizes a loss function similar to Eqn.2 using gradient descent with momentum.

Fig.8 demonstrates the results of the 3 inversion methods. ELDA gives a blurry and non-informative image with few details resembling the original image. ELDA is also computationally expensive as it requires comparison with all images in a large database. Since we trained it with

only a fraction of the PASCAL VOC dataset due to computational constraint, the result is expectedly poor. With the whole dataset, the Vondrick paper is able to reach better results in ELDA. Both paired dictionary and CNN based inversion give detailed results in much shorter time, with CNN providing more granular details. We compute the normalized sum square difference S_{sq} and normalized cross correlation C_{cr} between the original image I and inverted image J as follows:

$$S_{sq} = \frac{\sum(I[m,n] - J[m,n])^2}{\sqrt{\sum I[m,n]^2 \times \sum J[m,n]^2}}$$
$$C_{cr} = \frac{\sum(I[m,n] \times J[m,n])^2}{\sqrt{\sum I[m,n]^2 \times \sum J[m,n]^2}} \quad (5)$$

The average values of S_{sq} and C_{crr} across test images are in Table 1. ELDA achieves the worst performance on both metrics. CNN and the paired dictionary have performance much better than ELDA with CNN being the best among all. They are able to retain most information in the original image. In terms of the test time, CNN takes about 20 seconds

	S_{sq}	C_{cr}
ELDA	0.407	0.259
Paired Dictionary	0.342	0.316
CNN Based	0.339	0.320

Table 1. Evaluation metrics for different inversion methods. Lower is better for S_{sq}; a score of 0 is perfect. Lower is better for C_{cr}; a score of 1 is perfect.

to produce a inversion and the paired dictionary mostly less than 10 seconds.

5. Conclusion

Visualizations are powerful tools that help researchers understand feature extraction systems. Due to the effectiveness and popularity of the HOG descriptor, its visualizations is crucial in creating better object detection systems. In this paper, we explore the visualization algorithm proposed by Vondrick et al (2013) that attempts to find the inversion of HOG features, and we experiment with its key parameters. The comparison with other inversion algorithms shows that Vondrick's algorithm, though not the most detailed inversion, is fast and recovers most information. In the future, more visualization of the deep representation from NNs may be of interests due to their prevalence and growing efficiencies, and the CNN method tested in the project has already shown its potential in feature inversion. Another popular direction is the generalized visual explanation tools, such as GRAD-CAM (Selvaraju et al, 2016) that also allow researchers to inspect their model failures, resist models from adversarial images, and establish trust in predictions.

References

[1] A. Alahi, R. Ortiz, and P. Vandergheynst. Freak: Fast retina keypoint. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 510–517, 2012.

[2] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: Binary robust independent elementary features. In *Eur. Conf. Comput. Vis.*, volume 6314, pages 778–792, 09 2010.

[3] E. d'Angelo, A. Alahi, and P. Vandergheynst. Beyond bits: Reconstructing images from local binary descriptors. In *Proceedings - International Conference on Pattern Recognition*, pages 935–938, 01 2012.

[4] A. Dosovitskiy and T. Brox. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[5] M. Everingham, S.M. Ali Eslami, L. Van Gool, C.K.I Williams, J.M. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111:98–136, 2014.

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

[7] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *ECCV*, 2012.

[8] H. Kato and T. Harada. Image reconstruction from bag-of-visual-words. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[9] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2, 1999.

[10] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5188–5196, 2015.

[11] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th International Conference On Machine Learning, ICML 2009*, volume 382, page 87, 01 2009.

[12] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2004.

[13] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[14] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. HOGgles: Visualizing Object Detection Features. *ICCV*, 2013.

[15] S. Wang, L. Zhang, Y. Liang, and Q. Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2216–2223, 2012.

[16] P. Weinzaepfel, H. Jégou, and P. Pérez. Reconstructing an image from its local descriptors. *Computer Vision and Pattern Recognition*, 06 2011.

[17] J. Yang, J. Wright, T.S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873, 2010.

[18] M.D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

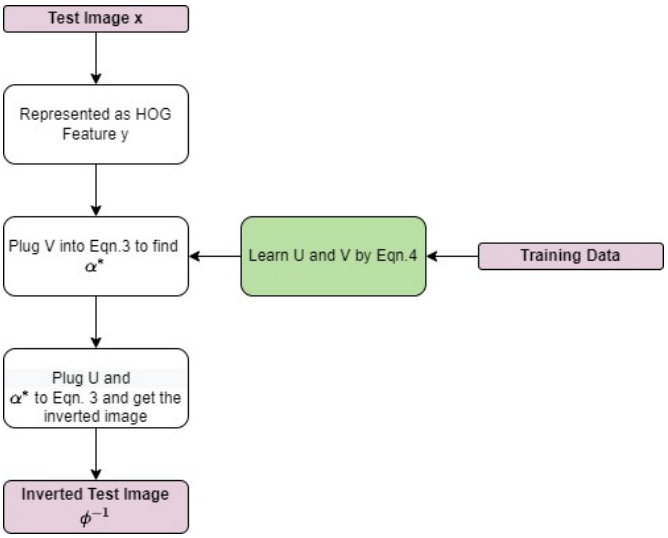
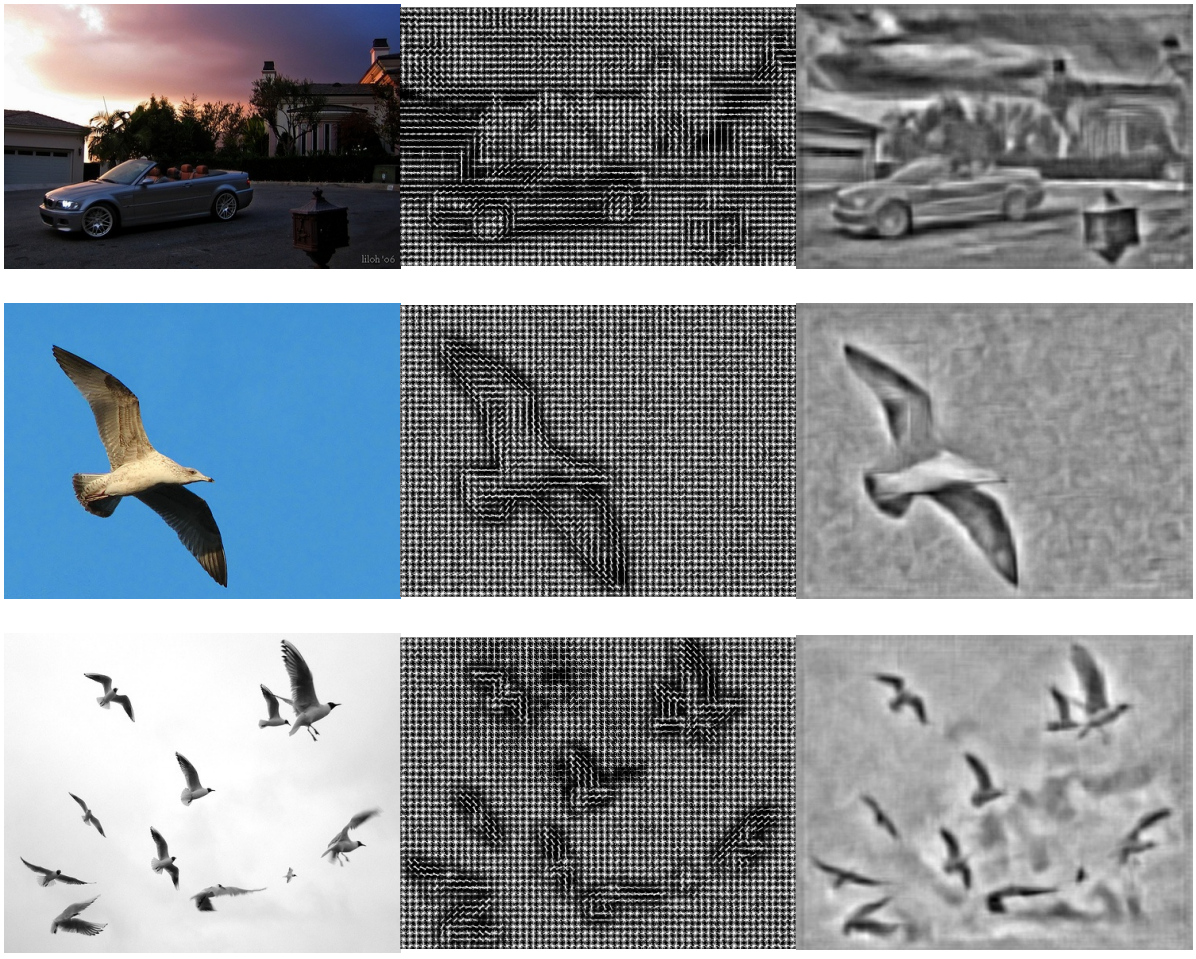


Figure 1. Inversion process of the paired dictionary algorithm.



(a) Original Image (b) HOG Descriptor (c) Single Inversion

Figure 2. Feature inversion results of images mislabeled as aeroplane by the FGMR and VGVZ object detectors (Hoiem et al, 2012)

(a) Original Image (b) HOG Descriptor (c) Single Inversion

Figure 3. Feature inversion results for images mislabeled as cow by the FGMR and VGVZ object detectors (Hoiem et al, 2012)

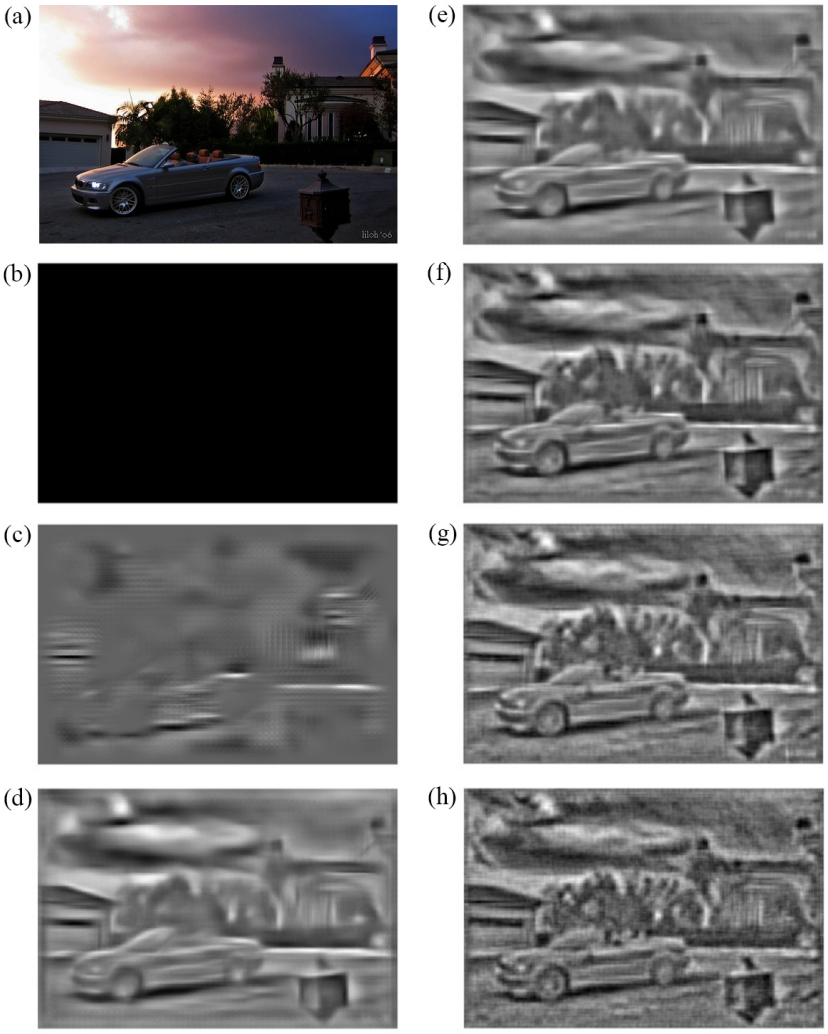


Figure 4. Feature inversion visualization with different regularization λ values. (a) original image; (b) $\lambda = 1$; (c) $\lambda = 0.5$; (d) $\lambda = 0.1$; (e) $\lambda = 0.05$; (f) $\lambda = 0.02$; (g) $\lambda = 0.01$; (h) $\lambda = 0.005$.

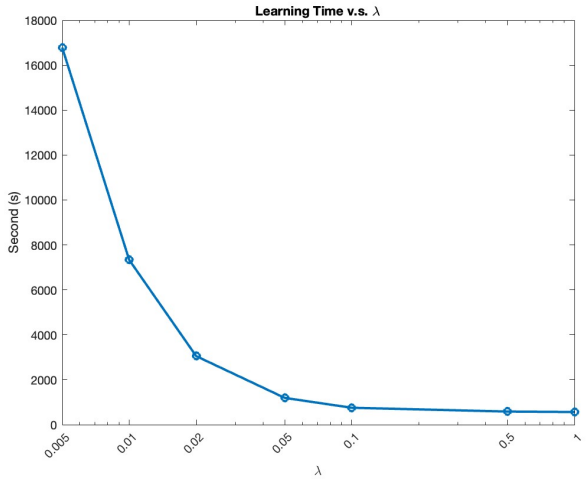


Figure 5. Total learning elapsing time versus different regularization λ values.

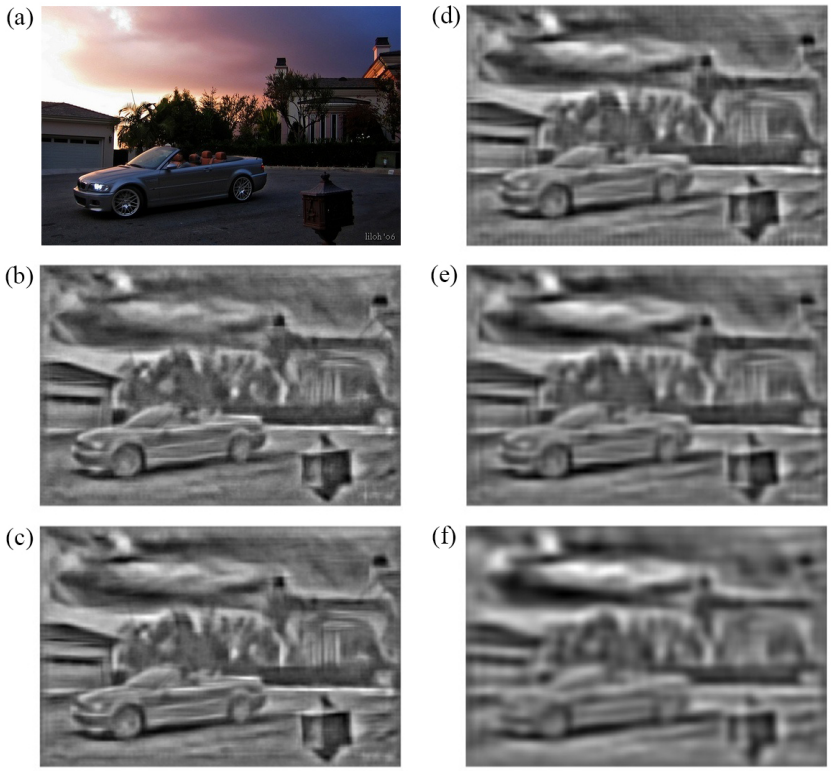
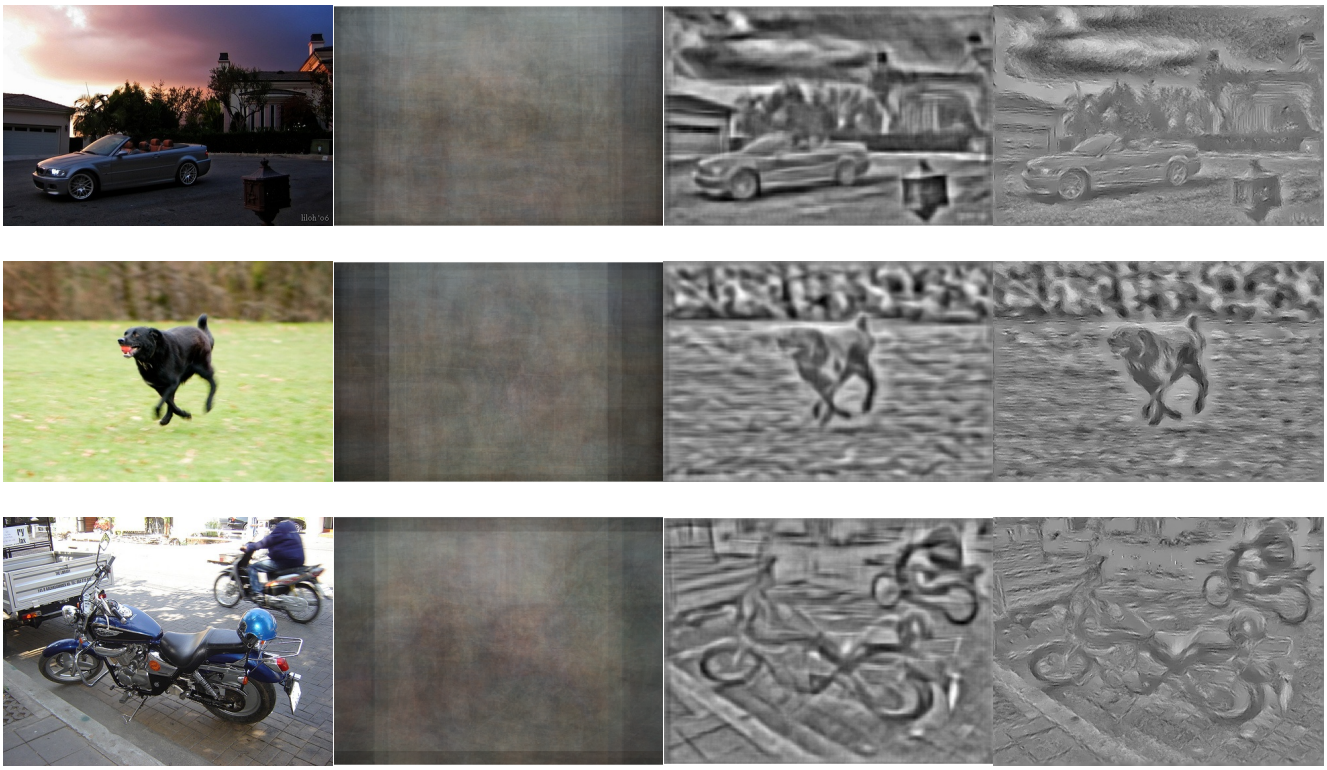


Figure 6. Feature inversion visualization with different dictionary sizes. (a) original image; (b) $size = 2000$; (c) $size = 1000$; (d) $size = 500$; (e) $size = 200$; (f) $size = 50$.

Figure 7. Total learning elapsing time versus different dictionary sizes.



(a) Original image (b) ELDA (c) Paired Dictionary (d) CNN based inversion

Figure 8. Comparison of results from different inversion algorithms.