ACTION RECOGNITION WITH GRADIENT BOUNDARY CONVOLUTIONAL NETWORK

Huafeng Chen[1,2], *Jun Chen*[1,2], *Chen Chen*[3*], *Ruimin Hu*[1,2]

[1]Research Institute of Shenzhen, Wuhan University, Shenzhen, China
[2]National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan, China
[3]Center for Research in Computer Vision, University of Central Florida, Orlando, USA

ABSTRACT

Deep learning features for video action recognition are usually learned from RGB/gray images, image gradients, and optical flows. The single modality of the input data can describe one characteristic of the human action such as appearance structure or motion information. In this paper, we propose a high efficient gradient boundary convolutional network (ConvNet) to simultaneously learn spatio-temporal feature from the single modality data of gradient boundaries. The gradient boundaries represent both local spacial structure and motion information of action video. The gradient boundaries also have less background noise compared to RGB/gray images and image gradients. Extensive experiments are conducted on two popular and challenging action benchmarks, the UCF101 and the HMDB51 action datasets. The proposed deep gradient boundary feature achieves competitive performances on both benchmarks.

Index Terms— Action recognition, convolutional network, gradient boundary

1. INTRODUCTION

Human action recognition aims to enable computer automatically recognize human action in video through related features [1, 2, 3, 4, 5]. Action features can be divided into two categories: hand-crafted features and deep-learned features. Significant progress has been achieved in recent years by hand-crafted features. For example, Wang et al. designed motion boundary histograms (MBH) feature based on dense trajectories (DT) [6], and further improved dense trajectories (iDT) in [7] through camera motion estimation and elimination. However, the hand-crafted descriptors are not optimized for visual representation and lack discriminative capacity for action recognition [8].

Encouraged by the success of deep learning methods in image classification [9], researchers have exploited the deep-learned features for video action recognition. Current deep learning works for action recognition can be divided into two categories: (1) ConvNets for action recognition, and (2) temporal structure modeling for action recognition [10].

ConvNets for action recognition. Several researchers have attempted to design effective ConvNet architectures for action recognition. Taylor et al. [11] proposed the convGRBM algorithm to learn unsupervised spatio-temporal features by using Gated Restricted Boltzmann Machine (GRBM). Ji et al. [12] extended 2D ConvNet to 3D ConvNet for action recognition. Karpathy et al. [13] evaluated several ConvNet architectures based on stacked RGB images for video classification. Tran et al. [14] explored 3D ConvNet [12] on realistic and large-scale video datasets. Simonyan and Zisserman [15] introduced two-stream architecture which exploits two ConvNets to model static appearance and motion variation of action respectively. Based on two-stream ConvNets and iDT, Wang et al. [8] designed Trajectory-pooled Deep-Convolutional descriptors (TDD) which enjoy the merits of ConvNets and trajectory based methods. Sun et al. [16] proposed a factorized spatio-temporal ConvNet and explored different ways to decompose 3D convolutional kernels. Wang et al. [17] evaluated the very deep two-stream ConvNets for action recognition.

Temporal structure modeling for action recognition. Many works have been devoted to modeling the temporal structure for action recognition based on ConvNet features. Recent works [18, 19, 20, 21] utilized the recurrent Long Short Term Memory (LSTM) architecture to capture temporal structure of consecutive frames. Wang et al. [22] designed a novel representation for actions by modeling action as a transformation which changes the state of the environment before the action happens (precondition) to the state after the action (effect). Feichtenhofer et al. [23] utilized 3D convolutional Pooling method to learn correspondences between highly abstract ConvNet features both spatially and temporally. Fernando et al. [24] used discriminative hierarchical rank pooling for encoding the temporal dynamics of video sequences. Bilen et al. [25] designed dynamic image networks to perform end-to-end training from videos combining both static appearance information from still frames, as well as short and long term dynamics from the whole video. Zhu et al. [26] proposed a key volume mining deep framework

The research was supported by the NSFC (61671332), the Hubei Province Technological Innovation Major Project (2016AAA015), and the Basic Research Program of Shenzhen City (JCYJ20150422150029090).
*Correspondence author

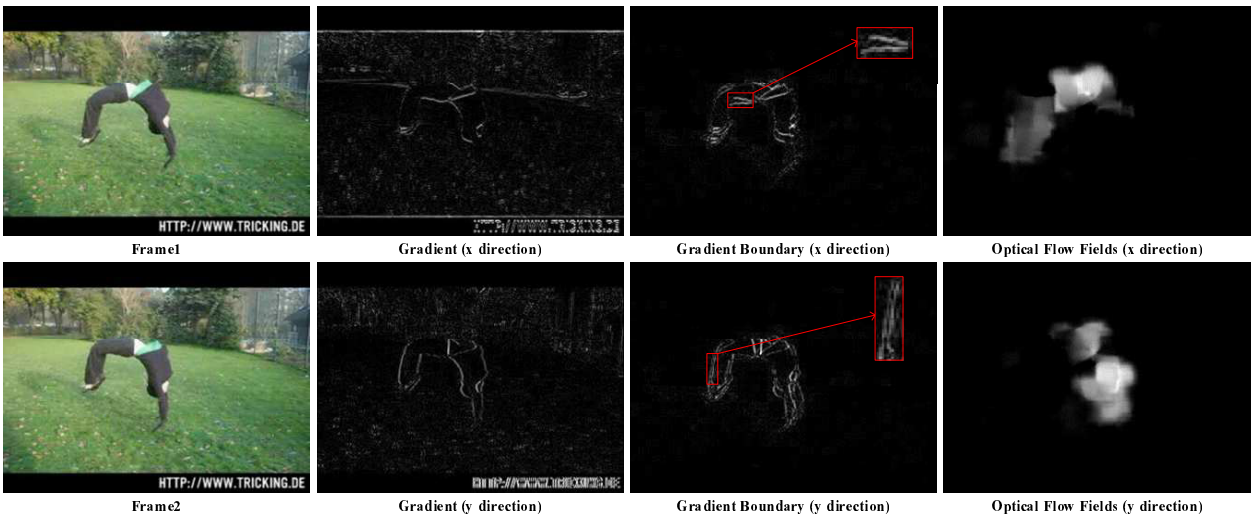

Fig. 1. Illustration of RGB images, image gradients, gradient boundaries, and optical flows for a "flic-flac" action. Compared to image gradients, gradient boundaries encode motion information and have less background noise. The areas in red boxes show the double edges at distances proportional to the speed of moving body parts. The gradient boundaries include appearance structure information compared to optical flows.

to identify key volumes in videos and conducted classification simultaneously. Wang et al. [10] introduced temporal segment network (TSN) for modeling long-range temporal structure in action video.

Motivation and contribution The aforementioned deep learning approaches learn deep features from multi-modality data such as RGB/gray images, image gradients, and optical flows. However, the single modality of the input data can only describe one characteristic of action: spatial structure or motion information. For example, the RGB images and RGB gradients can only describe static appearance of action [15]. The optical flow and its variations, e.g., warped optical flows, merely represent motion information of action [10]. *Can we encode both spatial structure and motion information of action into one single modality?* Inspired by the recent works [27, 15], in this paper we propose the deep gradient boundary ConvNet to simultaneously learn the spatio-temporal information of action based on gradient boundaries. The gradient boundaries encode both local spacial structure and motion information of action into a single modality data. Compared to RGB images and RGB gradients, the gradient boundaries have less background noise. It is also worth mentioning that the gradient boundaries are complementary to the above mentioned input modalities. The proposed deep gradient boundary ConvNet is extensively validated on two challenging action benchmarks, the UCF101 [28] and HMDB51 [29] action datasets and demonstrates its effectiveness as compared with the state-of-the-art deep learning approaches.

The rest of this paper is organized as follows. In Section 2, we present the proposed deep gradient boundary feature in detail. In Section 3, we report the experimental results and comparisons on the UCF101 [28] and HMDB51 [29] datasets. Finally, we conclude this paper in Section 4.

2. THE PROPOSED GRADIENT BOUNDARY CONVNET

Inspired by [27, 15], we develop the gradient boundary ConvNet for learning deep features from gradient boundaries of video sequences. We first introduce the gradient boundaries and then provide the details of the gradient boundary ConvNet architecture.

2.1. Gradient Boundary

In this section, we introduce the gradient boundary. We aim to encode both local static appearance and motion information into one modality. For each frame in a video, we follow [27] and first compute image gradients using simple 1-D [-1,0,1] Sobel masks on both x and y directions. Then, the [-1, 1] temporal filter is applied to two consecutive gradient images. Thus, for each pixel P at the location (u, v) in the gradient boundary of frame t, we have:

$$P_t^x(u,v) = \frac{\partial}{\partial t}(\frac{\partial P}{\partial x}), \quad P_t^y(u,v) = \frac{\partial}{\partial t}(\frac{\partial P}{\partial y}) \qquad (1)$$

The absolute values of gradient boundaries are discretized into the interval from 0 to 255 by a linear transformation.

Therefore, the range of gradient boundaries is the same as that of the RGB/gray images.

Figure 1 illustrates the comparison of RGB images, image gradients, gradient boundaries and optical flows. We have some important observations here. First, the subtraction of two consecutive images gradients results in the removal of image backgrounds. The two gradient images show a lot of background noise, while the gradient boundary images show clear human shapes with far less background noise. Moreover, gradient boundaries encode the moving human shapes. As demonstrated by the red bounding boxes in the figure, the double edges at various distances are proportional to the moving speed of the human body parts. For example, the distance between the haunch double edges is larger than the leg double edges, because the haunch moves faster than the leg at the moment. Compared to optical flows, the gradient boundaries capture local appearance structure information.

2.2. Gradient Boundary ConvNet

We follow optical flow ConvNet in [15] and design the gradient boundary ConvNet. The horizontal and vertical components of the gradient boundary, P_t^x and P_t^y, can be seen as gradient boundary channels (as shown in Figure 1). We stack the gradient boundary channels $P_t^{x,y}$ of L consecutive frames around frame t to form a total of $2L$ input channels by following [15]. Formally, let w and h be the width and height of a video frame, a gradient boundary ConvNet input volume $I_t \in \mathbf{R}^{w \times h \times 2L}$ for frame t is constructed as:

$$\begin{aligned} I_t(u, v, 2k-1) &= P^x_{t+k-1}(u, v), \\ I_t(u, v, 2k) &= P^y_{t+k-1}(u, v) \end{aligned} \tag{2}$$

where $u = [1, w]$, $v = [1, h]$, and $k = [1, L]$. The channels $I_t(u, v, c), c = [1, 2L]$ encode the appearance and motion synchronously over a sequence of L frames. Since the ConvNet requires a fixed-size input, we sample a $224 \times 224 \times 2L$ sub-volume from I_t and feed it into the ConvNet as input. In experiments, we set $L = 10$ by following [15].

3. EXPERIMENTAL RESULTS

In this section, we first introduce the two evaluation datasets and the implementation details used in the experiments. Then, we compare the gradient boundary with RGB image, optical flow, and warped optical flow as input modalities to ConvNet to demonstrate its effectiveness. We also compare the performance of our method with the state-of-the-art action recognition approaches.

3.1. Datasets and Implementation Details

We evaluate the proposed method on two popular action recognition datasets: UCF101 [28] and HMDB51 [29]. The UCF101 dataset contains 101 action classes and there are at least 100 video clips for each class. The whole dataset contains 13,320 video clips. The HMDB51 dataset is a large collection of realistic videos from movies and web videos. The dataset is composed of 6,766 video clips from 51 action categories, with each category containing at least 100 clips. We follow the original evaluation scheme of three training/testing splits and use the average accuracy over three splits as the final recognition performance.

We select recent BN-Inception [30] for gradient boundary ConvNet architecture by following [10]. The mini-batch stochastic gradient descent algorithm is used to learn the network parameters. The batch size is set to 256 and the momentum is set to 0.9. The initial network weights of the ConvNet come from the pre-trained models from ImageNet [31]. We set a smaller learning rate in the experiments by following [10]. For RGB image networks, we initialize the learning rate at 0.001, which decreases to its $1/10$ every 2000 iterations. The whole training procedure stops at 4500 iterations. For gradient boundary and optical flow networks, the learning rate is initialized as 0.005 and reduces to its $1/10$ after 12,000 and 18,000 iterations. The maximum iteration is set to 20,000. We follow [10, 15] and use the techniques of location jittering, horizontal flipping, corner cropping, and scale jittering for data augmentation. The action recognition system is implemented with Caffe [32] and OpenMPI to speed up the training process.

3.2. Evaluation of Gradient Boundary ConvNet

In this section, we focus on the evaluation of the gradient boundary ConvNet. We compare it in terms of computational efficiency and recognition accuracy with the RGB image, optical flow, and warped optical flow ConvNet. We select the TVL1 [33] for computing optical flow. According to [7], we extract the warped optical flow by estimating homography matrix and compensating camera motion.

Computational efficiency. We analyze the speed of computing the input modality data since the same ConvNet architecture is used in the experiments. The speed is measured as frames per seconds (fps) on a single-core CPU (E5-2640 v3) and a Titan X GPU. The results are reported in Table 1. We see that the speed of RGB image is very fast because it only depends on the speed of image decoder. Based on RGB image, the calculation of gradient boundary only adds a few simple operations such as Sobel and subtraction. So the speed of gradient boundary is also fast (695fps on UCF101 and 679fps on HMDB51), and satisfies the requirement of real-time applications. Optical flow calculation is a time-consuming operation, even it runs on the GPU (23fps on UCF101 and 19fps on HMDB51). The speed of warped optical flow is about half of optical flow since it doubles the process of optical flow calculation.

Action recognition accuracy. We evaluate the accuracy performance of gradient boundary ConvNet from three as-

Table 1. Speed comparison of computing the input modality data. UCF101: 320×240, HMDB51: 360×240.

Modality	UCF101	HMDB51
RGB image (CPU)	1533fps	1358fps
Gradient boundary (CPU)	695fps	679fps
Optical flow (GPU)	23fps	19fps
Warped optical flow (GPU)	11fps	9fps

Table 2. Accuracy comparison. RGB: RGB image, OF: optical flow, WOF: warped optical flow, GB: gradient boundary.

Network Architecture(Modality)	UCF101	HMDB51
ConvNet(RGB)	83.9%	49.6%
ConvNet(OF)	87.5%	58.2%
ConvNet(WOF)	87.0%	57.5%
ConvNet(GB)	**89.3**%	**60.8**%
ConvNet(GB+RGB)	91.3%	63.7%
ConvNet(GB+OF)	91.8%	65.4%
ConvNet(GB+WOF)	91.4%	65.1%
ConvNet(GB+RGB+OF+WOF)	**93.2**%	**68.5**%
TSN(RGB+OF+WOF) [10]	94.2%	69.4%
TSN(GB+RGB+OF+WOF)	**95.3**%	**71.9**%

pects: single modality, combination of multi-modalities, combination of multi-modalities plus TSN [10] for modeling temporal information among frames. The experimental results are reported in Table 2.

From the top of Table 2, we can see that the gradient boundary ConvNet performs the best among four single input modalities on both datasets since it encodes spatial and temporal information of action simultaneously. The highest accuracy improvements occur between the gradient boundary and the RGB image (i.e., 89.3% vs. 83.9% on the UCF101 and 60.8% vs. 49.6% on the HMDB51), which verify that the motion is the most important information of video action.

The complementary nature of the gradient boundary with other modalities is verified in the middle of Table 2. Compared to single gradient boundary modality, the combination with other modalities improves the recognition accuracy at least by 2% on UCF101 and 2.9% on HMDB51. The combination of all four modalities obviously boosts the performances over the combination of any two modalities, which indicates that the spacial/temporal informations from different modalities are still complementary with each other.

We also evaluate the combination performance of the gradient boundary on the recent TSN architecture [10] which models temporal structure among frames in action video. The results are listed at the bottom of Table 2. Compare to the best results in [10], the addition of the gradient boundary further improves the recognition performances by 1.1% on the UCF101 and 2.5% on the HMDB51. It clearly demonstrates the flexibility and effectiveness of gradient boundary for any deep learning framework.

Table 3. Comparison of the proposed method with the state-of-the-art approaches.

Method	UCF101	HMDB51
iDT [7]	86.0%	60.1%
Two Stream [15]	88.0%	59.4%
TDD [8]	90.3%	63.2%
MDI [25]	89.1%	65.2%
Hierarchical Rank Pooling [24]	91.4%	66.9%
Actions Transformations [22]	92.4%	62.0%
Convolutional Fusion [23]	92.5%	65.4%
Key-volume Mining [26]	93.1%	63.3%
TSN(RGB+OF+WOF) [10]	94.2%	69.4%
ConvNet(GB)	89.3%	60.8%
ConvNet(GB+RGB+OF+WOF)	93.2%	68.5%
TSN(GB+RGB+OF+WOF)	**95.3**%	**71.9**%

3.3. Comparison with the State-of-the-art

We also compare the proposed method against the state-of-the-art approaches in Table 3. The iDT [7] is the best action feature in traditional hand-drafted methods. The Two-Stream ConvNet [15] is the first work to learn the deep spatio-temporal feature by two-layer ConvNet architectures and has a great influence on later deep learning methods for action recognition. The TDD [8] enjoys the merits of deep ConvNets and hand-crafted dense trajectory. [25, 24, 22, 23, 26, 10] represent the latest and state-of-the-art deep learning approaches for action recognition.

Our ConvNet(GB) outperforms the Two Stream ConvNet [15] (using RGB image and optical flow), since we employ a deeper ConvNet architecture. This implies deeper network architectures are better than shallow networks. ConvNet(GB) also achieves competitive performances as compared with the state-of-the-art methods, e.g., MDI [25] and TDD [8]. However, by incorporating other data modalities (e.g., RGB, OF and WOF), the performance can be further improved. Specifically, ConvNet(GB+RGB+OF+WOF) outperforms other methods such as [24, 22, 23]. This is also verified by the comparison of TSN(RGB+OF+WOF) [10] and TSN(GB+RGB+OF+WOF) using the same TSN network, where TSN(GB+RGB+OF+WOF) obtains the new state-of-the-art results on both datasets.

4. CONCLUSION

In this paper, we proposed a high efficient gradient boundary ConvNet for action recognition. As an input modality, the gradient boundaries represent both local spacial structure and motion information of action. The gradient boundaries are also complementary to RGB images and optical flows. Experimental results on two public databases have demonstrated the superiority of the proposed method over some state-of-the-art methods.

5. REFERENCES

[1] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz, "Fusion of depth, skeleton, and inertial data for human action recognition," in *ICASSP*, 2016.

[2] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," in *WACV*, 2015, pp. 1092–1099.

[3] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *ICIP*, 2015.

[4] Chen Chen, Mengyuan Liu, Baochang Zhang, Jungong Han, Junjun Jiang, and Hong Liu, "3d action recognition using multi-temporal depth motion maps and fisher vector," in *IJCAI*, 2016, pp. 3331–3337.

[5] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz, "Improving human action recognition using fusion of depth camera and inertial sensors," *IEEE THMS*, vol. 45, no. 1, pp. 51–61, 2015.

[6] Heng Wang, Alexander Klser, Cordelia Schmid, and Cheng Lin Liu, "Dense trajectories and motion boundary descriptors for action recognition," *IJCV*, vol. 103, no. 1, pp. 60–79, 2013.

[7] Heng Wang, Oneata Dan, Jakob Verbeek, and Cordelia Schmid, "A robust and efficient video representation for action recognition," *IJCV*, vol. 36, no. 2, pp. 1–20, 2015.

[8] Limin Wang, Yu Qiao, and Xiaoou Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *CVPR*, 2015.

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein, "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[10] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *ECCV*, 2016.

[11] Graham W. Taylor, Rob Fergus, Yann Lecun, and Christoph Bregler, "Convolutional learning of spatio-temporal features," in *ECCV*, 2010.

[12] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, "3d convolutional neural networks for human action recognition," *IEEE TPAMI*, vol. 35, no. 1, pp. 221–31, 2013.

[13] Andrej Karpathy, George Toderici, Sanketh Shetty, and Thomas Leung, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[14] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015.

[15] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014.

[16] Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *ICCV*, 2015.

[17] Limin Wang, Yuanjun Xiong, Zhe Wang, and Yu Qiao, "Towards good practices for very deep two-stream convnets," *arXiv:1507.02159*, 2015.

[18] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici, "Beyond short snippets: Deep networks for video classification," in *CVPR*, 2015.

[19] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015.

[20] Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *ACM MM*, 2015.

[21] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov, "Unsupervised learning of video representations using lstms," in *ICML*, 2015.

[22] Xiaolong Wang, Ali Farhadi, and Abhinav Gupta, "Actions transformations," in *CVPR*, 2016.

[23] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman, "Convolutional two-stream network fusion for video action recognition," in *CVPR*, 2016.

[24] Basura Fernando, Peter Anderson, Marcus Hutter, and Stephen Gould, "Discriminative hierarchical rank pooling for activity recognition," in *CVPR*, 2016.

[25] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould, "Dynamic image networks for action recognition," in *CVPR*, 2016.

[26] Wangjiang Zhu, Jie Hu, Gang Sun, Xudong Cao, and Yu Qiao, "A key volume mining deep framework for action recognition," in *CVPR*, 2016.

[27] Feng Shi, Robert Laganiere, and Emil Petriu, "Gradient boundary histograms for action recognition," in *WACV*, 2015.

[28] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv:1212.0402*, 2012.

[29] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre, "Hmdb: a large video database for human motion recognition," in *ICCV*, 2011.

[30] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.

[31] Jia Deng, Wei Dong, Richard Socher, Li Jia Li, Kai Li, and Fei Fei Li, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[32] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM MM*, 2014.

[33] Christopher Zach, Thomas Pock, and Horst Bischof, "A duality based approach for realtime tv-l1 optical flow," in *JPRS*, 2007.