ICIC Express Letters
Part B: Applications
Volume **7**, Number **5**, May **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **993–998**

SPACE-TIME BINARIZED NORMED GRADIENT FOR ACTION RECOGNITION AND LOCALIZATION

Dongxue Wu and Weiwei Xing

School of Software Engineering
Bejing Jiaotong University
No. 3, Shangyuancun, Haidian District, Beijing 100044, P. R. China
{ 13121690; wwxing }@bjtu.edu.cn

Received November 2015; accepted February 2016

Abstract. *Recently, human action recognition in videos attracts increasing research interests in computer vision. As a result, visual features are becoming crucial for human action recognition in videos. In this paper, we propose Space-Time Binarized Normed Gradients (ST-BING) as a new feature for action recognition and localization. This feature comprises both the static information and the non-static information of an action performer. We use our ST-BING feature extracted from human action bounding boxes to learn an action recognition and localization model in a cascaded Support Vector Machine (SVM) framework. The binarized version of our feature is easy to extract with only a few atomic operations. Using our feature and just trained by a simple linear SVM, we attain better than state-of-the-art action recognition performance on a challenging dataset. At the same time, our method produces good action localization results.*
Keywords: Action recognition, Action localization, Computer vision, Space-time feature

1. **Introduction.** Human action recognition is a crucial research topic in computer vision due to its applications in surveillance, video search and retrieval. A lot of global and local representations have been proposed in action recognition task [1]. In recent years, many local space-time representations have been widely used, for example dense trajectories [2] and Space-Time Interest Points (STIP) [3]. Both of these two methods have a common issue that they only focus on non-static parts of video [4]. Therefore, they cannot attain the accurate action localization of the whole human body. We believe that non-static parts of videos are important for action recognition and the relevant static parts are not only important for action recognition but also vital for action localization. According to the above conclusions, Ma et al. [4] proposed a new representation for action recognition and localization and attained an impressive performance on some challenging datasets.

Different from their work, in this paper, we combine the non-static parts and the relevant static parts of videos as features in an efficient way. We propose a feature called Space-Time Binarized Normed Gradients (ST-BING), which is inspired by [5], for both action recognition and localization. The ST-BING feature comprises two parts: Space Binarized Normed Gradients and Time Binarized Gradients. The first part (Space Binarized Normed Gradients) comprises the static part of an action performer that may contain the whole human body in each frame. The second part (Time Binarized Gradients) comprises the non-static part of an action that contains the change of the action position in the current frame with the same position in the previous and back frame. Meanwhile, we use the general pattern of the two stages cascaded SVM to learn the action recognition and localization model with our ST-BING features.

Besides, in this paper, we evaluated the proposed ST-BING feature on a challenging benchmark dataset UCF-Sports [6], a representative dataset of sports in action localization. Using our feature and just trained by a simple linear SVM, we attain better or

comparable to state-of-the-art action recognition performance. At the same time, our method produces good action localization results.

The main contributions of this paper are: 1). A new ST-BING feature for both action recognition and localization; 2). We apply the accelerated version of NG feature into our ST-BING feature to speed up the feature extraction and testing process.

In this paper, we introduce some outstanding work in Section 2. In Section 3 we show our ST-BING feature and the training method. We present our experiments results in Section 4. Finally, Section 5 concludes our paper and make prospect.

2. **Related Work.** In the past decade, plenty of action recognition methods that use the bag of STIP [3] have been proved to achieve good performance on many challenging datasets, and some methods that use dense trajectories [2] also have performed well. However, the STIP or dense trajectories ignored the space-time relationships. Many methods have been proposed to enhance the space-time relationships for action recognition. Ma et al. [4] proposed the hierarchical space-time segment achieves good results. Our ST-BING feature also enhances the space-time relationships with reserving the action performer and motion information.

Action localization is not usually studied in action recognition, but it is quite common in the study of action detection [7,8]. [7] proposes a spatio-temporal localization method according to sliding window-based approaches in object detection. Based on [7], [8] learns the mapping between a video and a spatio-temporal action trajectory by using max-margin structured output regression. Nevertheless, relatively few works do both action recognition and localization. Use holistic representations of the human body in action recognition methods that have the foundation to localize the action performer, such as silhouettes of action performer [9], space-time shape models [10] and Motion History Images (MHI) [11]. These approaches may not be good enough to handle cluttered backgrounds and occlusions in realistic videos.

[12] proposes a representation which combined the bag-of-words style statistical representation and the figure-centric structural representation, which makes the representation can localize an action. [13] extracts a large set of features for action recognition in the videos and formulates them within a multiple instance learning framework, which sacrifices the efficiency for improving accuracy to some extent. [14] proposes a unified framework to solve the task of action recognition by combining human detection and pose estimation. Our ST-BING feature contains both action performer body and motion information that achieve better results in action localization, and at the same time the proposed method can perform efficiently by using our extracting and matching algorithm.

3. **Methodology.** As mentioned above, the non-static part of human body may contain the information of different actions and the static part may contain the pose of different actions. We also observe that people have well-defined closed boundary with objects. Based on this observation and the good result of [5] in object detection, we introduce a simple 128D Space-Time Normed Gradients (ST-NG) feature. For finding a human action in a frame, we use different sizes of window to scan over a frame. Each size of window is defined by fixed scales and aspect ratios.

Firstly, we defined the location l of a window as: $l = (i, x, y)$, where i and (x, y) are size and position of a window respectively. Secondly, we extract the ST-BING feature for each window and use a linear model $\mathbf{w} \in \mathbb{R}^{128}$ to calculate the score. Our linear SVM score is defined as: $s_l = \langle \mathbf{w}, \mathbf{stg_l} \rangle$, where s_l is the score of each window calculated by the linear SVM model $\mathbf{w}$, and $\mathbf{stg_l}$ is the ST-BING feature of each window at location l.

For each fixed window size i, we choose a small set of windows which are the most likely to contain a target action by using Non-Maximal Suppression (NMS). Taking into account the possibility that different size of window contains an action, we further define

the score of human action as: $act_l = a_i \cdot s_l + b_i$, where $a_i, b_i \in \mathbb{R}$ are learnt coefficient and a bias term by a linear SVM for different size of window and act_l is regarded as the possibility of each window belonging to the target action category. The score of human action act_l is used in our method to re-rank the set of proposals from location of a window l and make our results more robust.

3.1. **Space-Time Normed Gradients (ST-NG) and human action.** Inspired by [5], we find that both action performer bodies and objects share a same characteristic in the corresponding frame gradients even after resizing windows to a small fixed size (e.g., 8×8). The characteristic is that both of them are stand-alone things with well-defined closed boundaries and centers, even in a quite abstracted view. Since human is the performer of an action, we will calculate the current frame time gradients with its previous and back frame at the position where humans exist. This time gradients may contain the motion information of one person. Instead of sliding a frame with different size of window, we first resize the input frame to different quantized sizes, and for each three frame we calculate the Space-Time Normed Gradients. In these resized normed gradients maps, every value in an $8 \times 8 \times 2$ region is defined as a Space-Time Normed Gradients (ST-NG) feature in which dimensions are 128D for each fixed size.

Our new proposed ST-NG features are insensitive to change of translation, scale and aspect ratio. And the time gradients feature at the position where humans exist produces enough difference between different action categories.

3.2. **Learning action recognition model with ST-NG.** We follow the general idea of the two stages cascaded SVM [15] to learn an action recognition and localization for each frame of video.

Stage I. We use the ST-NG features of the ground truth action bounding boxes of one class as positive training samples. The ST-NG features of the ground truth action bounding boxes of other class are used as negative training samples. We also need to random sample some background windows and extract their ST-NG feature adding to the negative training samples. Using these training data, we can learn a single model $\mathbf{w}$ for s_l using linear SVM.

Stage II. Different size of window has different possibility to contain an action. Based on this reason, we learn a_i and b_i in act_l using a linear SVM [16]. After evaluating s_l at size i for training data, we get a set of proposals from different size. By using Non-Maximal Suppression (NMS), we select a small set of proposals at each size i as the training samples for Stage II and their Stage I linear SVM scores can be regarded as 1D feature.

Discussion. We show one of our Stage I linear model $\mathbf{w}$ in Figure 1. It is an $8 \times 8 \times 2$ matrix which we reshape to a two 8×8 matrix for display. The left 8×8 matrix has a

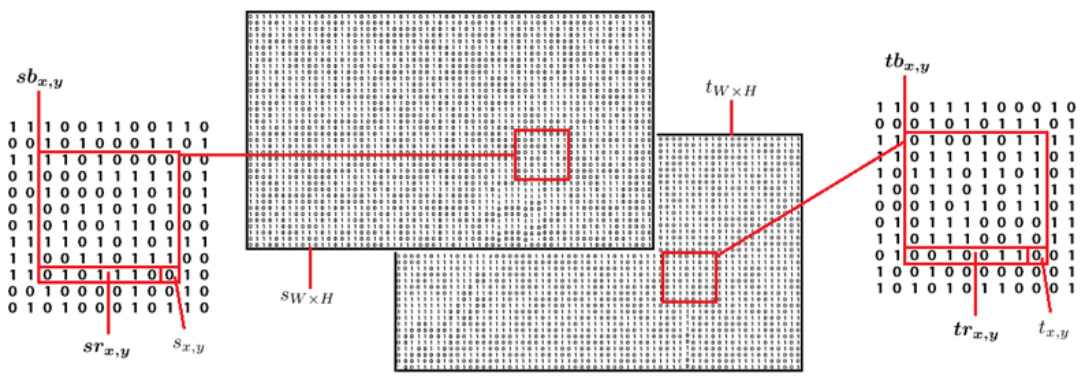

FIGURE 1. $\mathbf{sb_{x,y}}$ and $\mathbf{tb_{x,y}}$ constitute our ST-BING feature extracted from two binary normed gradient map $s_{W\times H}$ and $t_{W\times H}$ respectively.

large weights along the borders that separate a human body from its background. The right 8×8 one represents the action of one class.

3.3. **Space-Time Binarized Normed Gradients (ST-BING).** Inspired by [5] and recent brilliant works in model binary approximation [7,8], we propose a binarized version of ST-NG feature to speed up the feature extraction and testing process. Our learned linear model $\mathbf{w} \in \mathbb{R}^{128}$ can be divided into two 64D models, i.e., $\mathbf{w_1}$, $\mathbf{w_2}$. Both of them can be approximated with a set of basis vectors. We take $\mathbf{w_1}$ as an example, $\mathbf{w_1} \approx \sum_{j=1}^{N_{w_1}} \beta_{1j}\alpha_{1j}$, where N_{w_1} denotes the number of basis vectors, $\alpha_{1j} \in \{-1, 1\}^{64}$ indicates a basis vector, and $\beta_{1j} \in \mathbb{R}$ indicates the corresponding coefficient. We use a binary vector and its complement to represent each α_{1j} as: $\alpha_{1j} = \alpha_{1j}^{+} - \overline{\alpha_{1j}^{+}}$, where $\alpha_{1j}^{+} \in \{0, 1\}^{64}$.

A space-time binarized feature $\mathbf{b}$ can be divided into two parts, i.e., $\mathbf{b_1}$, $\mathbf{b_2}$. Our feature could be tested using fast BITWISE AND and BIT COUNT operations [5,7],

$$\langle \mathbf{w,b} \rangle = \sum_{m=1}^{2} \langle \mathbf{w_m}, \mathbf{b_m} \rangle \approx \sum_{m=1}^{2} \sum_{j=1}^{N_{w_m}} \beta_{mj} \left(2 \left\langle \boldsymbol{\alpha}_{\mathbf{mj}}^{+}, \mathbf{b_m} \right\rangle - |\mathbf{b_m}| \right). \tag{1}$$

The key challenge is to find out a way to binarize our ST-NG feature efficiently and make them easy to calculate. We divide our 128D ST-NG feature to two 64D features g_{l1}, g_{l2}. Each of both can be approximated by N_{g_m} half of Space-Time Binarized Normed Gradients (ST-BING) features.

Normally, getting an 8×8 BING feature needs a loop computing access to 64 positions. We can use a fast BING feature calculation algorithm, which uses atomic operations to avoid loop computing. Firstly, we divide the ST-BING feature into the Space BING and Time BING feature. After, we use a single INT64 and a BYTE variable to save a BING feature $\mathbf{b_{x,y}}$ and its last row $\mathbf{r_{x,y}}$, respectively. Then, we find a simple cumulative relation between adjacent BING features and their rows. We use the BITWISE SHIFT operation to shift $\mathbf{r_{x\text{-}1,y}}$ by one bit which does not belong to $\mathbf{r_{x,y}}$ and use BITWISE OR operation to insert the new bit $b_{x,y}$. In the same way, BITWISE SHIFT shifts $\mathbf{b_{x,y\text{-}1}}$ by 8 bits automatically through the bits which do not belong to $\mathbf{b_{x,y}}$, and insert $\mathbf{r_{x,y}}$ to update $\mathbf{b_{x,y}}$. We can efficiently test each ST-BING feature $\mathbf{b_{k,l}}$ of a window with the linear SVM model which we learned in Stage I using:

$$s_l = s_{l_1} + s_{l_2} \approx \sum_{j=1}^{N_{w_1}} \beta_{1_j} \sum_{k=1}^{N_{g_1}} C_{1_{j,k}} + \sum_{j=1}^{N_{w_2}} \beta_{2_j} \sum_{k=1}^{N_{g_2}} C_{2_{j,k}}, \tag{2}$$

where $C_{m_{j,k}} = 2^{8-k} \left(2 \left\langle \boldsymbol{\alpha}_{\mathbf{mj}}^{+}, \mathbf{b_{m_{k,l}}} \right\rangle - |\mathbf{b_{m_{k,l}}}| \right)$ also can be tested using fast BITWISE and POPCNT SSE operators.

We use the 1-D mask $[-1, 0, 1]$ to calculate three directions gradients g_x, g_y, g_t for each window in frames. g_x and g_y are gradients in horizontal and vertical directions. g_t is gradients in time axis direction. By default, we calculate gradients in RGB color space.

4. **Experiments.** We present our results on the UCF-Sports dataset [6] to demonstrate the effectiveness of our ST-BING feature in both action recognition and localization. This UCF-Sports dataset is a challenging dataset and is taken from real sports broadcasts. It consists of 150 videos of 10 different classes of actions. We split the dataset into training and testing data in the same way with [12]. We trained a separate classifier for each class of actions. For each frame of video, we use 36 quantized sliding window sizes $\{(W, H)\}$, where $W, H \in \{10, 20, 40, 80, 160, 320\}$. For each window, we will get a $2 \times 8 \times 8$ ST-BING feature and use our trained classifiers to discriminate which class of action it belongs or it does not contain an action. We chose the 10 highest score windows for each frame and use the most of action category label as the frame action category label.

Experimental Results. Firstly, we present our action recognition results on the UCF-Sports dataset at Table 1. We compare our method with four state-of-art methods [4,12,16,17]. Our method uses a quite simple feature compared with these methods and attains better than, or comparable to these method performance in action recognition. Table 1 shows our method performs better than [4,12,17], and achieves comparable classification performance with [16]. To assess the result of our method, we can intuitively see the importance of our ST-BING feature which contains both human body and motion information. In the experiment, we try different parameters of linear SVM and the performance of our feature is quite stable in action recognition.

Meanwhile, we present our action localization results on the UCF-Sports dataset at Table 2. It shows some localization result on part of video frames in the UCF-Sports dataset. We use the average IOU (intersection-over-union) to compute the localization score over tested frames. The method [12] can only cope with a subset of frames to present part of localization results and the method [4] can produce localization results on all frames, so we chose the highest average IOU of these two methods respectively. We can see that our method performs better than [12] at almost 10 classes, and also better than [4] on 7 out of 10 classes. The method [7] and [8] can only produce action localization results on running, diving and horse-riding of UCF-Sports dataset. Our method performs better than [7,8] in the diving class but in running and horse-riding classes we got lower performance. The size of sliding windows of our method is very important. We use as many as possible different sizes window to detect an action and in the future work we may use an SVM to choose the sliding window size instead of hand-crafted. Figure 2 shows more action localization results by using our ST-BING feature in the UCF-Sports video

Table 1. Pre-class classification accuracy on the UCF-Sports dataset

Method	Accuracy
Ma et al. [4]	81.7%
Wang and Sahbi [16]	85.2%
Raptis et al. [17]	79.4%
Lan et al. [12]	73.1%
Our method	85.4%

Table 2. Average IOU on the UCF-Sports dataset in action localization

	dive	golf	kick	lift	ride	run	skate	swing-b	swing-s	walk	Avg.
[7]	22.6	–	–	–	62.2	50.2	–	–	–	–	–
[8]	37.0	–	–	–	68.1	61.4	–	–	–	–	–
[12]	43.4	37.1	36.8	68.8	21.9	20.1	13.0	32.7	16.4	28.3	31.8
[4]	44.3	50.5	48.3	51.4	30.6	33.1	38.5	54.3	20.6	39.0	41.0
Ours	52.1	47.2	53.5	58.2	32.3	35.1	42.6	51.7	38.6	36.4	44.8

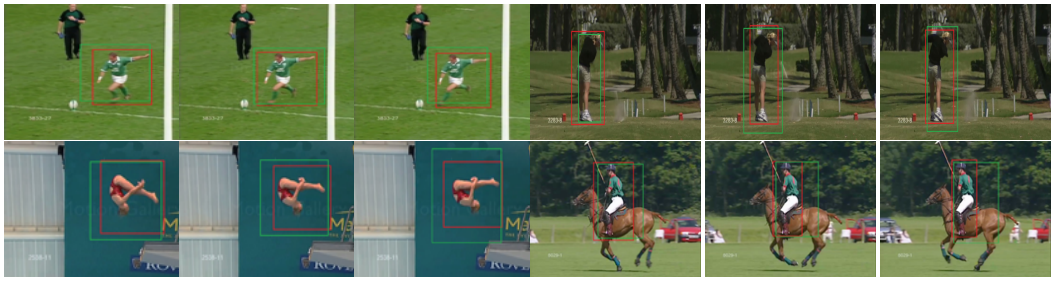

Figure 2. Action localization results in real videos

dataset. Two rectangle blocks in each photo represent the ground truth bounding box of action and our action localization results, respectively.

5. **Conclusion and Future Work.** In this paper, we propose a new ST-BING feature which can be easily extracted and effectively localize and recognize an action that happened in videos. Some previous methods such as STIPs and dense trajectories may be only able to recognize an action; our new ST-BING feature preserves both human body and motion information that make us also localize an action at the same time. Because of sliding window, our method can deal with not only one action in a frame. One direction is to make our method more robust to low frame rate video. A promising direction for future work is to apply more advanced machine learning techniques to achieve better performance on action recognition and localization.

Acknowledgement. This work is supported in part by National Natural Science Foundation of China (No. 61100143, 61370128), Program for New Century Excellent Talents in University (NCET-13-0659), Beijing Higher Education Young Elite Teacher Project (YETP0583).

REFERENCES

[1] P. Ronald, A survey on vision-based human action recognition, *Image and Vision Computing*, vol.28, no.6, pp.976-990, 2010.
[2] H. Wang, A. Kläser, C. Schmid and C. Liu, Action recognition by dense trajectories, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.3169-3176, 2011.
[3] I. Laptev, M. Marszalek, C. Schmid and B. Rozenfeld, Learning realistic human actions from movies, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.1-8, 2008.
[4] S. Ma, J. Zhang, N. Ikizler-Cinbis and S. Sclaroff, Action recognition and localization by hierarchical space-time segments, *Proc. of IEEE International Conference on Computer Vision*, pp.2744-2751, 2013.
[5] M. Cheng, Z. Zhang, W. Lin and P. Torr, BING: Binarized normed gradients for objectness estimation at 300fps, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.3286-3293, 2014.
[6] S. Khurram and A. R. Zamir, Action recognition in realistic sports videos, *Computer Vision in Sports*, Springer International Publishing, 2014.
[7] D. Tran and J. Yuan, Optimal spatio-temporal path discovery for video event detection, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.3321-3328, 2011.
[8] D. Tran and J. Yuan, Max-margin structured output regression for spatio-temporal action localization, *Proc. of Advances in Neural Information Processing Systems*, pp.350-358, 2012.
[9] Y. Wang, K. Huang and T. Tan, Human activity recognition based on R transform, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.1-8, 2007.
[10] M. Blank, L. Gorelick, E. Shechtman, M. Irani and R. Basri, Actions as space-time shapes, *Proc. of the 10th IEEE International Conference on Computer Vision*, pp.1395-1402, 2005.
[11] A. F. Bobick and J. W. Davis, The recognition of human movement using temporal templates, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.23, no.3, pp.257-267, 2001.
[12] T. Lan, Y. Wang and G. Mori, Discriminative figure-centric models for joint action localization and recognition, *Proc. of IEEE International Conference on Computer Vision*, pp.2003-2010, 2011.
[13] N. Ikizler-Cinbis and S. Sclaroff, Object, scene and actions: Combining multiple features for human action recognition, *Computer Vision*, pp.494-507, 2010.
[14] Y, Xie, H, Chang, Z. Li, L. Liang, X. Chen and D. Zhao, A unified framework for locating and recognizing human actions, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.25-32, 2011.
[15] Z. Zhang, J. Warrell and P. H. S. Torr, Proposal generation for object detection using cascaded ranking SVMs, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.1497-1504, 2011.
[16] L. Wang and H. Sahbi, Directed acyclic graph kernels for action recognition, *Proc. of IEEE International Conference on Computer Vision*, pp.3168-3175, 2013.
[17] M. Raptis, I. Kokkinos and S. Soatto, Discovering discriminative action parts from mid-level video representations, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.1242-1249, 2012.