PEDESTRIAN DETECTION VIA PCA FILTERS BASED CONVOLUTIONAL CHANNEL FEATURES

Wei Ke *Yao Zhang* *Pengxu Wei* *Qixiang Ye* *Jianbin Jiao*

School of Electronics, Electrical and Communication Engineering
University of Chinese Academy of Sciences, Beijing, China

ABSTRACT

In this paper, we propose a kind of image representation, named PCA filters based convolutional channel features (PCA-CCF) for pedestrian detection. The motivation is to use the convolutional network architecture with orthogonal PCA filters to enhance the state-of-the-art aggregate channel features (ACF). In PCA-CCF, the convolutional operation improves the feature robustness to pedestrian local deformation. The learned PCA filters reduce the correlations among features of each channel, and therefore, improve feature discrimination capability. With the proposed PCA-CCF features and cascaded AdaBoost classifiers, we develop a coarse-to-fine pedestrian detection approach. Experiments show that such approach achieves 3.04%, 17.87% and 6.28% performance gain on the INRIA, Caltech Reasonable and Caltech Overall pedestrian datasets, respectively.

Index Terms— Pedestrian detection, Channel features, P-CA, Convolutional network

1. INTRODUCTION

Pedestrian detection in natural scene images contributes to a variety of applications, including robotics, intelligent transportation and video surveillance systems. Although extensively investigated, the low performance on public benchmarks indicates that image based pedestrian detection remains an open problem [1–3]..

Feature representation has been considered one of the most critical factors in the pedestrian detection problem, and exploring image features that can effectively and efficiently discriminate pedestrians from the clutter backgrounds has been the focus of the community. Hand-craft image features developed for pedestrian detection include Haar-like features [4], Histogram of Oriented Gradients (HOG) [5], v-HOG [6],covariance features [7], Local Binary Pattern (LBP) [8], HOG-LBP [9], HOG-SURF [10], Edgelet [11], Shapelet [12], Multi-Scale Orientation (MSO) [13], and pose-invariant descriptors [14]. A recent research demonstrates superior detection performance and efficiency when using integral features from multiple channels of color, gradient and orientation [15]. The features extracted from multiple channels are called Aggregate Channel Features (ACF) [1, 15].

On the other hand, various learning based features have attracted attentions in recent years. P. Sermanet et al. proposed to learn multi-stage features for pedestrian detection using convolutional neural network (CNN) [16]. Ren et al. proposed to incorporate unsupervised learning to extract sparse coding histogram features [17]. Lim et al. proposed to use unsupervised learning to extract mid-level features that precisely capture pedestrian contours [18]. Learning based features leverage unsupervised or back-propagation algorithms to specify object representations, and demonstrate excellent robustness and adaptability. Despite superior performance, these learned features are not as descriptive as hand-craft features, and are often computationally expensive. Some back-propagation based feature learning algorithms, i.e., CNN, often require a large number of training samples to guarantee performance.

In this paper, we propose the PCA filters based convolutional channel features (PCA-CCF) for pedestrian detection. Such work is rooted in the success of integration of multiple channel features, and is also inspired by the success of a simple deep learning method, PCANet [19], which extracts effective features using PCA filters and convolutional operations. Our contribution is applying PCA filters to reduce the correlations among feature channels, and use convolutional operations to improve robustness of features. Without using the back-propagation, PCA-CCF is not sensitive to the number of training samples. We also propose using channel pooling to compensate conventional spatial pooling. By these strategies, PCA-CCF incorporates both the discriminative property of learned features and the descriptive capability of hand-craft features.

In pedestrian detection, the conventional ACF is firstly used to train a coarse detector, which is used to generate a set of candidate windows. The candidate windows have a high recall rate on pedestrians, but include lots of false detections. PCA-CCF is then extracted on these windows and fed to a cascaded AdaBoost classifier to perform fine classification.

The remainder of this paper is organized as follows. In section 2, the PCA filters based convolutional channel features extraction is described. In section 3, the pedestrian de-

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

tection framework is elaborated. In section 4, experimental results are presented, and in section 5 we conclude the paper with discussion of future directions.

2. FEATURE EXTRACTION

To make the paper self-contained, the aggregate channel features (ACF) are first reviewed, and the procedures of learning PCA filters and extraction of convolutional channel features are then described.

2.1. Aggregate Channel Features

Given a sample image I, ACF computes channels using linear or non-linear transformation of I as

$$f = \Omega(I), \tag{1}$$

where Ω is a first-order function as a sum of pixels in a rectangular image region or higher-order functions that are computed using multiple first-order in a single channel. Downsampling by 2×2 is used to reduce the size of channel maps. As shown in Fig.1, ACF has ten channels: three color channels (L, U and V channels), a gradient magnitude channel (|G| channel), histogram of oriented gradients channels (G_1-G_6 channels). The ACF code is available online[1].

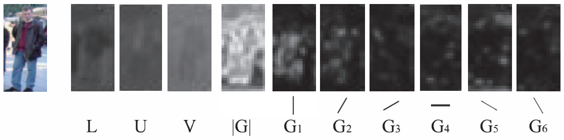


Fig. 1. Aggregate Channel Features of a pedestrian sample.

Because the first-order feature can be computed efficiently using integral images, extracting ACF takes about 0.05~0.2s per 640×480 image, depending on the Ω selected.

2.2. Learning PCA filters

Compared with PCANet filters derived from raw image patches, PCA-CCF learns a filter group for every ACF channel map. Given N training pedestrian sample images $\mathbf{I} = \{I_i\}, i = 1, 2, \cdots, N$ of size $w \times h$, their ACF is denoted as $\{f_{i,k}\}, i = 1, 2, \cdots, N, k = 1, 2, \cdots, K$. For the kth channel of a pedestrian sample image, we take a $m \times m$ patch $x_{i,k}$ around each pixel to obtain patches $X_{i,k} = [x_{1,k}, x_{2,k}, \cdots x_{wh,k}] \in \mathbf{R}^{mm \times wh}$. Collecting all patches of the kth channel of $\mathbf{I}$, we get

$$X_k = [X_{1,k}, X_{2,k}, \cdots X_{N,k}] \in \mathbf{R}^{mm \times Nwh}. \tag{2}$$

PCA uses an orthogonal transformation to convert a set of possibly correlated variables into a set of values of linearly uncorrelated. It minimizes the reconstruction error within a set of orthogonal vectors. Applying PCA for each feature channel, we get

$$\begin{aligned} &\min_k \left\| X_k - VV^T X_k \right\|^2 \\ &s.t. VV^T = I_L, k = 1, 2, \cdots, K, \end{aligned} \tag{3}$$

where I_L is an identity matrix of size $L \times L$. The solution V of Eq.(3) is a matrix combining principle eigenvectors. The eigenvectors are reshaped to 2-dimensional PCA filters of $m \times m$ pixels, which are expressed as

$$P_{l,k} = mat(V_i) \in R^{m \times m}, l = 1, 2, \cdots L, \tag{4}$$

where $mat(\cdot)$ is a function that reshapes a vector to a matrix.

We experimentally set the size of PCA filters as 5×5 pixels ($m \times m = 5 \times 5$). Ten groups of learned PCA filters are shown in Fig. 2. According to PCA theory, top PCA filters contain the most information of the samples (channel features). It can be seen that the orientations of the first PCA filters in the HOG channels (G_1-G_6 channels) are consistent with the quantized orientations of G_1-G_6 channels.

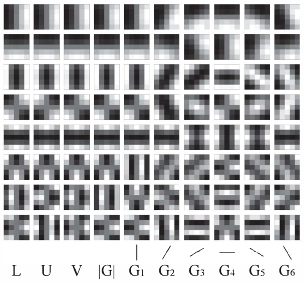


Fig. 2. Visualization of top PCA filters of ten channels. Each group (column) shows top eight eigenvectors.

2.3. Convolutional Channel Features

With learned PCA filters, the ACF $\{f_{i,k}\}, i = 1, 2, ..., N, \ k = 1, 2, ..., K$ are convoluted to calculate the PCA-CCF. This procedure has three steps: convolution with top L PCA filters, spatial local pooling and channel pooling, as shown in Fig. 3. We take spatial local pooling and channel pooling instead of hashing and histogram as in PCANet.

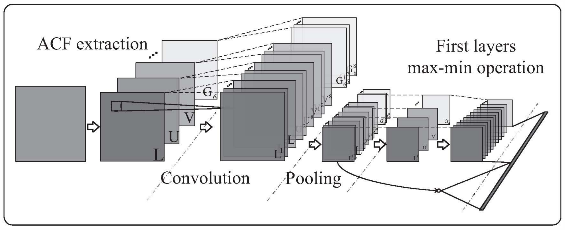


Fig. 3. PCA-CCF extraction.

Convolution: Each ACF channel convolves with L PCA filters, and outputs L convolutional layer maps, as

$$CONV_{l,k} = P_{l,k} \circ f_k, l = 1, 2, ..., L, \ k = 1, 2, ..., K \tag{5}$$

[1] http://vision.ucsd.edu/ pdollar/toolbox/doc/

where $P_{l,k}$ is the lth PCA filter of the kth channel, f_k is the kth channel feature map. After convolution, the number of ACF feature maps (channels) increases to $L \times K$.

Spatial Local Pooling: Similar to the CNN, a pooling operation is applied to reduce the feature dimensionality, as well as to achieve robustness to local variations. On the convolutional layer maps, the max and min pooling operations [20]are performed to calculate spatially pooled feature maps, as

$$\begin{cases} FS\_MAX_{l,k}(i,j) = \max\{CONV_{l,k}(R(i,j))\} \\ FS\_MIN_{l,k}(i,j) = \min\{CONV_{l,k}(R(i,j))\} \end{cases}, \quad (6)$$

where $R(i,j)$ is a region surrounding pixel (i,j) in the lth convolutional layer of the kth channel. As ACF conducts 2×2 down-sampling, the dimensionality of all spatially pooled maps $FS = \{FS\_MAX, FS\_MIN\}$ after spatial local pooling is $2 \times L \times K \times W \times H/4$, where $W \times H$ is the size of an input image.

Channel Pooling: Convolution and pooling are carried out on each channel, which ignores the relations among channels. We propose to use a max-min operation, i.e., channel pooling, between channels to capture such relations. The max operation is similar to operator OR, capturing complementary information of the two channels. The min operation is similar to operator AND, which combines the two channels.

Given $L \times K$ channels, its expensive to use all channel pairs for pooling, which can produce $K \times (K-1)$ channels. Considering that the first PCA filters are the most important, we choose the first convolutional layer maps, corresponding to the first PCA filters, to calculate channel pooling features, as

$$\begin{cases} FC\_MAX_{k^{(1)},k^{(2)}} = \max\{FS_{1,k^{(1)}}, FS_{1,k^{(2)}}\} \\ FC\_MIN_{k^{(1)},k^{(2)}} = \min\{FS_{1,k^{(1)}}, FS_{1,k^{(2)}}\} \end{cases}, \quad (7)$$

where $FS_{1,k^{(1)}}$ and $FS_{1,k^{(2)}}$ denote a channel pair with two first convolutional layer maps (obtained by first PCA filters) of $k^{(1)}$ and $k^{(2)}$ ACF channels.

Finally, the feature maps from spatial pooling and channel pooling are concentrated to form the PCA-CCF as

$$F = \{FS, FC\} \quad (8)$$

3. PEDESTRIAN DETECTION

The calculation of PCA-CCF is much more computationally expensive than the ACF features for the usage of convolutional operations. Therefore, we propose a coarse-to-fine detection strategy, as shown in Fig. 4.

The coarse detector used to localize candidates has the following characteristics: high efficiency, precise localization, and high recall rate. A 3-stage cascaded Adaboost classifier trained with the ACF is employed as the coarse detector. Sliding window classification is applied on image pyramids to perform detection. Using Integral image and a 3-stage cascaded ACF detector, the coarse detection has a high detection efficiency [1]. It also has a high recall rate on pedestrians, but include lots of false detections.

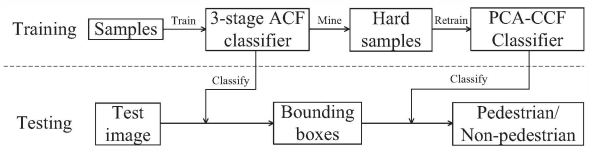


Fig. 4. Pedestrian detection framework.

PCA-CCF is then extracted from these candidate windows and fed to another cascaded AdaBoost classifier to perform final detection. The weak classifiers used for the AdaBoost classifier are decision trees. The reason why we choose decision trees is that channel features are linearly de-correlated by the PCA filters, and therefore, could be coupled to the decision trees with orthogonal splits. When training the fine detector, hard negative samples are mined from the images without pedestrians, and are used to update the training set.

4. EXPERIMENTAL RESULTS

We investigate the performance of the proposed method on two public pedestrian datasets: INRIA and Caltech.

In the coarse pedestrian detection, a 3-stage cascaded ACF detector generates candidate windows, which combines 32, 128 and 512 decision trees in each cascade. Experiments show that the coarse detection procedure has a 95% recall rate on the INRIA dataset, and the speed is 14.3fps. In comparison, the recall rates of three other popular object localization approaches, i.e., Selective Search [21], BING [22] and Edge Boxes [23], are 23%, 61% and 93%, respectively.

In the fine detection, the number of PCA filters is empirically set as $L = 8$, and the size of filters $m = 5$. As shown in Fig.5, with PCA convolution the miss detection rate (1.0-Recall) of the proposed approach is 15.77%. With spatial local pooling and channel pooling, the miss rate reduces to 14.79% and 14.24%, respectively. In comparison with the 3-stage ACF detector of a miss rate 22.00%, the performance

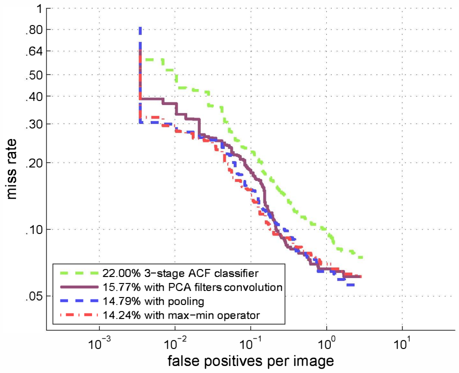
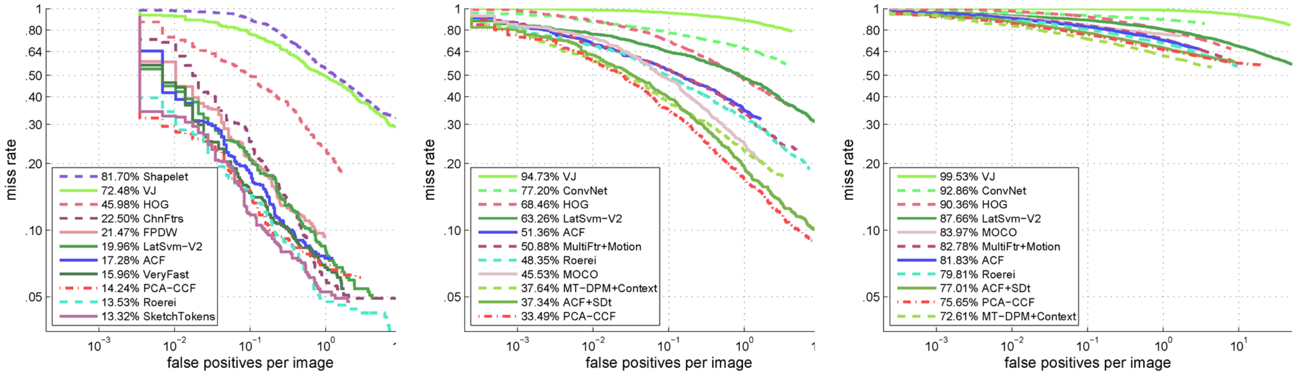


Fig. 5. Validation of the convolutional and pooling operations.



(a) INRIA (b) Caltech Reasonable (c) Caltech Overall

Fig. 6. Comparison of detection performance on the INRIA and the Caltech datasets.

gain is about 8%.

For final detection performance, DET curves, the miss rate versus false positive per image, is employed as the protocol in [1]. We conduct experiments on INRIA and Caltech pedestrian datasets. For all datasets, we compare with three classical methods which are VJ [4], HOG [5], LatSvm-V2 [24], baseline detector ACF [1], and other six top 10 methods [1] which involve Shapelet [12], ChnFtrs [15], ConvNet [16], Sketch Tokens [18], VaryFast [25], FPDW [26], Roerei [27], MultiFtr-Motion [28], MOCO [29], MT-DPM+Context [30], ACF-SDt [31].

The miss rate on INRIA is 14.24% which has an improvement over ACF baseline by 3.04%, as show in Fig.6a. It is comparable with the performance of two best approaches. Caltech pedestrian dataset is a more complicated dataset, with clutter background and large scale variation. In Fig.6b and Fig.6c it can be seen that our approach achieves 17.87% and 6.28% performance gain compared with the ACF based approach on the Reasonable and Overall pedestrians, respectively. On the Caltech Reasonable (pedestrian resolution is reasonable) it reports the best result. On the Caltech Overall (with pedestrians of very low resolution) it reports the second best result. It should be noted that the best approach uses context information. Some of our detection examples are shown in Fig.7.

5. CONCLUSION AND FUTUREWORKS

We propose PCA filters based convolutional channel features (PCA-CCF) using a forward convolutional network. Without any back-propagation operation, PCA-CCF achieves higher discriminative capability than the convolutional neural network (CNN) and the hand-craft Aggregate Channel Features. Based on the proposed PCA-CCF, we propose a coarse-to-fine pedestrian detection framework, which is validated to have comparable performance to several representative approaches on the INRIA dataset, and a significant performance gain on the Caltech dataset.

In the future, it is useful to try other orthogonal filters, i.e., Wavelet and Gabor filters, to further improve the performance of PCA-CCF. It is also interesting to extend the PCA-CCF to other detection or recognition tasks.

6. ACKNOWLEDGEMENT

This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2011CB706900, and in part by the National Science Foundation of China under Grant 61271433 and Grant 61202323.

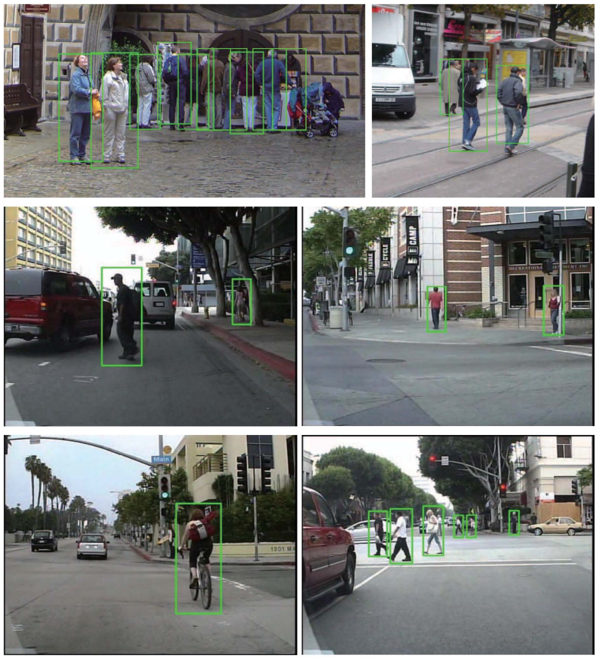


Fig. 7. Detection examples from INRIA dataset (first row), and Caltech dataset (second and third rows).

7. REFERENCES

[1] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *TPAMI*, 2014.

[2] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *TPAMI*, vol. 34, no. 4, pp. 743–761, 2012.

[3] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *TPAMI*, vol. 31, no. 12, pp. 2179–2195, 2009.

[4] P. Viola and M. J. Jones, "Robust real-time face detection," *IJCV*, vol. 57, no. 2, pp. 137–154, 2004.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005, vol. 1, pp. 886–893.

[6] R. Xu, B. Zhang, Q. Ye, and J. Jiao, "Cascaded L1-norm minimization learning (CLML) classifier for human detection," in *CVPR*, 2010, pp. 89–96.

[7] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on Riemannian manifolds," *TPAMI*, vol. 30, no. 10, pp. 1713–1727, 2008.

[8] Y. Mu, S. Yan, Y. Liu, T. Huang, and B. Zhou, "Discriminative local binary patterns for human detection in personal album," in *CVPR*, 2008, pp. 1–8.

[9] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *CVPR*, 2009, pp. 32–39.

[10] J. Liang, Q. Ye, J. Chen, and J. Jiao, "Evaluation of local feature descriptors and their combination for pedestrian representation," in *ICPR*, 2012, pp. 2496–2499.

[11] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors," in *ICCV*, 2005, vol. 1, pp. 90–97.

[12] P. Sabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features," in *CVPR*, 2007, pp. 1–8.

[13] Q. Ye, J. Jiao, and B. Zhang, "Fast pedestrian detection with multi-scale orientation features and two-stage classifiers," in *ICIP*, 2010, pp. 881–884.

[14] Z. Lin and L. S. Davis, "A pose-invariant descriptor for human detection and segmentation," in *ECCV*, 2008, pp. 423–436.

[15] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features.," in *BMVC*, 2009, vol. 2, p. 5.

[16] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multistage feature learning," in *CVPR*, 2013, pp. 3626–3633.

[17] X. Ren and D. Ramanan, "Histograms of sparse codes for object detection," in *CVPR*, 2013, pp. 3246–3253.

[18] J. J. Lim, C. L. Zitnick, and P. Dollár, "Sketch tokens: A learned mid-level representation for contour and object detection," in *CVPR*, 2013, pp. 3158–3165.

[19] T. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "P-CANet: A simple deep learning baseline for image classification," *arXiv preprint arXiv:1404.3606*, 2014.

[20] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *CVPR*, 2014.

[21] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *IJCV*, vol. 104, no. 2, pp. 154–171, 2013.

[22] M. Cheng, Z. Zhang, W. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *CVPR*, 2014.

[23] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *ECCV*, 2014, pp. 391–405.

[24] F. Pedroand, G. Ross, M. David, and R. Deva, "Object detection with discriminatively trained part-based models," *TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[25] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, "Pedestrian detection at 100 frames per second," in *CVPR*, 2012, pp. 2903–2910.

[26] P. Dollár, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *BMVC*, 2010.

[27] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool, "Seeking the strongest rigid detector," in *CVPR*, 2013, pp. 3666–3673.

[28] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," pp. 1030–1037, 2010.

[29] G. Chen, Y. Ding, J. Xiao, , and T. Han, "Detection evolution with multi-order contextual co-occurrence," in *CVPR*, 2013, pp. 1798–1805.

[30] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li, "Robust multi-resolution pedestrian detection in traffic scenes," in *CVPR*, 2013, pp. 3033–3040.

[31] D. Park, C. Lawrence Zitnick, D. Ramanan, and P. Dollár, "Exploring weak stabilization for motion feature extraction," in *CVPR*, 2013, pp. 2882–2889.