DETECTOR WITH FOCUS:
NORMALIZING GRADIENT IN IMAGE PYRAMID

Yonghyun Kim   *Bong-Nam Kang*   *Daijin Kim*

Department of Computer Science and Engineering, POSTECH, Korea

ABSTRACT

An image pyramid can extend many object detection algorithms to solve detection on multiple scales. However, interpolation during the resampling process of an image pyramid causes gradient variation, which is the difference of the gradients between the original image and the scaled images. Our key insight is that the increased variance of gradients makes the classifiers have difficulty in correctly assigning categories. We prove the existence of the gradient variation by formulating the ratio of gradient expectations between an original image and scaled images, then propose a simple and novel gradient normalization method to eliminate the effect of this variation. The proposed normalization method reduce the variance in an image pyramid and allow the classifier to focus on a smaller coverage. We show the improvement in three different visual recognition problems: pedestrian detection, pose estimation, and object detection. The method is generally applicable to many vision algorithms based on an image pyramid with gradients.

Index Terms— normalization, detection, gradient

1. INTRODUCTION

Gradient and image pyramid are one of the essential parts for computer vision. Well-known methods based on magnitudes and orientations of gradients are Histogram of Oriented Gradients (HOG) [1], Scale-Invariant Feature Transform (SIFT) keypoint [2], and Aggregated Channel Feature (ACF) [3]. An image pyramid [4] is a collection of resampled images from an original image; the pyramid is used to make a computer vision problem invariant over multiple scales. Many object detectors (e.g., ACF-AdaBoost [3], and Viola and Jones [5]), scan a detection window of a fixed size over an image pyramid.

However, interpolation while constructing the image pyramid usually causes a difference between the gradients of the original image and the scaled image [6]. When pixels in downsampled images are computed using a bilinear function over corresponding pixels, the intensities of the pixels have similar distribution and magnitude, but the skipped pixels in downsampled images increase gradients (the first derivative of intensity). In contrast, the inserted pixels in upsampled images decrease the gradients. We define this difference between original image and scaled image as gradient variation.

Our method is inspired by the gradient variation that causes the decrease in accuracy of the classifiers. The increased variance of gradients over the image pyramid increases the coverage of the classifier. Thus, the increased coverage decreases the accuracy and precision of the classifiers [7, 8, 9, 10].

Hence, we propose a simple and novel gradient normalization method by analyzing the gradient variation in the viewpoint of the classifier (Fig. 1). The proposed method defines the original image as reference, and normalizes gradients from other resampled images to the reference image. The normalized gradient, which is similar to the gradients of original images, reduces the variance, and increases the performance of the classifiers with negligible increase in computing time.

2. GRADIENT VARIATION IN MULTI-SCALE

In this section, we discuss the change of gradients that occurs in an image pyramid, which is used to apply the fixed-size detector to multi-scale detection.

2.1. Analysis of Gradient in Multi-Scale

We compared the difference between original images and scaled images that include objects of the same identity, and observed that the first derivatives of intensity are greater in downsampled images than in the original images, even if the distributions of intensity are similar [3, 6, 11].

We theoretically show the gradient difference by computing the ratio of gradient expectations between the two images under three conditions: computing gradient using a central difference method [12], sampling images using a bilinear interpolation [12], and decomposing the problem to a one-dimensional form.

Let an image f_s be a sequence that consists of the pixels $f_s(x)$ from an upsampled image with scale s. $f_{r=1.0}$ is a reference image, which is original and the only natural data in an image pyramid. A linear interpolation computes an upsampled image f_s by inserting z new pixels between two adjacent
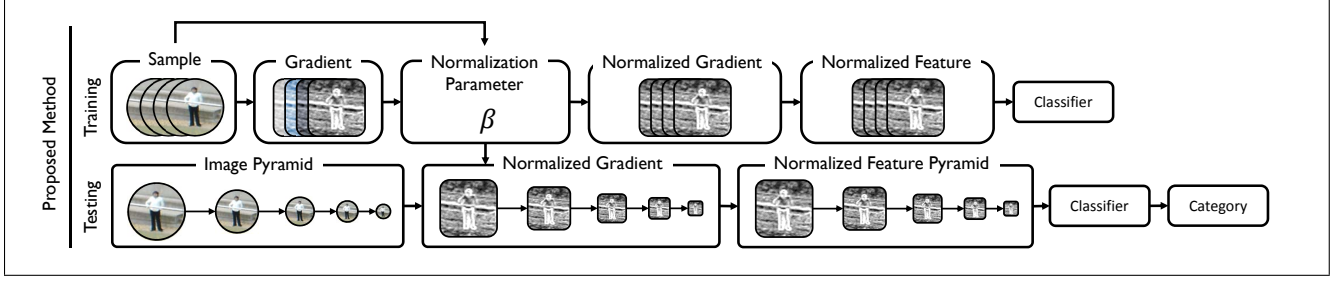
Fig. 1: The proposed method constructs an image pyramid, and computes normalized gradients using the proposed normalization function. Unlike an image pyramid and a fast feature pyramid, the proposed method enhances the quality of samples in both training and testing to improve the accuracy of classifiers.

pixels on the original image. The pixels on the upsampled image are partitioned into inherited pixels and interpolated pixels. The number of inserted pixels between two adjacent inherited pixels is $z = s - 1$ and is therefore an integer ≥ 0. The pixels in an upsampled image consist of a set of inherited pixels and z sets of inserted pixels, so the total number of pixels is $n_s = (n_r - 1)z + n_r$. The pixels in an upsampled image is approximated as

$$f_s(x) = \frac{(z + 1 - d)f_r(x) + df_r(x+1)}{z+1}, \quad (1)$$

where d is the distance between x and the nearest inherited pixel leftward.

We use a central difference function $\phi(f_s)$ as a gradient function and an intermediate difference function $\tilde{\phi}(f_s)$ is appeared in the calculation of gradient expectation, by substituting. When a gradient is computed at x, $\tilde{\phi}(f_s)$ subtracts the pixel at x from the adjacent pixel at $x+1$, and $\phi(f_s)$ subtracts the neighbor pixels at $x - 1$ and $x + 1$:

$$
\begin{aligned}
\tilde{\phi}(f_s) &= \left[\frac{\partial f_s}{\partial x}\right]_{\Delta=1} \approx |f_s(x+1) - f_s(x)| \\
\phi(f_s) &= \left[\frac{\partial f_s}{\partial x}\right]_{\Delta=2} \approx |f_s(x+1) - f_s(x-1)|,
\end{aligned}
\quad (2)
$$

where Δ is an interval of a differential.

To prove the existence of the gradient difference, we compute the gradient expectation $E[\phi(f_s)]$ at scale s:

$$
\begin{aligned}
E[\phi(f_s)] =& \frac{1}{(z+1)n_r - z - 2} \left\{ \sum_{x=2}^{n_r-1} \frac{|f_r(x+1) - f_r(x-1)|}{z+1} \right. \\
& \left. + \sum_{x=1}^{n_r-1} \frac{2z|f_r(x+1) - f_r(x)|}{z+1} \right\} \\
=& \frac{(n_r - 2)E[\phi(f_r)] + (2z(n_r - 1))E\left[\tilde{\phi}(f_r)\right]}{(z+1)((z+1)n_r - z - 2)}.
\end{aligned}
\quad (3)
$$

The input images that are used in object detection typically have enough pixels to assume that n_r is infinite. $E[\phi(f_s)]$ is approximated as $\lim_{n_r \to \infty} E[\phi(f_s)]$:

$$
\begin{aligned}
E[\phi(f_s)] &\approx \lim_{n_r \to \infty} E[\phi(f_s)] \\
&= \frac{E[\phi(f_r)] + 2(s-1)E\left[\tilde{\phi}(f_r)\right]}{s^2}.
\end{aligned}
\quad (4)
$$

Eq. 4 reveals that a gradient difference between the upsampled and reference image exists, and is determined by scale and the gradient expectation of an intermediate difference function $\tilde{\phi}(f_s)$.

2.2. Formulation of Gradient Variation

We define gradient variation as the difference of gradient expectation between an original image and a scaled image, and formulate the variation as the ratio of the gradient expectations. We formulate the equations for the integer variable z, however, the practical algorithm estimates the real value of z through nearest neighbor or linear approximation. With the same concept, we expand the equations of the gradient variation to a real value. Gradient variation $\rho(f_s|f_r)$ between the upsampled image f_s and the reference image f_r is computed as

$$\rho(f_s|f_r) = \frac{E[\phi(f_s)]}{E[\phi(f_r)]} = \frac{2cs - 2c + 1}{s^2} \quad (5)$$

where $c = E[\tilde{\phi}(f_r)]/E[\phi(f_r)]$ is a constant. Because Eq. 5 is only available for upsampled images due to the definition of f_s, we replace the reference image and the scaled image with each other to represent downsampling. We invert s to re-define it to the range $(0, 1]$, then calculate the inverse of $\rho(f_s|f_r)$ as

$$\overline{\rho}(f_s|f_r) = \frac{1}{\rho(f_{1/s}|f_r)} = \frac{1}{(1 - 2c)s^2 + 2cs}. \quad (6)$$

The practical interval $[1, 2)$ of s for upsampling has a smaller rate of change than the interval $(0, 1]$ of s for downsampling, and the constant c is close enough to 0.5 for degree reduction ($\mu_c = 0.62$ and $\sigma_c = 0.05$ in INRIA dataset). We

approximate the last term as $1 - 2c \approx 0$ in the numerator for upsampling, to simplify the gradient variation $\rho(f_s|f_r)$.

The gradient variation for resampled images are computed as

$$
\begin{aligned}
&\rho(f_s|f_r) \\
&= \begin{cases} \rho(f_s|f_r) \approx (2c/s) & 1 < s \\ \bar{\rho}(f_s|f_r) = 1/\{(1-2c)s^2 + 2cs\} & 0 < s \le 1 \end{cases} . \quad (7)
\end{aligned}
$$

Eq. 7 shows that $\rho(f_s|f_r)$ is a decreasing function. These trends imply that upsampling decreases gradients and downsampling increases gradients. This phenomenon implies that the gradient distribution of the resampled images is different from the gradient distribution of the reference images; the increased variance increases the difficulty of training the classifiers [7, 13].

3. GRADIENT NORMALIZATION

We propose a normalization method to eliminate the gradient variation. The proposed method normalizes the gradients of the resampled image to the gradients of the reference image to reduce the variance of gradients. The reduced variance makes the classifier concentrate on a small coverage, and improves overall precision and accuracy of detection [7, 8, 9, 10]. We obtain the gradient normalization function $g(s)$ as the inverse of the gradient variation $\rho(f_s|f_r)$ as

$$
\begin{aligned}
g(s) &= \frac{1}{\rho(f_s|f_r)} \\
&= \begin{cases} s/(2c) \approx a_1 s + b_1 & 1 < s \\ (1-2c)s^2 + 2cs \approx a_2 s^2 + b_2 s + c_2 & 0 < s \le 1 \end{cases} .
\end{aligned} \quad (8)
$$

with a bias term: b_1 and c_2.

The normalization function consists of polynomials of degree 1 for upsampling and of degree 2 for downsampling. We compute the optimal coefficients of $g(s)$ for the training set. Given a training image f^k, we define an error criterion $\mathcal{E}$, which is a mean squared error to minimize the difference between the normalized gradient and the reference gradient:

$$
\mathcal{E} = \sum_{s \in S} \sum_{k \in K} \left[\phi(f_s^k) g(s) - \phi(f_r^k) \right]^2, \quad (9)
$$

where S is a set of scales and K is a set of training images.

The separate training of the normalization functions $g(s)$ for upsampling and downsampling requires an equality constraint. We impose an equality constraint between original and normalized gradients at the reference scale. The equality constraint prevents gradient normalization at reference images and keeps the continuity of the gradient normalization function at reference image, and is defined as

$$
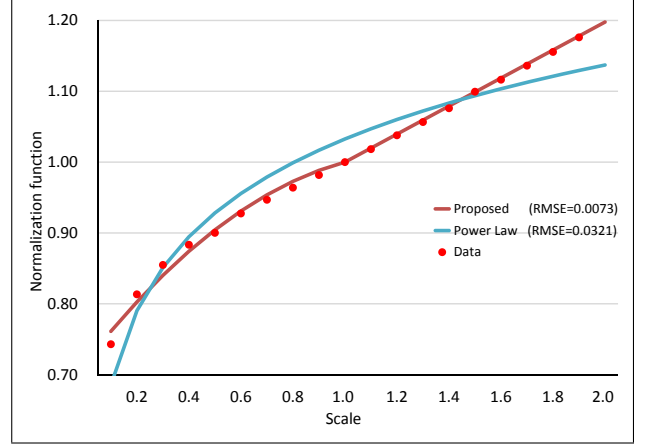\sum_{k \in K} \left[\phi(f_r^k) g(r) - \phi(f_r^k) \right] = 0. \quad (10)
$$

Fig. 2: Illustration on the collected data for the normalization function from scale 0.1 to scale 2.0, and on the estimated value using the proposed function and a power law. Our normalization function fits the data over every scales, whereas the power law fails the extremes; our function also has smaller RMSE (0.0073) than RMSE (0.0321) of the power law. The data is collected in INRIA dataset.

The error criterion and the equality constraint is combined into a Lagrangian

$$
\begin{aligned}
&L(\beta_d, \beta_u, \lambda_d, \lambda_u) \\
&= \begin{cases} \|X_d \beta_d - y_d\|^2 + \lambda_d (X_r \beta_d - y_r), & 1 < s \\ \|X_u \beta_u - y_u\|^2 + \lambda_u (X_r \beta_u - y_r), & 0 < s \le 1 \end{cases},
\end{aligned} \quad (11)
$$

where subscripts $i = d, u$ and r represent downsampled, upsampled and reference, respectively, X_i are Vandermonde matrices of scales, β_i are coefficients of the proposed polynomial equation, y_i are vectors of the ratio of gradients, and λ_i are Lagrange multipliers. The optimal coefficients β^* of $g(s)$ are computed by minimizing the Lagrangian [14].

We compared the fitting accuracy of the gradient normalization between the proposed function and a power law function (Fig. 2). A power law was dealt with to represent the study of natural image statistics by Ruderman and Bialek [6] and Dollar et al. [3].

4. EXPERIMENTS

We show the effectiveness of the gradient normalization in object detection with three applications: pedestrian detection, pose estimation, and object detection.

4.1. Pedestrian Detection

ACF [3, 15] is widely used for pedestrian detection [16, 17]. In this paper, we build **ACF++**, which is a simplified version of the filtered channel features based detector [17]. We combine the original ACF and the differences between two
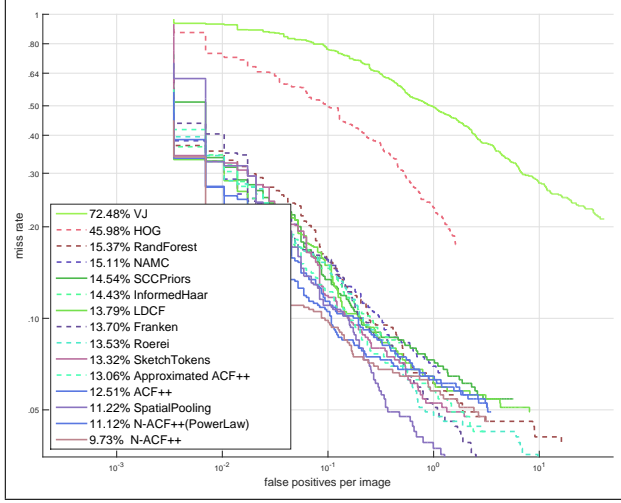
Fig. 3: The log-average miss rate of **ACF++**, **Approximated ACF++**, **N-ACF++(PowerLaw)**, **N-ACF++** on INRIA dataset.

neighboring features, which are part of the checkerboards filters. **Approximated ACF++** is a version of **ACF++** with a fast feature pyramid. We evaluate ACF++ using normalized gradient (**N-ACF++**) on INRIA dataset [1]. **N-ACF++** is trained in the same way as **ACF++** without gradient normalization. To train the gradient normalization function, we collected all gradient expectations of both positive and negative images over scales from 0.1 to 2.0 in increments of 0.1. We applied our normalization method in both training and testing, and we only normalized gradient magnitudes to naturally spread out over the gradient-based features such as HOG. **N-ACF++(PowerLaw)** is a version of **N-ACF++** trained by a power law function. The proposed normalization method with **ACF++** shows the improvement from 12.51% to 9.73% log average miss rate (Fig.3).

4.2. Pose Estimation

Yang et al. [18, 19] proposed flexible mixtures of parts model (**FMM**) to estimate human poses. Each appearance model is trained as a filter of HOG [1] based features that consist of contrast-sensitive HOG, contrast-insensitive HOG, and magnitudes. We evaluate the normalized **FMM** (**N-FMM**) on PARSE dataset [20]. As the negative images, we used the INRIA dataset [1]. We achieved 2%p overall improvement on probability of correct keypoint (Table 1).

	Avg	Head	Shou	Elbo	Wris	Hip	Knee	Ankle
FMM [18]	72.3	89.0	85.3	66.0	46.3	76.5	76.3	66.3
N-FMM	74.2	91.0	86.8	67.6	49.5	80.2	77.6	66.8

Table 1: Probability of correct keypoint for **FMM** and **N-FMM** (using normalized gradients) on PARSE dataset.

4.3. Object Detection

The deformable part model (**DPM**) from Felzenszwalb et al. [21, 22] is a representative approach for object detection. **DPM** consists of mixtures of multiscale deformable part models that are trained using partially labeled data, and each part model includes appearance and spatial models. Appearance models are trained as a filter of HOG [1] based features that consist of contrast-sensitive HOG, contrast-insensitive HOG, and magnitudes. We evaluate the normalized **DPM** (**N-DPM**) on PASCAL 2007 dataset [23]. We achieve 1%p overall improvement and 4.4%p maximum improvement in average precision scores [23] (Table 2).

	DPM	N-DPM		DPM	N-DPM
plane	33.3	34.2	table	24.6	**27.3**
bike	59.7	**60.7**	dog	12.2	12.5
bird	10.4	10.8	horse	56.4	57.0
boat	15.5	**16.6**	mbike	47.7	**48.9**
bottle	27.1	27.2	person	42.6	43.2
bus	51.2	**52.8**	plant	14.3	14.5
car	58.2	58.2	sheep	18.6	**23.0**
cat	23.9	**25.5**	sofa	37.6	37.8
chair	19.9	**21.3**	train	45.5	**46.8**
cow	25.1	25.7	tv	43.4	43.5

Table 2: Average precision scores for **DPM** and **N-DPM** (using normalized gradients) on PASCAL VOC 2007.

5. CONCLUSION

Our research reinterprets the gradient variation in the viewpoint of the classifier. Unlike conventional approaches concentrating on computing resized images, our approach concentrates on decreasing the coverage of the classifier to enhance the focus of the classifier. We prove the existence of the gradient variation by formulating the ratio of gradient expectations between an original image and scaled images, then estimate a normalization function to eliminate the effect of this variation. Our calculations and experiments prove the validity of the gradient normalization function. The proposed method is not restricted to object-detection applications, but can be applied in many gradient-based studies with negligible cost of computing time. We will adopt our study to deep learning based features.

6. ACKNOWLEDGEMENT

7. REFERENCES

[1] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 886–893.

[2] David G Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. Ieee, 1999, vol. 2, pp. 1150–1157.

[3] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona, "Fast feature pyramids for object detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 8, pp. 1532–1545, 2014.

[4] Edward H Adelson, Charles H Anderson, James R Bergen, Peter J Burt, and Joan M Ogden, "Pyramid methods in image processing," *RCA engineer*, vol. 29, no. 6, pp. 33–41, 1984.

[5] Paul Viola and Michael J Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[6] Daniel L Ruderman and William Bialek, "Statistics of natural images: Scaling in the woods," *Physical review letters*, vol. 73, no. 6, pp. 814, 1994.

[7] Stuart Geman, Elie Bienenstock, and René Doursat, "Neural networks and the bias/variance dilemma," *Neural computation*, vol. 4, no. 1, pp. 1–58, 1992.

[8] Michael Jones and Paul Viola, "Fast multi-view face detection," *Mitsubishi Electric Research Lab TR-20003-96*, vol. 3, pp. 14, 2003.

[9] Dennis Park, Deva Ramanan, and Charless Fowlkes, "Multiresolution models for object detection," in *Computer Vision–ECCV 2010*, pp. 241–254. Springer, 2010.

[10] Junjie Yan, Xucong Zhang, Zhen Lei, Shengcai Liao, and Stan Z Li, "Robust multi-resolution pedestrian detection in traffic scenes," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3033–3040.

[11] Piotr Dollár, Serge Belongie, and Pietro Perona, "The fastest pedestrian detector in the west.," in *BMVC*. Citeseer, 2010, vol. 2, p. 7.

[12] Rafael C Gonzalez, *Digital image processing*, Pearson Education India, 2009.

[13] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An introduction to statistical learning*, vol. 112, Springer, 2013.

[14] Richard O Duda, Peter E Hart, and David G Stork, *Pattern classification*, John Wiley & Sons, 2012.

[15] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona, "Fast feature pyramids for object detection, code," http://vision.ucsd.edu/ pdollar/toolbox/doc/.

[16] Woonhyun Nam, Piotr Dollár, and Joon Hee Han, "Local decorrelation for improved pedestrian detection," in *Advances in Neural Information Processing Systems*, 2014, pp. 424–432.

[17] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele, "Filtered feature channels for pedestrian detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1751–1760.

[18] Yi Yang and Deva Ramanan, "Articulated human detection with flexible mixtures of parts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 12, pp. 2878–2890, 2013.

[19] D. Ramanan Y. Yang, "Articulated pose estimation with flexible mixtures of parts, code," http://www.ics.uci.edu/ dramanan/software/pose/.

[20] Deva Ramanan, "Learning to parse images of articulated bodies," in *Advances in neural information processing systems*, 2006, pp. 1129–1136.

[21] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.

[22] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester, "Discriminatively trained deformable part models, release 5," http://people.cs.uchicago.edu/ rbg/latent-release5/.

[23] 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.