The Open University of Israel
Department of Mathematics and Computer Science

Real-Time Detection of Violent Crowd Behavior

Thesis submitted as partial fulfillment of the requirements
towards an M.Sc. degree in Computer Science
The Open University of Israel
Computer Science Division

By
Yossi Itcher

Prepared under the supervision of Dr. Tal Hassner

January 2013

Contents

List of Figures

List of Tables

Abstract

Although surveillance video cameras are now widely employed to monitor much of the world around us, their effectiveness in crime prevention is questionable. Here, we focus on a particularly challenging surveillance task: monitoring crowded events for outbreaks of violence. Such scenes require a human surveyor to vigilantly monitor multiple video screens, presenting crowds of people in a constantly changing sea of activity, and to identify any sign of breaking violence early enough to alert help. With this task in mind, we propose the following contributions: (1) We describe a novel approach to real-time detection of breaking violence in crowded scenes. Our method considers statistics of how flow-vector magnitudes change over time. These statistics, collected for short frame sequences, are represented using the VIolent Flows (ViF) descriptor. ViF descriptors are then efficiently classified as either violent or non-violent using linear SVM. (2) We present a unique data set of real-world surveillance videos, along with standard benchmarks designed to test both violent/non-violent classification, as well as real-time detection accuracy. Finally, (3) we provide empirical tests, comparing our method to state-of-the-art techniques, and demonstrating its effectiveness.

Acknowledgements

I wish to thank my thesis supervisor, Dr. Tal Hassner, for his valuable guidance, ideas and helpful remarks throught the thesis. His assistance, attention to detail, hard work and great ideas enriched my knowledge and made this thesis possible.

1 Introduction

There is no question that video surveillance equipment can be easily and cheaply deployed to monitor practically any environment. The effectiveness of doing so, however, is indeed questioned [2]. Surveillance systems are often ineffective due to the insufficient number of trained human supervisors watching the footage produced by these systems (see, e.g., [13]) and the natural limits of human attention capabilities [17]. This is understandable, when considering the huge numbers of cameras that require supervision, the monotonic nature of the footage, and the alertness required to pick up on events and provide an immediate response. In fact, even the seemingly simple task of searching recorded videos, off-line, for events that are known to have happened, requires the aid of Computer Vision systems for video retrieval (e.g., [33]) and summarization [34].

Here, we focus on the task of detecting outbreaks of crowd violence, as it happens, from surveillance video cameras. Such videos typically do not have audio tracks, and, of course, subtitles and other contextual sources of information are non-existent. The footage is often far below motion picture quality, and so color cues are not reliable and neither are the details required for fine-scale action recognition. Some action recognition techniques are designed to analyze a single dominant action in the video. Here, however, videos present crowds, and we do not know a-priori who will participate in the violence. Finally, crowd scenes are especially challenging as they present constant, often monotonous, spatially unconstrained, human motion. This may not only reduce the effectiveness of a human observing the videos over long periods of time, but it can also flood a Computer Vision system with large quantities of motion information, making methods relying on interest points too time consuming. Figure 1 illustrates the type of scenarios we consider here by providing some examples from our database of both violent and non-violent crowd behavior.

Figure 1: Examples of violent (down-left) and non-violent (up-right) crowd behavior in "real-world" videos

In order to design a system capable of operating in real-time, we forgo high-level shape and motion analysis (e.g., [1]) and instead follow the example of methods for dynamic texture recognition, such as(e.g., [16]) ,in collecting statistics of densely sampled, low-level features. For the purpose of violence detection in crowded scenes, however, we show that accuracy can be achieved, without compromising processing speed, by considering how flow-vector magnitudes change through time. We collect this information, over short frame sequences, in a representation which we call the VIolent Flows (ViF) descriptor (Sec. 3.1). These ViF descriptors are then efficiently labeled as either violent or non-violent using a standard linear Support Vector Machine (SVM).

In order to test the accuracy of our method we require suitable data and benchmarks. Few video collections are available for testing violence detection performance, and none that we are aware of focus on the problem described here. We have therefore assembled our own collection of videos, presenting both violent and non-violent crowd behaviors. Our videos were all downloaded from the web and therefore represent unconstrained, "in-the-wild" conditions and scenes. We tested both our own method, as well as existing state-of-the-art techniques on violence classification and violence detection benchmarks designed using this collection. Our tests clearly demonstrate the wide performance margin, in favor of the method proposed here.

2 Previous work

Action recognition. Violence may be considered a particular type of action. Violence detection, is therefore a particular problem within the greater problem of action recognition. Action recognition techniques can roughly be classified as either local, interest-point based approaches, or global, frame-based methods. Methods employing local information begin by first detecting space-time interest points [10, 21, 31, 40]. Descriptive information is then extracted at each of these locations using one of several space-time descriptors (see for example: [14, 18, 20, 22]). A whole video can then be represented using, e.g., Bag-of-Feature techniques (as in [22, 26, 30]). These methods are often very resilient to camera motion and have been shown to provide excellent performance on a number of challenging benchmarks [19, 22, 26, 28], however, when videos contain too few space-time interest points (e.g., little motion) or too much motion (e.g., as in our case here), they may fail to provide meaningful representations.

The alternative of considering whole frames, or frame parts, often builds on dense flow estimation between successive frames [3, 9, 11, 42] or high-level appearance models [43]. Related to crowd videos are the methods of [1, 15] and more recently Rodriguez et al [36]. Both these methods are data-driven and require matching parts of the query video – frame segments in [15] and spatiol-temporal cubes in [36] – to exemplars in a pre-collected database. Searching the database for matching exemplars would be impractical for the applications considered here.

Violence detection. Often, "violence detection" refers to detecting violent scenes in motion pictures and TV broadcasts. In such cases, "violence" may include anything from explosions to more subtle actions. In such cases, audio may provide important additional information for detection [8, 23, 29]. Sometimes a significant change in the scene (a "surprising event") may be considered an act of violence. Boiman and Irani proposed an approach for detecting unexpected events in videos by using a data-driven approach [6]. It is not straightforward to apply their method for real-time processing. Hendel et al. [12], on the other hand, describe a more efficient, probabilistic technique. Their method, however, assumes that the scene can be characterized using multiple space-time tubes, each containing an object moving in the scene. This requirement is often impractical in videos of crowds.

Dynamic textures. Videos of crowd scenes may be described as produced by a stochastic process, stationary in both space and time. Such videos are often referred to as dynamic textures [16]. Although the videos we focus on here are not necessarily stationary – different parts of the frames may have dif-

ferent characteristic motion patterns – it is reasonable to consider analyzing them using dynamic texture recognition techniques [27]. Indeed, over the past decade, such methods have been successfully applied to varying scenes, from pure textures (e.g., running water, smoke, etc., [38]) to facial expression recognition [45]. Recently, Local Binary Patterns (LBP), originally proposed for face and texture recognition in 2D images [32] and extended for 3D videos, have proven both effective and efficient in recognizing motion patterns [16, 45]. Inspired by these methods, the Local Trinary Patterns (LTP) of [44] has demonstrated state-of-the-art performance on action recognition tasks.

Benchmarks for action recognition. Following the recent trends in image test sets, video benchmarks have recently shifted focus, presenting more and more videos obtained "in-the-wild", typically downloaded from online repositories such as YouTube. For a recent, comprehensive survey of such benchmarks, see [19]. Few data sets, however, provide surveillance footage (e.g., [4]) and none provide surveillance footage capturing violent crowd behavior. Although some test sets have been assembled for the purpose of violence detection, these typically focus on violence occurring between two (or very few) people [5] or contain high quality motion picture and TV footage (the e.g., the "slaps and kisses" data-set [37]). The videos assembled here, described in Sec. 4, present challenging, real-world scenes. We design both a straightforward, five-fold, cross validation test for violence classification accuracy, as well as tests for violence action detection.

3 Violence in crowded scenes

We make the following general assumptions on the surveillance footage and the problem at hand: (1) The viewpoint is typically far from the scene, and therefore captures many people appearing in low resolution. (2) Processing must be kept at real-time; frame processing should require less than 1/25 seconds per frame on a standard computer and a detection should be made within a few seconds of the outbreak of violence.

Given a video sequence $\mathcal{S}$ of frames $\{f_1, f_2, \ldots\}$ we consider two related but different tasks. The first is *violence classification*: The video $\mathcal{S}$ is assumed to be segmented temporally, containing T frames portraying either violent or non-violent crowd behavior. The goal is to classify $\mathcal{S}$ accordingly. The second is *violence detection*: Here, we assume an input stream of frames and the goal is to detect the change from non-violent to violent behavior, with the shortest delay from the time (frame) that the change occurred. Moreover, as mentioned above, this goal must be achieved with processing performed faster than frame-rate.

Existing work [39] has shown that under certain circumstances, less than ten video frames are required for reliable action classification. We consider such sub-second delays acceptable for a detection system and so reduce the second problem to the first by processing short frame sequences separately, classifying each one as either violent or non-violent; a detection is reported once a violent sub-sequence of frames is thus encountered. We next describe how each frame sequence is represented and classified.

3.1 The ViF representation

Given a sequence of frames, $\mathcal{S}$, we begin producing the VIolence Flows (ViF) descriptor by estimating the optical flow between each pair of consecutive frames. This provides for each pixel $p_{x,y,t}$, where t is the frame index, a flow vector $(u_{x,y,t}, v_{x,y,t})$, matching it to a pixel in the next frame $t+1$. Here, we consider only the magnitudes of these vectors: $m_{x,y,t} = \sqrt{(u_{x,y,t}^2 + v_{x,y,t}^2)}$. Doing so is in some sense a throwback to some early action recognition techniques which also relied on flow-vector magnitudes for processing actions [24]. There are some important differences, however, between those earlier approaches and the one proposed here.

Unlike previous methods, we do not consider the flow magnitudes themselves, but rather how they *change* over time. Our rational is that although flow vectors encode meaningful temporal information, their magnitudes are arbitrary quantities: they can depend on frame resolution, different motions in different spatio-temporal scene locations, and more. By comparing consecutive magnitudes we obtain

a meaningful measure of the significance of the observed motion magnitude in this frame compared to the previous frame. This is somewhat related to the self-similarity descriptor of [41] and its extension to action recognition using the LTP descriptors of [44]. Unlike them, however, we consider similarities of *flow-magnitudes in time*, rather than local appearances.

Specifically, for each pixel in each frame we obtain a binary indicator $b_{x,y,t}$, reflecting the significance of the change of magnitude between frames:

$$b_{x,y,t} = \begin{cases} 1 & \text{if} \quad |m_{x,y,t} - m_{x,y,t-1}| \geq \theta \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Where θ is a threshold adaptively set in each frame to the average value of $|m_{x,y,t} - m_{x,y,t-1}|$. Doing so provides us with a binary, magnitude-change, significance map B_t for each frame f_t. We next compute a mean magnitude-change map by simply averaging these binary values over all the frames $f_t \in \mathcal{S}$:

$$\bar{b}_{x,y} = \frac{1}{T} \sum_t b_{x,y,t}. \tag{2}$$

In its simplest form, the ViF descriptor is a vector of frequencies of quantized values $\bar{b}_{x,y}$. If the crowd motion patterns were indeed spatially stationary, this may suffice. In practice, however, we found that different spatial regions have different characteristic behaviors. The ViF descriptor is therefore produced by partitioning $\bar{b}$ into $M \times N$ non-overlapping cells and collecting magnitude change frequencies in each cell separately. The distribution of magnitude changes in each such cell is represented by a fixed-size histogram. These histograms are then concatenated into a single descriptor vector.

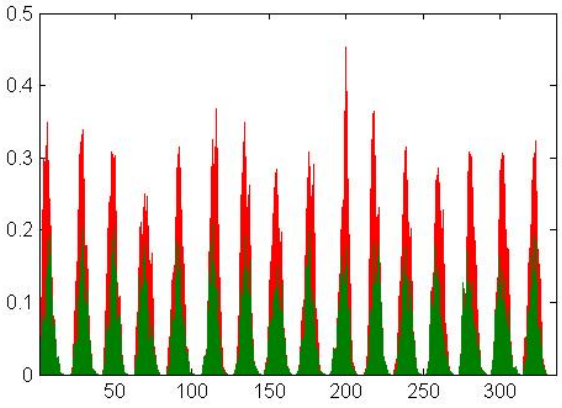

Figure 2: Two ViF descriptors produced for a violent scene (red) and a non-violent scene (green). 4×4 cells, each with 20-bin histograms concatenated into a 340-D ViF representation. See text for details.

What do the ViF descriptors capture? Figure 2 presents a comparison between the ViF profiles of violent and non-violent sequences. In both cases, ViF descriptors are L1 normalized. Clearly, the violent sequence produced a smaller variation in the number of times magnitudes changed. This reflects the rather arbitrary changes in flow-field magnitudes in non-violent scenes. Different parts of the frame move in different directions; changing direction and distance independently of other pixels. When violence breaks, however, pixels change their magnitudes in a more global manner; reflecting sharp motion changes in many pixels at once.

3.2 Classification with ViF descriptors

We use the ViF descriptors for classification in two distinct manners: (1) As global descriptors, extracted for a frame sequence as a whole or (2) as proxies used to produce a Bag-of-Features representation for each sequence.

Global descriptors. For a given sequence $\mathcal{S}$ we produce its ViF representation. Each such vector is then classified as representing an either violent or non-violent video. In practice, we found the ViF representation to capture meaningful, descriptive information, thus providing high classification scores even using simple linear support vector machines (SVM) [7] as the underlying classifier. As a consequence, real-time violence detection is achieved by considering short frame sequences, encoding each using its ViF descriptor and then immediately classifying it.

Bag-of-Features. Although ViF descriptors were designed with violent crowd behavior videos in mind, it is natural to consider how well they perform on "non-textured" actions and general action recognition tasks. In Section 5.2 we present such results using benchmarks other than the one described here. or large enough training sets, we take the frequency vectors produced for each cell, given a small number of frames, as local video descriptors. These are analogous to the descriptors produced by using existing STIP techniques. Here, however, we produce our own descriptors in a uniform, $M \times N$ grid. These descriptors are then quantized into a visual vocabulary using k-means. A whole video sequence is then represented using the frequencies of the ViF words it includes. The bags of words are then used according to the application at hand (Section 5.2).

4 The violent crowds data-set and benchmarks

4.1 Database assembly and details

Although data-sets which include videos for action recognition are by no means rare, we know of none suitable for testing violent crowd behavior. We therefore assembled our own database of videos for use in both violence classification and violence detection tasks[1]. To avoid introducing biases for particular scenes or behaviors, and at the same time provide a wide range of challenging real-world viewing conditions, our data is collected from YouTube. It therefore includes videos produced under uncontrolled, in-the-wild conditions, presenting a wide range of scene types, video qualities and surveillance scenarios. Table 1 presents some statistical details for our database, as well as provides examples for the search terms used to retrieve the videos.

The movies themselves are all de-interlaced and stored as AVI files. All the videos were compressed using the DivX codec (mpeg4), with the frames resized to 320×240 pixels.

[1] Our benchmark data set and protocols are all publicly available from the following URL: www.openu.ac.il/home/hassner/data/violentflows.

Table 1: **Violence/Non Violence Database Statistics**

General statistics:	
♯ of videos	246
♯ unique urls	214
♯ unique YouTube titles	218
Video statistics:	
Shortest video duration	1.04 sec.
Longest video duration	6.52 sec.
Average video duration	3.60 sec.
Example search terms:	
crowd violence	
balcony football violence	
fans violence	
football violence	
Hooligans violence	
soccer violence	

4.2 Benchmark protocols

We design two separate benchmarks on our video set.

Classification. The first benchmark is a five-fold cross-validation, classification test. We split the video set into five sets: half the videos in each set portray violent crowd behavior and half non-violent behavior. In some cases, different videos originated from the same YouTube clip or the same scene. In such cases, these videos are all included in the same set (the sets were mutually scene-exclusive).

The classification test protocol is a simple five-fold cross validation test. Five tests are performed; in each test, four of the sets are used for training (including SVM training and vocabulary generation, when required). Violence classification is then performed on the remaining set. Results are reported as both mean prediction accuracy (ACC) $\pm$ standard deviation (SD) as well as the area under the ROC curve (AUC).

Table 2: Classification results on our crowd violence database, mean over 5-folds cross validation. The average accuracy $\pm$ standard error as well as the AUC are given for a list of methods (see text for details).

Method	Accuracy $\pm$ SE	AUC
LTP [44]	71.53 % $\pm$ 0.17	0.79
HOG [22]	57.43 % $\pm$ 0.37	0.61
HOF [22]	58.53 % $\pm$ 0.32	0.57
HNF [22]	56.52 % $\pm$ 0.33	0.59
ViF	**81.30 % $\pm$ 0.21**	**0.85**

Detection. To evaluate the accuracy and reaction-time of a violence detection method, we consider videos which begin with non-violent behavior which turns to violence mid-way through the video. 21 such videos exist in our collection. We manually mark the frame in each video where this transition happens. The goal is to detect the violence as close to its manually specified outbreak as possible. Methods are required to process the videos, with frames provided sequentially. We require results on this test to present, in a graph, the percent of violence detections (percent of videos where violence was correctly detected) for increasing delays in time from violence outbreak. Different methods can then be compared by their accuracy vs. the time they require to detect the violence. Here, all training is performed on the videos which were *not* included in the detection test set.

5 Experiments

Our method was implemented in MATLAB, using code available on-line for computing optical flow [25], and linear SVM [7]. We have made little attempts to optimize the few parameters of our method, and so improved performance may be obtained by exploring other values. Here, we report the values used throughout our tests: We use a grid size of $M \times N = 4 \times 4$. For the violence/non-violence classification task we consider the whole video at once, i.e. Equation 2 averages over all the frames in the video to produce a single ViF descriptor. We use 20-bin histograms no matter the number of frames in the video. For real-time detection, Equation 2 averages frames in five-frame temporal windows, classifying each one separately and appropriately using six-bin histograms. We further found that it is enough to process one in every three frames for accurate temporal detection.

We compare our method to existing state-of-the-art techniques, representing two different approaches to action recognition. The first is the interest-point driven method of [22] as used in [19]. We use the implementation of [22] and test all three spatio-temporal descriptors it provides: HOG, HOF, and HNF. We use the videos included in the training set to produce a vocabulary of $6,000$ visual words using k-mean. Each video is then represented using a single frequency vector of size $6,000$ normalized to sum to one.

The second method we compare with is the LTP descriptor of [44], using a MATLAB implementation graciously provided by its authors. LTP, like ViF, is a frame-based descriptor. We therefore report its performance using the same pipeline used for our ViF descriptor.

5.1 Crowd violence database tests

We begin by presenting ViF performance on the database and benchmarks we have assembled for the purpose of violence classification and detection in crowds. Table 2 presents classification results on our five-fold cross validation test as described in Section 4.2. The ViF representation far outperforms the other methods tested. Unsurprisingly, the STIP representations, better suited for "structured videos", rather than the more textural videos in our data set, performed at almost chance. ROC curves of all tested methods are provided in Figure 3.

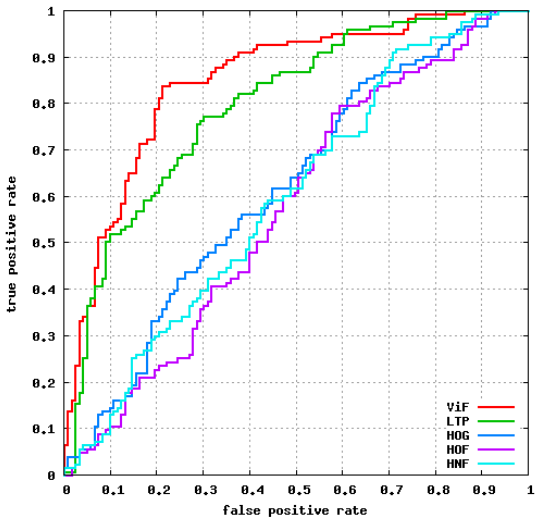

Figure 3: ROC curve for the various methods, averaged over 5-folds of our benchmark.

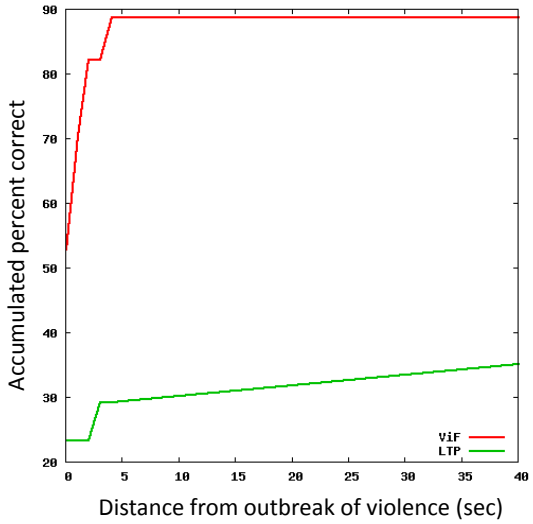

Figure 4: Real-Time detection results: ViF detects more violent scenes than [44] and does so sooner to the violence outbreak

To gain further insight on our results, Figure 5 presents the most confident classifications examples based on our ViF descriptor. The figure present the most confident correct classifications and the most confident incorrect classifications for both the violence and non-violence classes. Confident was measure as the distance from the SVM hyperplane.

Our real-time detection tests were performed on a 3Gb RAM, Intel core i7 computer running Windows Vista. These results are presented in Figure 4. We compare only ViF to LTP; STIP approaches performed too slowly for real-time processing, requiring, 0.28 seconds per-frame just for STIP feature extraction. Evidently, ViF detected far more violent scenes correctly, compared to LTP. It was further-

more far faster to detect the violence, typically in less than a second from the outbreak of violence. Table 3 summarizes these scores, providing also run-times for the two methods. Both operated at faster than frame-rate on our computer.

Table 3: Detection results on our benchmark (see text for details).

Method	LTP [44]	ViF
Success	35.29%	88.23%
Processing time per frame (ms)	10	30
Relative success by time to detection:		
1 Frame	23.53%	52.94%
1 Sec	23.53%	70.59%
2 Sec	23.53%	82.35%
3 Sec	29.41%	82.35%
4 Sec	29.41%	88.23%
10 Sec	35.29%	88.23%
Unable to detect at all	64.71%	11.77%

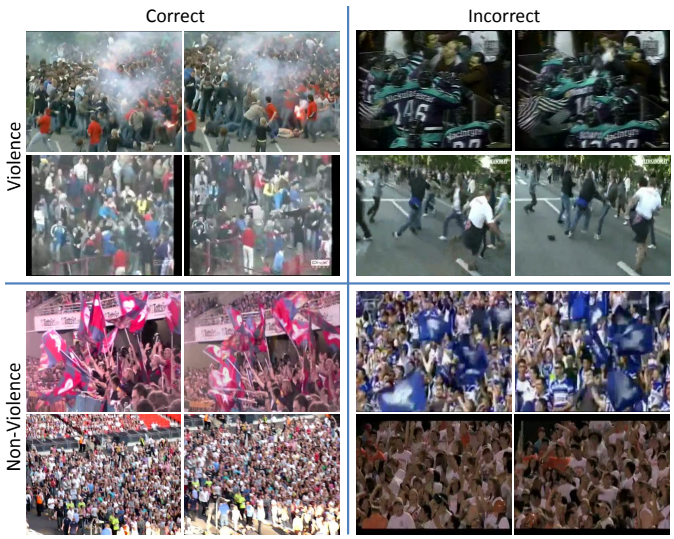

Figure 5: Most confident classification results based on ViF descriptor on our crowd violence database. The Violence/Non-Violence are the ground truth labels and the Correct/Incorrect labels indicate whether the method predicted correctly. For example the top-right quadrant displays frames of violent that were most confidently classified as non-violent.

ASCMN database tests. The Abnormal Surveillance Crowd Moving Noise (ASCMN) [35] was recently presented for testing a framework for dynamic saliency model evaluation. This data set contains 24 videos divided into 5 classes: Abnormal, Surveillance, Crowd, Moving and Noise. Some examples from this set are presented in Figure 6. We choose 4 movies that contain violence or some other abnormal and surprising event. We manually mark the frame in each video where this violence occurred. Our goal is to preform real time detection of violence as close to its manually specified outbreak as possible. The results show that 88.57% of the violent scenes were detected in less than a second from the outbreak of violence. Table 4 summarizes these scores.

Table 4: Detection results on ASCMN database

Success	94.29%
Relative success by time to detection:	
1 Sec	88.57%
3 Sec	94.29%
Unable to detect at all	5.71%

Figure 6: Examples of 5 video classes from the ASCMN database of [35].

5.2 Non-crowd behavior tests

Although the ViF representation was designed with the detection of crowd violence in mind, it is nevertheless natural to ask: how well ViF performs in action classification tasks involving one (or few) actors in "non-textures" video scenes? In this section we report the performance of the ViF descriptor on a benchmark for violence classification in non-crowded scenes and a more general benchmark for action similarity.

Hockey violence classification. The Hockey data set [5] was recently presented for testing methods designed to classify videos as violent or non-violent; not in crowd scenes, but instead, between two (or a few) participants. This data set contains 1000 video clips devided into five splits, each containing 100 violent and 100 non-violent sequences. Some examples from this set are presented in Figure 7. Methods are required to detect violence in a 5-folds cross validation test. Existing state-of-the-art results on this set were obtained using STIP descriptors, representing each video using a Bag-of-Features. In Table 5 we show our own result, the one obtained with LTP, and the state-of-the-art performances obtained by [5] with STIP [21, 22]. Clearly, ViF obtain performance comparable to using small STIP vocabularies. With larger vocabularies, STIP outperform ViF. This improved performance comes at a computational price, making such methods impractical for real-time processing.

Figure 7: Examples of two video sequences from the hockey violence database of [5].

Table 5: Classification results of various methods on the hockey violence set of [5], averaged over five-folds cross validation scheme. All the STIP results are as reported in [5].

Method	Accuracy ± SE
STIP(HOG) Vac50 [5]	87.8%
STIP(HOF) Vac50 [5]	83.5%
moSTIP Vac50 [5]	87.5%
STIP(HOG) Vac1000 [5]	91.7%
STIP(HOF) Vac1000 [5]	88.6%
moSTIP Vac1000 [5]	90.9%
LTP [44]	71.90 % ± 0.49
ViF	**82.90 % ± 0.14**

The ASLAN benchmark. To our knowledge, the Action Similarity Labeling Challenge (ASLAN) data set [19] is the most recent and comprehensive data set for testing action recognition methods. It includes thousands of videos portraying hundreds of human-performed actions. The goal of its accompanying benchmark is to decide if two videos present actors performing the same action, or not ("same" / "not-same" classification). Examples of the diversity of human actions in the ASLAN video collection are presented in Figure 8. Due to the un-textured nature of all the videos in the ASLAN set, and the high variability of the actions included in this set, ASLAN is highly *un*-suitable for the ViF descriptor. Nevertheless, the results reported in Table 6 demonstrate that ViF performance is comparable to other single-descriptor based methods reported in [19], while being far faster to extract.

Figure 8: Examples of the diversity of "real world" actions as presented in the ASLAN video collection of [19].

Table 6: Same/not-Same classification results of ViF, LTP and STIP the ASLAN video collection of [19], averaged over 10-folds cross validation scheme. All the STIP results are as reported in [19].

Method	Accuracy $\pm$ SE	AUC
HOG [19]	59.82 % $\pm$ 0.82	0.63
HOF [19]	56.68 % $\pm$ 0.56	0.58
HNF [19]	59.47 % $\pm$ 0.66	0.63
LTP [44]	55.45 % $\pm$ 0.60	0.57
ViF	56.57 % $\pm$ 0.25	0.58

6 Conclusions

Timely detection of violent outbreaks in crowds may mean the difference between life and death. Despite the significance of this task, it has received little attention in the past. Here, we make several important contributions towards the design of a system for detecting such events: We describe a novel means for efficient crowd violence detection. To test our system, as well as existing and future methods, we assemble a challenging data-set of related videos along with standard benchmarks. Finally, we demonstrate performance of both our own technique as well as existing ones on our own benchmarks and other video benchmarks.

Interestingly, the ViF representation presented here outperforms existing techniques by relying on magnitudes of the optical-flow fields alone. Although action recognition techniques have in the past been designed based on flow field magnitudes, more elaborate methods have since evolved, utilizing additional sources of information. Here, however, we turn to optical-flow magnitudes and show that when considered within a suitable frame of reference – by comparing their values from one frame to the next – coupled with spatial pooling, an accurate, computationally efficient representation emerges. We show that this representation is particularly potent when applied to the problem of detecting abnormal, specifically violent, crowd behavior.

References

[1] M. F. Abdelkader, W. Abd-Almageed, A. Srivastava, and R. Chellappa. Silhouette-based gesture and action recognition via modeling trajectories on riemannian shape manifolds. *Comput. Vis. Image Underst.*, 115(3):439–455, Mar. 2011. 8, 9

[2] R. Akers and C. Sellers. *Criminological theories: introduction, evaluation, and application*. Oxford University Press, 2008. 7

[3] S. Ali and M. Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(2):288–303, 2010. 9

[4] D. Baltieri, R. Vezzani, and R. Cucchiara. 3DPes: 3D people dataset for surveillance and forensics. In *Proc. ACM Workshop on Multimedia access to 3D Human Objects*, pages 59–64, Nov. 2011. http://imagelab.ing.unimore.it/visor/3dpes.asp. 10

[5] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar. Violence detection in video using computer vision techniques. In *Comput. Anal. Images and Patterns*, pages 332–339, 2011. 10, 21, 22

[6] O. Boiman and M. Irani. Detecting irregularities in images and in video. *Int. J. Comput. Vision*, 74(1):17–31, 2007. 9

[7] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, 2011. www.csie.ntu.edu.tw/∼cjlin/libsvm. 13, 17

[8] M. Cristani, M. Bicego, and V. Murino. Audio-visual event recognition in surveillance video sequences. *IEEE Trans. Multimedia*, 9(2):257–267, 2007. 9

[9] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. *European Conf. Comput. Vision*, pages 428–441, 2006. 9

[10] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance Performance Evaluation of Tracking and Surveillance*, pages 65–72, 2005. 9

[11] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proc. IEEE Int. Conf. Comput. Vision*, pages 726–733, 2003. 9

[12] A. Hendel, D. Weinshall, and S. Peleg. Identifying surprising events in videos using bayesian topic models. *Asian Conf. Comput. Vision*, pages 448–459, 2011. 9

[13] C. Hope. 1,000 CCTV cameras to solve just one crime, Met Police admits. The Telegraph, Aug. 2009. www.telegraph.co.uk/news/uknews/crime/6082530/1000-CCTV-cameras-to-solve-just-one-crime-Met-Police-admits.html. 7

[14] M. Kaâniche and F. Brémond. Recognizing gestures by learning local motion signatures of hog descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012. 9

[15] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *Proc. IEEE Int. Conf. Comput. Vision*, pages 1–8, 2007. 9

[16] V. Kellokumpu, G. Zhao, and M. Pietikainen. Human activity recognition using a dynamic texture based method. In *Proc. British Mach. Vision Conf.*, pages 1–10, 2008. 8, 9, 10

[17] H. Keval. *Effective, design, configuration, and use of digital CCTV*. PhD thesis, University College London, 2009. 7

[18] A. Klaeser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *Proc. British Mach. Vision Conf.*, pages 1–8, 2008. 9

[19] O. Kliper-Gross, T. Hassner, and L. Wolf. The action similarity labeling challenge. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012. 9, 10, 17, 22, 23

[20] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, pages 2046–2053, 2010. 9

[21] I. Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64(2):107–123, 2005. 9, 21

[22] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, pages 1–8, 2008. 9, 16, 17, 21

[23] J. Lin and W. Wang. Weakly-supervised violence detection in movies with audio and video based co-training. *Advances in Multimedia Information Processing-PCM 2009*, pages 930–935, 2009. 9

[24] J. Little and J. Boyd. Recognizing people by their gait: the shape of motion. *J. of Comp. Vision Research*, 1(2):1–32, 1998. 11

[25] C. Liu. *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis.* PhD thesis, Massachusetts Institute of Technology, May 2009. 17

[26] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos 'in the wild'. In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, pages 1996–2003, 2009. 9

[27] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, pages 1975–1981, 2010. 10

[28] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, pages 2929–2936, 2009. 9

[29] J. Nam, M. Alghoniemy, and A. Tewfik. Audio-visual content-based violent scene characterization. In *Int. Conf. Image Process.*, volume 1, pages 353–357, 1998. 9

[30] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *Int. J. Comput. Vision*, 79(3):299–318, 2008. 9

[31] J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, pages 1–8, 2007. 9

[32] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7), 2002. 10

[33] N. Petrovic, N. Jojic, and T. Huang. Adaptive video fast forward. *Multimedia Tools and Applications*, 26(3):327–344, 2005. 7

[34] Y. Pritch, S. Ratovitch, A. Hendel, and S. Peleg. Clustered synopsis of surveillance video. In *Advanced Video and Signal Based Surveillance*, pages 195–200, 2009. 7

[35] N. Riche, M. Mancas, D. Ćulibrk, V. Ćrnojevic, B. Gosselin, and T. Dutoit. Dynamic saliency models and human attention: a comparative study on videos. *Asian Conf. Comput. Vision*, November 2012. 20, 21

[36] M. Rodriguez, S. Ali, and T. Kanade. Tracking in unstructured crowded scenes. In *Proc. IEEE Int. Conf. Comput. Vision*, pages 1389–1396, 2009. 9

[37] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, pages 1–8, 2008. 10

[38] P. Saisan, G. Doretto, Y. Wu, and S. Soatto. Dynamic texture recognition. In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, volume 2, pages 58–63, 2001. 10

[39] K. Schindler and L. V. Gool. Action snippets: How many frames does human action recognition require? In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, pages 1–8, 2008. 11

[40] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *Int. Conf. Pattern Recognition*, volume 3, pages 32–36, 2004. 9

[41] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, pages 1–8, 2007. 12

[42] H. Sidenbladh and M. Black. Learning the statistics of people in images and video. *Int. J. Comput. Vision*, 54(1):183–209, 2003. 9

[43] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, pages 379–385, 1992. 9

[44] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *Proc. IEEE Int. Conf. Comput. Vision*, pages 492–497, 2009. 10, 12, 16, 17, 18, 19, 22, 23

[45] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. In *IEEE Trans. Pattern Anal. Mach. Intell.*, volume 29, pages 915–928, 2007. 10

תקציר

בעולם המודרני, מצלמות מעקב פזורות בכל מקום ומנטרות בקביעות את הסובב אותנו, אולם יעילותן במניעת פשעים תלויה בספק. בעבודת תזה זו נתמקד במשימת מעקב מאתגרת ביותר: ניטור התפרצויות של אלימות בסביבה מרובת קהל. משימה כזו דורשת משאבים רבים, בהם ניטור אנושי על מספר רב של מצלמות מעקב בסביבה דינאמית, משתנה ועמוסת קהל, תוך זיהוי התפרצות של אלימות בזמן מהיר מספיק על מנת לאפשר תגובה במהירות המרבית. במשימה כזו, מהירות התגובה יכולה להוות עניין של חיים ומוות. כהמשך למטרה זו, נתאר גישה חדשנית של גילוי אלימות בסביבה מרובת קהל בזמן אמת. השיטה שלנו מנתחת שינויים סטטיסטים של עוצמת וקטורי התנועה בתלות בזמן. הסטטיסטיקות הללו נאספות על פני סדרה קצרה של פריימים ומיוצגות ע"י ייצוג חדש למידע ויזואלי בסרטי וידאו שייקרא Violent Flows (VIF). באמצעות ייצוג זה, נסווג את הפעולות האלימות והלא אלימות תוך שימוש במסווג מסוג SVM ליניארי. בנוסף, נציג אוסף ייחודי של סרטים שנלקחו ממצלמות מעקב של אירועים אמיתיים שקרו ברחבי העולם, על מנת לספק אמת מידה לסיווג של פעולות אלימות מול פעולות לא אלימות. כמו כן נספק מבנה נתונים נוסף על מנת לבחון את רמת הדיוק של הסיווג בזמן אמת. לבסוף, נציג מבחנים ניסיוניים המשווים את השיטה שלנו מול שיטות מתחרות עדכניות המובילות היום בעולם ומציגים את יעילות השיטה שלנו על פניהם.

תוכן העניינים

האוניברסיטה הפתוחה
המחלקה למתמטיקה ולמדעי המחשב

זיהוי אלימות בזמן אמת
בסביבה מרובת קהל

עבודת תזה זו הוגשה כחלק מהדרישות לקבלת תואר
"מוסמך למדעים" M.Sc. במדעי המחשב
באוניברסיטה הפתוחה
החטיבה למדעי המחשב

על-ידי
יוסי איצ'ר

העבודה הוכנה בהדרכתו של ד"ר טל הסנר

ינואר 2013