Violence detection using Oriented VIolent Flows

Yuan Gao, Hong Liu, Xiaohu Sun, Can Wang, Yi Liu

PII: S0262-8856(16)30006-3
DOI: doi: 10.1016/j.imavis.2016.01.006
Reference: IMAVIS 3461

To appear in: *Image and Vision Computing*

Received date: 12 March 2015
Revised date: 13 January 2016
Accepted date: 14 January 2016

Please cite this article as: Yuan Gao, Hong Liu, Xiaohu Sun, Can Wang, Yi Liu, Violence detection using Oriented VIolent Flows, *Image and Vision Computing* (2016), doi: 10.1016/j.imavis.2016.01.006

Violence detection using Oriented VIolent Flows ☆

Yuan Gao[a], Hong Liu[a,*], Xiaohu Sun[a], Can Wang[a], Yi Liu[a,b]

[a] *Key Laboratory of Machine Perception*
Shenzhen Graduate School, Peking University, China
[b] *IMSL Shenzhen Key Lab, China*

Abstract

Nowadays, with so many surveillance cameras having been installed, the market demand for intelligent violence detection is continuously growing, while it is still a challenging topic in research area. Therefore, we attempt to make some improvements of existing violence detectors. The primary contributions of this paper are two-fold. Firstly, a novel feature extraction method named Oriented VIolent Flows (OViF), which takes full advantage of the motion magnitude change information in statistical motion orientations, is proposed for practical violence detection in videos. The comparison of OViF and baseline approaches on two public databases demonstrates the efficiency of the proposed method. Secondly, feature combination and multi-classifier combination strategies are adopted and excellent results are obtained. Experimental results show that using combined features with AdaBoost+Linear-SVM achieves improved performance over the state-of-the-art on the Violent-Flows benchmark.

Keywords: Violence Detection, Oriented VIolent Flows, AdaBoost, SVM

☆This work is supported by National Natural Science Foundation of China (NSFC, No.61340046), National High Technology Research and Development Program of China (863 Program, No.2006AA04Z247), Science and Technology Innovation Commission of Shenzhen Municipality (No. JCYJ20120614152234873, No.JCYJ20130331144716089), Specialized Research Fund for the Doctoral Program of Higher Education (No.20130001110011).

*Corresponding author

Email addresses: ygao@sz.pku.edu.cn (Yuan Gao), hongliu@pku.edu.cn (Hong Liu), xiaohusun@pku.edu.cn (Xiaohu Sun), canwang@pku.edu.cn (Can Wang), yi.liu@imsl.org.cn (Yi Liu)

1. Introduction

Violence detection is a particular problem within a greater problem of action recognition. In the last years, automation recognition of human actions in realistic videos has become increasingly important for applications such as video surveillance, human-computer interaction and content-based video retrieval [1, 2]. Recent proposed methods for action recognition can be roughly grouped into local, interest-point based, or global, frame-based methods.

In local methods, spatio-temporal feature points [3] are detected to represent human activity in a video [4]. An unsupervised method similar to the bag-of-words approach is proposed to learn the probability of distribution of these feature points [5]. Then a video can be represented with bag-of-feature techniques [6]. However, when there are only few interest points or too much motion, they may fail to provide enough meaningful information. On the other hand, global methods use global features such as optical flow to represent the state of motion in the frame at a particular instant of time [7, 8]. In [7], optical flow histograms, based on horizontal and vertical directions, were used as action descriptors to address the problem of human action recognition. Wang *et al.* in [8] employed optical flow to obtain densely tracking sampled points, and then they utilized these dense trajectories to calculate local descriptors for action recognition. Optical flow can describe coherent motion of moving objects, which is a good feature for motion detection and tracking. As a consequence, it has been widely used for object tracking and motion representation.

Relation to prior work: For violence detection, there existed some studies utilizing both vision and acoustic technologies [9, 10, 11]. However, in lots of surveillance systems, audio cues are usually unavailable. Therefore, in this paper we focus on violence detection in videos using pure vision methods. Datta *et al.* made an early attempt to address the violence detection problem based on background substraction [12]. Nevertheless, for violence happening in crowded scenarios, this approach may fail. In [13] and [14], the presence or absence of blood is an important cue for violence recognition. However, when the surveillance

cameras only output gray-scale videos, the performances of these approaches could be affected. More recently, Clarin *et al.* used local interest-point based approaches to detect fights on their own designed dataset [15]. Nievas *et al.* proposed a novel descriptor called ViF for real-time crowd violence detection [16]. Two databases, Hockey Fights and Violent-Flows, which were proposed in [15] and [16] separately, are benchmarks in our experiments for that both of them contain real-world, unconstrained, and violent or non-violent videos. Deniz *et al.* applied the Radon transform to get extreme acceleration patterns, which are the main feature of their method [17]. After using AdaBoost as the classifier, the violence recognition rate improved compared with BoW(STIP) and BoW(MoSIFT) methods on the Hockey Fights dataset. In 2015, Rota *et al.* used the improved trajectories [18] to get a feature codebook and the interpersonal space to detect violent interaction [19]. The disadvantage of their method is that it heavily relies on an accurate pedestrian detector or tracker and only analyses the circumstance of the interaction between two people, which could hardly be applied to analyse videos containing turbulent crowds in the Violent-Flows dataset.

In this paper, a new feature, OViF, is proposed for violence detection. The original motivation of designing OViF is to make full use of the orientation information of optical flow, which is omitted by ViF. Moreover, since feature combination and multi-classifier combination are common manners in the classification process, they are also adopted by us in experiments. During the period of multi-classier combination, a Linear-SVM classifier, which is trained on the features selected by AdaBoost, achieves noticeable improvement in violence recognition rate. This indicates the advantage of using AdaBoost as a feature selector. Finally, the combined features, ViF+OViF, obtain the state-of-the-art violence detection performance when using AdaBoost+Linear-SVM, which also shows the effectiveness of our proposed OViF features.

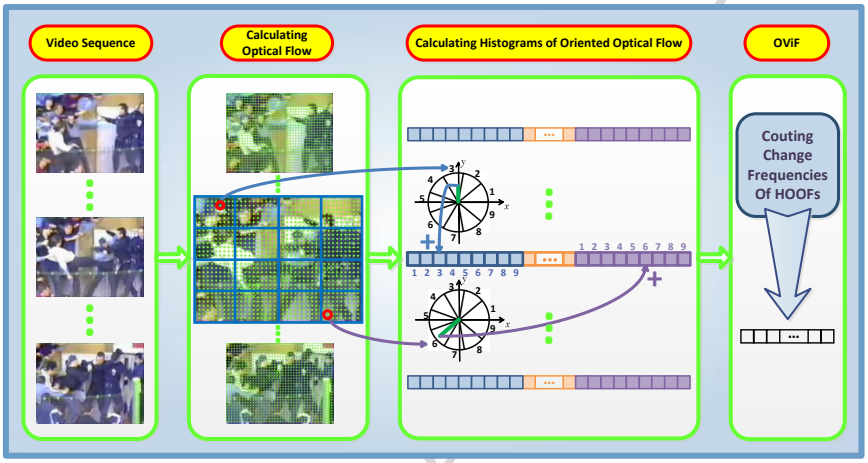

Figure 1: Flow chart of producing the OViF descriptor for a video sequence. The first step lists three example frames in an input video. The second one is the visualization of calculating optical flow for each frame. Step 3 describes how to calculate a Histogram of Oriented Optical Flow(HOOF) using these optical flow pictures. The last step obtains our proposed OViF by counting change frequencies of HOOFs, and the details are described in Eq. (4)(5).

2. Feature Extraction

In this section, two kinds of feature extraction methods, ViF and our proposed OViF, will be introduced orderly.

2.1. VIolent Flows (ViF)

The ViF descriptor is initially designed for crowd violence detection in [16]. In order to get a ViF vector for a video sequence, there are three steps. Firstly, through computing optical flow between pairs of consecutive frames, the magnitude of the flow vector, which corresponds to any pixel in a frame, can be computed. Then, by comparing these motion magnitudes between sequential frames, the magnitude-change maps are calculated, which helps obtaining a mean-magnitude map. Lastly, the obtained mean-magnitude map is divided into $M \times N$ non-overlapping regions, in each of which the frequencies of magnitude changes are collected and represented as a fixed-size histogram. And the

final ViF vector is the concatenation of these histograms.

2.2. Oriented VIolent Flows (OViF)

As introduced above, ViF is a feature descriptor which can describe the changes of observed motion magnitudes well. However, in some cases, it may loss some important information. For example, when the flow vectors of the same pixel in two sequential frames have the same magnitude but only differ in their directions, the effect of ViF seems to be restricted. This is because that ViF thinks that there is no difference between these two flow vectors but actually they differ a lot. Therefore, we propose a new feature representation method, OViF, which depicts the information involving both of motion magnitudes and motion orientations. The visualization of extracting OViF for a video sequence is illustrated in Fig. 1. And the details of this process is introduced as below:

Firstly, the optical flow should be calculated between pairs of sequential frames in the input video sequence. The flow vector of each pixel can be represented as:

$$|V_{i,j,t}| = \sqrt{(V^x_{i,j,t})^2 + (V^y_{i,j,t})^2} \tag{1}$$

$$\Phi_{i,j,t} = arctan(V^y_{i,j,t}/V^x_{i,j,t}) \tag{2}$$

Here, t means the t-th frame in a video sequence, and $(i,\ j)$ indicates the pixel location. Then, for each flow map which corresponds to a frame, we partition it into $M \times N$ non-overlapping blocks. After this, since 360° can be equally divided into B sectors and each sector corresponds to a bin of a histogram, the flow-vector magnitude $|V_{i,j,t}|$ is added into the bin where the flow-vector angle $\Phi_{i,j,t}$ locates. It is clear that, for each block, we get a histogram. These histograms are then concatenated into a single vector H, which is called Histogram of Oriented Optical Flow (HOOF) with X-dimensions:

$$X = M \times N \times B \tag{3}$$

This calculation step is exhibited in Step 3 of Fig. 1. Note that the HOOF here is special designed for violence detection task and is a little different from that in

[20]. There is no normalization here and the ways of counting orientations also differ. Subsequently, the HOOF vectors are used to obtain binary indicators:

$$b_{x,t} = \begin{cases} 1 & if\ |H_{x,t} - H_{x,t-1}| \geq \theta \\ 0 & otherwise \end{cases} \tag{4}$$

Here, x is the x-th dimension of the feature vector H, and θ is the average value of $|H_{x,t} - H_{x,t-1}|, x \subseteq [1, X]$. The above equation explicitly reflects the magnitude changes in different sectors. The mean magnitude-change vector is donated as:

$$\overline{b_x} = \frac{1}{T} \sum_t b_{x,t} \tag{5}$$

Finally, $\overline{b}$ is the final OViF vector for a sequence of frames, which counts the motion magnitude change frequencies in both direction sectors and spatial regions. Another reason for us to design OViF is that, based on observations of violent and nonviolent videos, we find that movements in nonviolent videos usually have the same direction with little deviation, while in violent videos movements are disordered with large deviation. Consequently, we encode the orientation information of motion and propose OViF.

3. Classifiers

Two types of traditional and popular machine learning algorithms, SVM and AdaBoost, are utilized in this work.

- As for the first classifier, a linear SVM [21] is selected in consideration of its simplicity, effectiveness and, last but not least, the speed.
- Regarding the second classifier, Gentle AdaBoost [22] is chosen for that it is one of the most practically efficient boosting algorithms.
- Apart of using these two algorithms individually, the combination of AdaBoost and SVM is also an effective way to improve classification performance. In particular, AdaBoost is only applied to select features and then a SVM classifier is trained on the selected features.

To know the specified implementation of these two algorithms, readers can refer to LIBLINEAR[1] and GML AdaBoost Matlab Toolbox[2].

4. Experiments and Discussions

4.1. Violence Databases

In order to demonstrate the effectiveness of the proposed OViF features, experiments are conducted on two public violence databases in both non-crowd and crowd scenarios.

Hockey Fight database is specially designed for evaluating violence detection systems [15]. It consists of 1000 videos (500 violence and 500 non-violence) of activities happening in the ice hockey rink. Each video exits two or very few people with 50 frames, and it is a non-crowd violence dataset.

Violent-Flows database is an evaluation benchmark for crowd violence detection [16]. There are 246 videos in this database (123 violence and 123 non-violence). All the videos are downloaded from the web with average 3.60 seconds and are under uncontrolled, in-the-wild conditions.

4.2. Experimental Settings

There are two baseline approaches, LTP [23] and ViF [16], for our OViF to compare against. The details are briefly introduced as follows:

LTP is a frame-based descriptor which achieves excellent performance on action recognition tasks. There are few parameters to set and they are the same as what in [23].

ViF is another important baseline approach and two parameters need to be set. The first one is the grid size, which equals 4×4. The second is the number of bins of a histogram, which equals 20.

OViF descriptor, as described in Section 2, also has two parameters. The grid

[1] http://www.csie.ntu.edu.tw/∼cjlin/liblinear/

[2] http://graphics.cs.msu.ru/ru/science/research/machinelearning/adaboost toolbox/

Table I: Classification results on the Hockey Fight database. Accuracy(± Standard Deviation) and AUC are reported (give in percentage).

Method	Classifier	Accuracy (± SD)	AUC
LTP [23, 16]	SVM	71.90 ± 0.49	-
ViF [16]	SVM	81.60 ± 0.22	88.01
	AdaBoost	73.70 ± 3.35	-
OViF	SVM	84.20 ± 3.33	90.32
	AdaBoost	78.30 ± 1.68	-
ViF+OViF	SVM	**86.30 ± 1.57**	**91.93**
	AdaBoost	82.30 ± 2.75	-
	AdaBoost+SVM	**87.50 ± 1.70**	**92.81**

of OViF owns the same size as ViF and B equals 9. Therefore, OViF is a feature vector with 144 ($4 \times 4 \times 9$) dimensions.

Five-fold cross validation is a common validation manner in previous work, which is also adopted here in experiments. In any one of the two databases, all the videos are divided into five heaps with the same ratio between violent and non-violent ones. At each time we select one distinct heap for testing and use the other four heaps for training. This procedure is then repeated for four times.

Besides, AdaBoost classification method has been applied in experiments, which is about selecting the most discriminative weak learners. Here, for any kind of feature(ViF or OViF), the number of weak learners is set to 100. And, for combined features(ViF+OViF), the number is set to 200. Moreover, the feature weights for ViF and OViF are 1:4 for the feature combination condition.

Lastly, with regard to the implementation of optical flow, we refer to an efficient implementation of coarse-to-fine optical flow in Piotr's Computer Vision Matlab Toolbox [24].

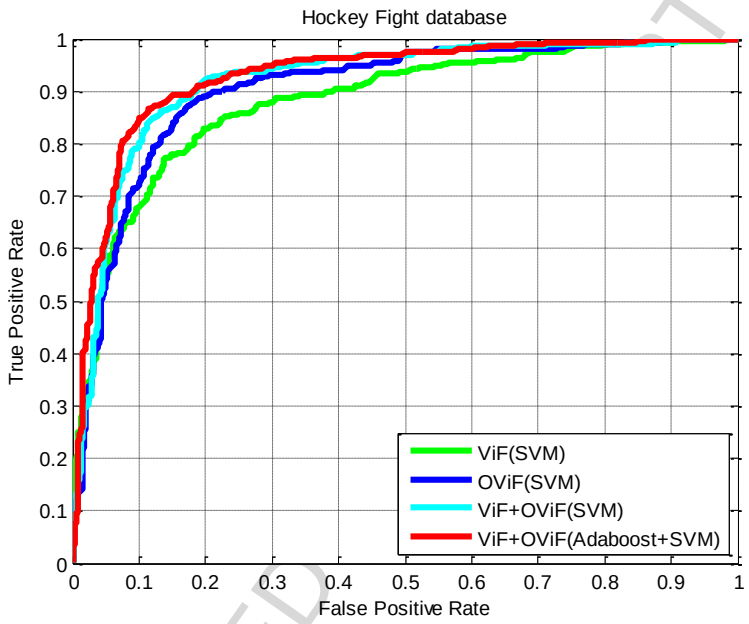

Figure 2: ROC curves of violence detection on 1000-video Hockey Fight database using OViF and baseline approaches.

4.3. Results and Analysis

Experiments are conducted on two databases separately. In addition to comparing our proposed OViF with baseline approaches, the combination of ViF and OViF is also implemented in consideration of the fact that they are both based on pre-calculating of optical flow, which occupies masses of computation time both in ViF and OViF. The calculation of ViF+OViF takes only a bit more time than that of using ViF or OViF individually. For two different databases, Hockey Fight and Violent-Flows, the ROC curves are illustrated in Fig. 2 and Fig. 3 respectively. Moreover, the detailed experimental results are represented in Table I and Table II, which are analyzed as follows:

There are three common characteristics existing in results on both two violence databases. 1) When using any type of feature, SVM performs better than

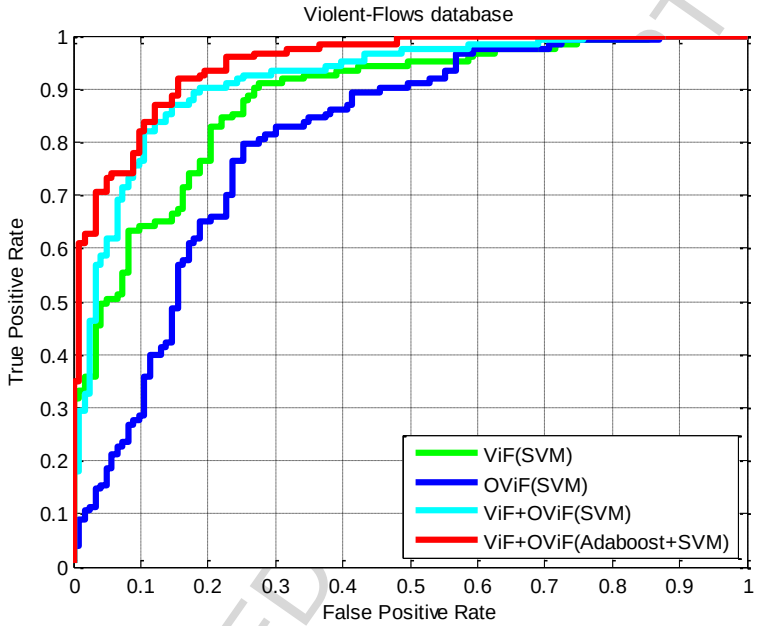

Figure 3: ROC curves of violence detection on 246-video Violent-Flows database using OViF and baseline approaches.

AdaBoost. This indicates that SVM classification approach is more suitable for the violent classification task than AdaBoost. 2) When using the same classification method, SVM or AdaBoost, ViF+OViF achieve better violence recognition rate than using ViF or OViF alone. It suggests that the feature combination manner is especially beneficial for improving the performance of the violence detector. 3) Uniting two traditional and famous classification methods contributes to the performance of the last violence detector, which implies that multi-classifier combination way is also very important for better performance. What's more, AdaBoost's ability of selecting discriminative features is powerful. After the common characteristics having been described, some differences are represented as bellow:

On the Hockey Fight database, the performance of OViF is better than that

Table II: Classification results on the Violent-Flows database. Accuracy(± Standard Deviation) and AUC are reported (give in percentage).

Method	Classifier	Accuracy (± SD)	AUC
LTP [23, 16]	SVM	71.53 ± 0.17	79.86
ViF [16]	SVM	81.20 ± 1.79	88.04
	AdaBoost	77.60 ± 3.29	-
OViF	SVM	76.80 ± 3.90	80.47
	AdaBoost	74.00 ± 4.90	-
ViF+OViF	SVM	**86.00 ± 1.41**	**91.82**
	AdaBoost	82.40 ± 3.58	-
	AdaBoost+SVM	**88.00 ± 2.45**	**94.84**

of ViF, which shows that OViF is a good choice for non-crowded violence detection applications. Nevertheless, on the Violent-Flows database, the performance of OViF is not very satisfying, which shows OViF may not be a decent selection for the violence detection of crowded scenes, but it overcomes the LTP baseline.

The final best violence detection rates are 87.50% and 88.00% on Hockey Fight and Violent-Flows separately using ViF+OViF with Adaboost+SVM.

5. Conclusion

This paper focuses on proposing a reliable and discriminative feature, OViF, for practical violence detection problem. The OViF features describe the changes of motion magnitudes based on the statistics of motion orientations. In comparison with baseline approaches, it is more appropriated for violence detection in non-crowded scenarios than that in crowded scenarios. In addition, feature combination and multi-classifier combination strategies are found to be beneficial for improving the performance of the violence detector. The combination of OViF and ViF using AdaBoost+Linear-SVM outperforms the state-of-the-art on the Violent-Flows datebase.

6. References

[1] R. Poppe, A survey on vision-based human action recognition, Image and vision computing (IVC) 28 (6) (2010) 976–990.

[2] Q. Sun, H. Liu, Learning spatio-temporal co-occurrence correlograms for efficient human action classification, in: International Conference on Image Processing (ICIP), 2013, pp. 3220–3224.

[3] I. Laptev, On space-time interest points, International Journal of Computer Vision (IJCV) 64 (2-3) (2005) 107–123.

[4] J. Niebles, F.-F. Li, A hierachical model of shape and appearance for human actioin classification, in: Computer Vision on Pattern Recognition (CVPR), 2007, pp. 1–8.

[5] J. Niebles, H. Wang, F.-F. Li, Unsupervised learning of human action categories using spatio-temporal words, International Journal of Computer Vision (IJCV) 79 (2008) 229–318.

[6] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos in the wild, in: Computer Vision on Pattern Recognition (CVPR), 2009, pp. 1996–2003.

[7] S. Danafar, N. Gheissari, Action recognition for surveillance applications using optic flow and svm, in: Asian Conference on Computer Vision (ACCV), 2007, pp. 457–466.

[8] H. Wang, A. Klaser, C. Schmid, C.-L. Liu, Action recognition by dense trajectories, in: Computer Vision and Pattern Recognition (CVPR), 2011, pp. 3169–3176.

[9] J. Nam, M. Alghoniemy, A. H. Tewfik, Audio-visual content-based violent scene characterization, in: International Conference on Image Processing (ICIP), Vol. 1, 1998, pp. 353–357.

[10] W. Zajdel, J. D. Krijnders, T. Andringa, D. M. Gavrila, CASSANDRA: audio-video sensor fusion for aggression detection, in: IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), 2007, pp. 200–205.

[11] J. Lin, W. Wang, Weakly-supervised violence detection in movies with audio and video based co-training, in: Advances in Multimedia Information Processing-PCM, 2009, pp. 930–935.

[12] A. Datta, M. Shah, N. Da Vitoria Lobo, Person-on-person violence detection in video data, in: International Conference on Pattern Recognition (ICPR), Vol. 1, 2002, pp. 433–438.

[13] L.-H. Chen, H.-W. Hsu, L.-Y. Wang, C.-W. Su, Violence detection in movies, in: International Conference on Computer Graphics, Imaging and Visualization (CGIV), 2011, pp. 119–124.

[14] C. Clarin, J. Dionisio, M. Echavez, P. Naval, Dove: Detection of movie violence using motion intensity analysis on skin and blood, PCSC 6 (2005) 150–156.

[15] E. B. Nievas, O. D. Suarez, G. B. García, R. Sukthankar, Violence detection in video using computer vision techniques, in: Computer Analysis of Images and Patterns (CAIP), 2011, pp. 332–339.

[16] T. Hassner, Y. Itcher, O. Kliper-Gross, Violent flows: Real-time detection of violent crowd behavior, in: Computer Vision on Pattern Recognition Workshops (CVPRW), 2012, pp. 1–6.

[17] O. Deniz, I. Serrano, G. Bueno, T. Kim, Fast violence detection in video, in: International Conference on Computer Vision Theory and Applications (VISAPP), 2014, pp. 478–485.

[18] H. Wang, C. Schmid, Action recognition with improved trajectories, in: International Conference on Computer Vision (ICCV), 2013, pp. 3551–3558.

[19] P. Rota, N. Conci, N. Sebe, J. M. Rehg, Real-life violent social interaction detection, in: International Conference onImage Processing (ICIP), 2015, pp. 3456–3460.

[20] R. Chaudhry, A. Ravichandran, G. Hager, R. Vidal, Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions, in: Computer Vision and Pattern Recognition (CVPR), 2009, pp. 1932–1939.

[21] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (2011) 1–27.

[22] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting, The Annals of Statistics 28 (2) (2000) 337–407.

[23] L. Yeffet, L. Wolf, Local trinary patterns for human action recognition, in: International Conference on Computer Vision (ICCV), 2009, pp. 492–497.

[24] P. Dollár, Piotr's Computer Vision Matlab Toolbox (PMT), `http://vision.ucsd.edu/ pdollar/toolbox/doc/index.html`.