Human Violence Recognition and Detection in Surveillance Videos

Piotr Bilinski and Francois Bremond
INRIA Sophia Antipolis, STARS team
2004 Route des Lucioles, BP93, 06902 Sophia Antipolis, France
{Piotr.Bilinski,Francois.Bremond}@inria.fr

Abstract

In this paper, we focus on the important topic of violence recognition and detection in surveillance videos. Our goal is to determine if a violence occurs in a video (recognition) and when it happens (detection). Firstly, we propose an extension of the Improved Fisher Vectors (IFV) for videos, which allows to represent a video using both local features and their spatio-temporal positions. Then, we study the popular sliding window approach for violence detection, and we re-formulate the Improved Fisher Vectors and use the summed area table data structure to speed up the approach. We present an extensive evaluation, comparison and analysis of the proposed improvements on 4 state-of-the-art datasets. We show that the proposed improvements make the violence recognition more accurate (as compared to the standard IFV, IFV with spatio-temporal grid, and other state-of-the-art methods) and make the violence detection significantly faster.

1. Introduction

Video surveillance cameras are part of our lives. They are used almost everywhere, *e.g.* at streets, subways, train and bus stations, airports, and sport stadiums. Today's increase in threats to security in cities and towns around the world makes the use of video cameras to monitor people necessary. The attacks on humans, fights, and vandalism are just some cases where detection, particularly violence detection, systems are needed.

In this paper, we focus on the important topic of violence recognition and detection in surveillance videos. Our goal is to determine if a violence occurs in a video (recognition) and when it happens (detection).

Over the last years, several violence recognition and detection techniques have been proposed. [6] have used motion trajectory information and orientation information of person's limbs for person-on-person fight detection. One of the main drawbacks of this approach is that it requires precise segmentation, which is very difficult to obtain in real world videos. Instead, [7, 9, 21, 23] have focused on local features and the bag-of-features approach; the main difference between these techniques lies in the type of features used. [23] have applied the STIP and SIFT, and [7] have used the STIP and MoSIFT features. [9] have proposed the VIolence Flows descriptor, encoding how flow-vector magnitudes change over time. [21] have proposed a video descriptor based on substantial derivative. Despite recent improvements in violence recognition and detection, effective solutions for real-world situations are still unavailable.

The Improved Fisher Vectors (IFV) [24] is a bag-of-features-like video encoding strategy which has shown to outperform the standard bag-of-features. It is a video (and image) descriptor obtained by pooling local features into a global representation. It describes local features by their deviation from the "universal" generative Gaussian Mixture Model. The IFV has been widely applied for recognition tasks in videos [1, 2, 11, 27, 28]. One of the main drawbacks of the IFV is that it simplifies the structure of a video assuming conditional independence across spatial and temporal domains; it computes global statistics of local features only, ignoring spatio-temporal positions of features.

Clearly, spatial information may contain useful information. A common way to use spatio-temporal information with IFV is to use either spatio-temporal grids [16] or multi-scale pyramids [17]; however, these methods are still limited in terms of a detailed description providing only a coarse representation. There are several other state-of-the-art methods [13, 14, 19, 25], but as they were proposed for images (for image categorization and object recognition), they cannot be directly applied for videos; moreover, [13, 14] achieve similar results as compared to the spatial grids/pyramids, and [19] is parameter sensitive and requires additional parameter learning.

As opposed to the existing violence recognition and detection methods (which focus mainly on new descriptors), we focus on a video representation model due to two reasons: to make it more accurate for violence recognition, and to make it faster for violence detection. Firstly, we propose an extension of the IFV for videos (Sec. 2.2),

978-1-5090-3811-4/16/$31.00 ©2016 IEEE

which allows to use spatio-temporal positions of features with the IFV. The proposed extension boosts the IFV and achieves better or similar accuracy (keeping the representation more compact) as compared to the spatio-temporal grids. Then, we study and evaluate the popular sliding window approach [12] for violence detection. We re-formulate the IFV and use the summed area table data structure to speed up the sliding window method (Sec. 2.3). Then, we present an extensive evaluation, comparison and analysis of the proposed improvements on 4 state-of-the-art datasets (Sec. 3 and Sec. 4). Finally, we conclude in Sec. 5.

Abnormal behavior detection: There are several methods for abnormal behavior and anomaly detection [4, 18, 20, 22]. However, abnormalities do not represent a compact and well defined concept. Abnormality detection is a different research topic, with different constraints and assumptions, and therefore we do not focus on these techniques.

2. Boosting the Improved Fisher Vectors (IFV)

2.1. State-of-The-Art: Improved Fisher Vectors

This section provides a brief description of the Improved Fisher Vectors, introduced in [24]. The mathematical notations and formulas provided here are in accordance with [24], and we refer to it for more details.

Let $\mathbf{X} = \{\mathbf{x}_t, t = 1 \dots T\}$ be a set of T local features extracted from a video, where each local feature is of dimension D, $\mathbf{x}_t \in \mathbb{R}^D$. Let $\boldsymbol{\lambda} = \{w_i, \mu_i, \mathbf{\Sigma}_i, i = 1 \dots K\}$ be parameters of a Gaussian Mixture Model (GMM): $u_{\boldsymbol{\lambda}}(\mathbf{x}) = \sum_{i=1}^{K} w_i u_i(\mathbf{x})$ fitting the distribution of local features, where $w_i \in \mathbb{R}$, $\mu_i \in \mathbb{R}^D$ and $\mathbf{\Sigma}_i \in \mathbb{R}^{D \times D}$ are respectively the mixture weight, mean vector and covariance matrix of the i-th Gaussian u_i. We assume that the covariance matrices are diagonal and we denote by σ_i^2 the variance vector, *i.e.* $\mathbf{\Sigma}_i = \mathrm{diag}(\sigma_i^2)$, $\sigma_i^2 \in \mathbb{R}^D$.

Moreover, let $\gamma_t(i)$ be the soft assignment of a descriptor $\mathbf{x}_t$ to a Gaussian i:

$$\gamma_t(i) = \frac{w_i u_i(\mathbf{x}_t)}{\sum_{j=1}^{K} w_j u_j(\mathbf{x}_t)}, \tag{1}$$

and let $\boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\mu,i}$ (resp. $\boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\sigma,i}$) be the gradient w.r.t. the mean μ_i (resp. standard deviation σ_i) of a Gaussian i:

$$\boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\mu,i} = \frac{1}{T\sqrt{w_i}} \sum_{t=1}^{T} \gamma_t(i) \left(\frac{\mathbf{x}_t - \mu_i}{\sigma_i} \right), \tag{2}$$

$$\boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\sigma,i} = \frac{1}{T\sqrt{2w_i}} \sum_{t=1}^{T} \gamma_t(i) \left[\frac{(\mathbf{x}_t - \mu_i)^2}{\sigma_i^2} - 1 \right], \tag{3}$$

where the division between vectors is as a term-by-term operation. Then, the gradient vector $\boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\boldsymbol{\lambda}}$ is the concatenation of all the K gradient vectors $\boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\mu,i} \in \mathbb{R}^D$ and all the K gradient vectors $\boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\sigma,i} \in \mathbb{R}^D$, $i = 1 \dots K$:

$$\boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\boldsymbol{\lambda}} = [\, \boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\mu,1}, \boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\sigma,1}, \dots, \boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\mu,K}, \boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\sigma,K} \,]'. \tag{4}$$

The IFV (Improved Fisher Vectors) representation $\boldsymbol{\Phi}^{\mathbf{X}}_{\boldsymbol{\lambda}}$, $\boldsymbol{\Phi}^{\mathbf{X}}_{\boldsymbol{\lambda}} \in \mathbb{R}^{2DK}$, is the gradient vector $\boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\boldsymbol{\lambda}}$ normalized by the power normalization and then the L2 norm:

$$\boldsymbol{\Phi}^{\mathbf{X}}_{\boldsymbol{\lambda}} = [\, \boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\mu,1}, \boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\sigma,1}, \dots, \boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\mu,K}, \boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\sigma,K} \,]'_{\sqrt{+l_2}}. \tag{5}$$

2.2. Boosting the IFV with spatio-temporal inf.

The Improved Fisher Vectors encoding simplifies the structure of a video assuming conditional independence across spatial and temporal domains (see Sec. 2.1). It computes global statistics of local features only, ignoring spatio-temporal positions of features. Thus, we propose an extension of the Improved Fisher Vectors which incorporates spatio-temporal positions of features into the video model.

Firstly, we represent positions of local features in a video normalized manner. In this paper, we focus on local trajectories only; however, the following representation can also be applied to spatio-temporal interest points [8, 15, 29] (with assumptions: $\mathbf{p}_t = (a_{t,1}, b_{t,1}, c_{t,1})$ is the spatio-temporal position of a point and $n_t = 1$).

Let $\mathbf{P} = \{\mathbf{p}_t, t = 1 \dots T\}$ be a set of T trajectories extracted from a video sequence and $\mathbf{p}_t = ((a_{t,1}, b_{t,1}, c_{t,1}), \dots, (a_{t,n_t}, b_{t,n_t}, c_{t,n_t}))$ is a sample trajectory, where a feature point detected at a spatial position $(a_{t,1}, b_{t,1})$ in a frame $c_{t,1}$ is tracked in $n_t \geq 1$ subsequent frames until a spatial position (a_{t,n_t}, b_{t,n_t}) in a frame c_{t,n_t}. We define the video normalized position $\hat{\mathbf{p}}_t$ of a center of a trajectory $\mathbf{p}_t$ as:

$$\hat{\mathbf{p}}_t = \left[\frac{1}{v_w n_t} \sum_{i=1}^{n_t} a_{t,i}, \frac{1}{v_h n_t} \sum_{i=1}^{n_t} b_{t,i}, \frac{1}{v_l n_t} \sum_{i=1}^{n_t} c_{t,i} \right]', \tag{6}$$

where v_w is the video width (with the units in pixels), v_h is the video height (in pixels), and v_l is the video length (number of frames). We normalize the position of a center of a trajectory, so that the video size does not significantly change the magnitude of the feature position vector.

Once positions of local features are represented in a video normalized manner, we also consider using the unity based normalization to reduce the influence of motionless regions at the boundaries of a video, so that the large motionless regions do not significantly change the magnitude of the feature position vector. Let $\hat{p}_{t,i}$ be the i-th dimension of the vector $\hat{\mathbf{p}}_t$ and $\min(\hat{\mathbf{p}}_{:,i})$ (resp. $\max(\hat{\mathbf{p}}_{:,i})$) be the minimum (resp. maximum) value of the i-th dimension among all the video normalized position vectors extracted from the training videos. When the condition $\forall i : \min(\hat{\mathbf{p}}_{:,i}) \neq \max(\hat{\mathbf{p}}_{:,i})$ is true, we can apply the unity based

normalization to calculate the vector $\tilde{\mathbf{p}}_t$. The i-th dimension of the vector $\tilde{\mathbf{p}}_t$ is:

$$\tilde{p}_{t,i} = \frac{\hat{p}_{t,i} - \min(\hat{\mathbf{p}}_{:,i})}{\max(\hat{\mathbf{p}}_{:,i}) - \min(\hat{\mathbf{p}}_{:,i})}. \quad (7)$$

Then, we incorporate the normalized positions of local features into the Improved Fisher Vectors model, so that videos are represented using both local descriptors and their spatio-temporal positions.

Let $\mathbf{Y} = \{\mathbf{y}_t = [\tilde{\mathbf{p}}_t, \mathbf{x}_t], t = 1 \dots T\}$ be a set of local features, where $\mathbf{x}_t \in \mathbb{R}^D$ is a local feature descriptor and $\tilde{\mathbf{p}}_t \in \mathbb{R}^E$ is its corresponding normalized position, typically $E = 3$, calculated as above. Let $\tilde{\boldsymbol{\lambda}} = \{\tilde{w}_i, \tilde{\mu}_i, \tilde{\boldsymbol{\Sigma}}_i, i = 1 \dots K\}$ be parameters of a GMM $u_{\tilde{\boldsymbol{\lambda}}}(\mathbf{y}) = \sum_{i=1}^{K} \tilde{w}_i u_i(\mathbf{y})$ fitting the distribution of local features, where $\tilde{w}_i \in \mathbb{R}$, $\tilde{\mu}_i \in \mathbb{R}^{D+E}$ and $\tilde{\boldsymbol{\Sigma}}_i \in \mathbb{R}^{(D+E)\times(D+E)}$ are respectively the mixture weight, mean vector and covariance matrix of the i-th Gaussian. As before, we assume that the covariance matrices are diagonal and we denote by $\tilde{\sigma}_i^2$ the variance vector, *i.e.* $\tilde{\boldsymbol{\Sigma}}_i = \mathrm{diag}(\tilde{\sigma}_i^2)$, $\tilde{\sigma}_i^2 \in \mathbb{R}^{D+E}$. We calculate $\boldsymbol{\mathcal{G}}^{\mathbf{Y}}_{\tilde{\mu},i}$ (Eq. 2) and $\boldsymbol{\mathcal{G}}^{\mathbf{Y}}_{\tilde{\sigma},i}$ (Eq. 3) for all K Gaussian components, and concatenate all the gradient vectors into a vector $\boldsymbol{\mathcal{G}}^{\mathbf{Y}}_{\tilde{\boldsymbol{\lambda}}}$. Finally, the new Improved Fisher Vectors representation is the gradient vector $\boldsymbol{\mathcal{G}}^{\mathbf{Y}}_{\tilde{\boldsymbol{\lambda}}}$ normalized by the power normalization and then the L2 norm:

$$\boldsymbol{\Phi}^{\mathbf{Y}}_{\tilde{\boldsymbol{\lambda}}} = \left[\, \boldsymbol{\mathcal{G}}^{\mathbf{Y}}_{\tilde{\mu},1}, \boldsymbol{\mathcal{G}}^{\mathbf{Y}}_{\tilde{\sigma},1}, \dots, \boldsymbol{\mathcal{G}}^{\mathbf{Y}}_{\tilde{\mu},K}, \boldsymbol{\mathcal{G}}^{\mathbf{Y}}_{\tilde{\sigma},K} \,\right]'_{\sqrt{+l_2}}. \quad (8)$$

2.3. Fast IFV-based Sliding Window

Our goal is to determine if a violence occurs in a video and when it happens; therefore, we search for a range of frames which contains violence. We base our approach on the temporal sliding window [12] which evaluates video sub-sequences at varying locations and scales.

Let v_l be a video length (in frames), $s > 0$ be the window step size (in frames), and $\mathbf{w} = \{is\}_{i=1\dots m}$ be temporal window sizes (scales) for the sliding window algorithm. Moreover, let $v = ns$ be an approximated video length (where: $ns \geq v_l > (n-1)s$ and $n \geq m \geq 1$). Visualization of a sample video and sample sliding windows is presented in Fig. 1. Note the IFV are calculated for features from the same temporal segments multiple times, *i.e.* $m(n-m+1)$ times for m segments (*e.g.* Fig. 1: 20 times for 8 segments). Therefore, to speed up the detection framework, we re-formulate the IFV and use the summed area table data structure, so that the IFV are calculated for features from the temporal segments only ones.

Let $\mathbf{X} = \{\mathbf{x}_t, t = 1 \dots T\}$ be a set of T local features extracted from a video. Let $\mathbf{X}' = \{\mathbf{X}_j, j = 1 \dots N\}$ be a partition of a set $\mathbf{X}$ into N subsets $\mathbf{X}_j = \{\mathbf{x}_{j,k}\}_{k=1}^{|\mathbf{X}_j|}$ such that: $|\mathbf{X}_j|$ is the cardinality of the set $\mathbf{X}_j$, $\mathbf{X} = \bigcup_{j=1}^{N} \mathbf{X}_j$,

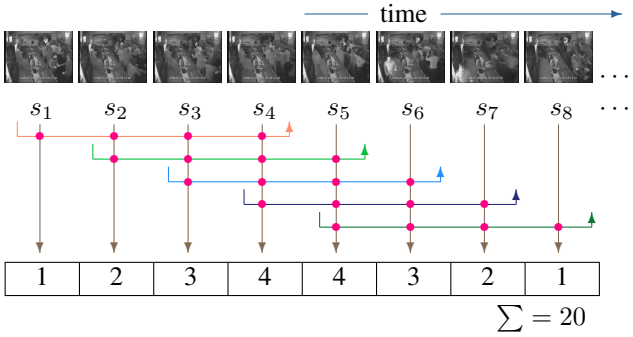

Figure 1. Temporal sliding window: a sample video is divided into $n \geq 8$ segments. We use $m = 4$ window scales. Note that the IFV are calculated for features from the same segments multiple times (20 times for 8 segments).

$\forall_{j\neq k,\ j,k=1}^{N} \mathbf{X}_j \cap \mathbf{X}_k = \emptyset$ and $\phi(j,k) \rightarrow t$ is the mapping function such that $\mathbf{x}_{j,k} = \mathbf{x}_t$.
We re-write Eq. (2):

$$\begin{aligned}
\boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\mu,i} &= \frac{1}{T\sqrt{w_i}} \sum_{t=1}^{T} \gamma_t(i) \left(\frac{\mathbf{x}_t - \mu_i}{\sigma_i} \right) \\
&= \frac{1}{T\sqrt{w_i}} \sum_{j=1}^{N} \sum_{k=1}^{|\mathbf{X}_j|} \gamma_{\phi(j,k)}(i) \left(\frac{\mathbf{x}_{\phi(j,k)} - \mu_i}{\sigma_i} \right) \\
&= \frac{1}{T} \sum_{j=1}^{N} \boldsymbol{\mathcal{G}}^{\mathbf{X}_j}_{\mu,i} |\mathbf{X}_j| = \frac{1}{T} \sum_{j=1}^{N} \boldsymbol{\mathcal{H}}^{\mathbf{X}_j}_{\mu,i}, \\
\boldsymbol{\mathcal{H}}^{\mathbf{X}_j}_{\mu,i} &= \boldsymbol{\mathcal{G}}^{\mathbf{X}_j}_{\mu,i} |\mathbf{X}_j| \\
&= \frac{1}{\sqrt{w_i}} \sum_{k=1}^{|\mathbf{X}_j|} \gamma_{\phi(j,k)}(i) \left(\frac{\mathbf{x}_{\phi(j,k)} - \mu_i}{\sigma_i} \right).
\end{aligned} \quad (9)$$

Similarly, we re-write Eq. (3):

$$\begin{aligned}
\boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\sigma,i} &= \frac{1}{T\sqrt{2w_i}} \sum_{t=1}^{T} \gamma_t(i) \left[\frac{(\mathbf{x}_t - \mu_i)^2}{\sigma_i^2} - 1 \right] \\
&= \frac{1}{T\sqrt{2w_i}} \sum_{j=1}^{N} \sum_{k=1}^{|\mathbf{X}_j|} \gamma_{\phi(j,k)}(i) \left[\frac{(\mathbf{x}_{\phi(j,k)} - \mu_i)^2}{\sigma_i^2} - 1 \right] \\
&= \frac{1}{T} \sum_{j=1}^{N} \boldsymbol{\mathcal{G}}^{\mathbf{X}_j}_{\sigma,i} |\mathbf{X}_j| = \frac{1}{T} \sum_{j=1}^{N} \boldsymbol{\mathcal{H}}^{\mathbf{X}_j}_{\sigma,i}, \\
\boldsymbol{\mathcal{H}}^{\mathbf{X}_j}_{\sigma,i} &= \boldsymbol{\mathcal{G}}^{\mathbf{X}_j}_{\sigma,i} |\mathbf{X}_j| \\
&= \frac{1}{\sqrt{2w_i}} \sum_{k=1}^{|\mathbf{X}_j|} \gamma_{\phi(j,k)}(i) \left[\frac{(\mathbf{x}_{\phi(j,k)} - \mu_i)^2}{\sigma_i^2} - 1 \right].
\end{aligned} \quad (10)$$

Then, let's define the gradient vector $\boldsymbol{\mathcal{H}}^{\mathbf{X}_j}_{\boldsymbol{\lambda}}$ as a concatenation of all the K gradient vectors $\boldsymbol{\mathcal{H}}^{\mathbf{X}_j}_{\mu,i}$ and all the K gradi-

ent vectors $\boldsymbol{\mathcal{H}}_{\sigma,i}^{\mathbf{X}_j}$, $i = 1 \dots K$:

$$\boldsymbol{\mathcal{H}}_{\boldsymbol{\lambda}}^{\mathbf{X}_j} = [\, \boldsymbol{\mathcal{H}}_{\mu,1}^{\mathbf{X}_j}, \boldsymbol{\mathcal{H}}_{\sigma,1}^{\mathbf{X}_j}, \dots, \boldsymbol{\mathcal{H}}_{\mu,K}^{\mathbf{X}_j}, \boldsymbol{\mathcal{H}}_{\sigma,K}^{\mathbf{X}_j} \,]'. \quad (11)$$

The Improved Fisher Vectors representation $\boldsymbol{\Phi}_{\tilde{\boldsymbol{\lambda}}}^{\tilde{\boldsymbol{X}}}$ of local features $\tilde{\boldsymbol{X}} = \bigcup_{j=M}^{N} \mathbf{X}_j$, where $1 < M \leq N$, can be calculated using:

$$\boldsymbol{\mathcal{G}}_{\boldsymbol{\lambda}}^{\bigcup_{j=M}^{N} \mathbf{X}_j} = \frac{\boldsymbol{\mathcal{H}}_{\boldsymbol{\lambda}}^{\bigcup_{j=1}^{N} \mathbf{X}_j} - \boldsymbol{\mathcal{H}}_{\boldsymbol{\lambda}}^{\bigcup_{j=1}^{M-1} \mathbf{X}_j}}{\sum_{j=1}^{N} |\mathbf{X}_j| - \sum_{j=1}^{M-1} |\mathbf{X}_j|}, \quad (12)$$

and applying the power normalization and then the L2 norm to the obtained gradient vector. The obtained representation is exactly the same as if we use Eq. (2)-(5). However, in contrast to the original IFV, the above equations can be directly used with data structures such as summed area table (Integral Images) and KDD-trees.

For the task of violence localization, we use the above formulation of the IFV (Eq. (9)-(12)), and directly apply the summed area table (Integral Images [26]). The 2 main advantages of this solution are: (1) it allows to speed up the calculations, as every feature is assigned to each Gaussian exactly once; *e.g.* we detected 25k features in a 84 frames long video. With $m = 4$ and $s = 5$, every feature was assigned to each Gaussian $4 - 10$ times; this is like 224k features were assigned to each Gaussian. In our algorithm, each feature is assigned to each Gaussian exactly once. This means nearly 9 times less calculations. (2) it allows to reduce the memory usage, especially when a video contains a lot of motion and dense features are extracted [27]; *e.g.* we extracted ∼130k features in a 35 seconds long video (∼3.7k features per second on average). With Improved Dense Trajectories [27] (each trajectory is represented using 426 floats), this means ∼1.6M floats to store per second (segment), which is 29 times more than the IFV representation with 128 Gaussians calculated for this segment.

3. Experimental Setup: Approach Overview

Firstly, we extract **local spatio-temporal features** in videos, and we use the Improved Dense Trajectories (IDT) [27] for that; we apply a dense sampling and track the extracted interest points using a dense optical flow field. Then, we extract local spatio-temporal video volumes around the detected trajectories, and we represent each trajectory using: Histogram of Oriented Gradients (HOG) capturing appearance information, and Trajectory Shape (TS), Histogram of Optical Flow (HOF) and Motion Boundary Histogram (with MBH-x and MBH-y components) descriptors capturing motion information. The extracted IDT features provide a good coverage of a video and ensure extraction of meaningful information. As the results, they have shown to achieve excellent results for various recognition tasks in videos and they have been widely used in literature [1, 27].

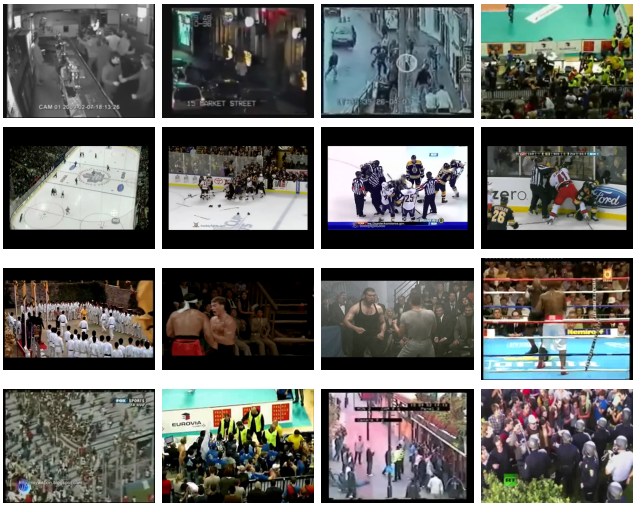

Figure 2. Sample video frames from the Violent-Flows (first row), Hockey Fight (second row), Movies (third row), and Violent-Flows 21 (fourth row) datasets.

To **represent a video**, we calculate a separate video representation for each descriptor independently (*i.e.* HOG, *etc.*) using the IFV / proposed Spatio-Temporal IFV (Sec. 2.2), and we concatenate the obtained representations using late fusion (*i.e.* per video: we concatenate the IFV-based video representation from HOG with video representation from HOF, *etc.*).

For **violence recognition**, we use linear Support Vector Machines (SVMs) [3] classifer, which has shown to achieve excellent results with high-dimensional data such as Fisher Vectors; as typically if the number of features is large, there is no need to map data to a higher dimensional space [10]. Moreover, linear SVMs have shown to be efficient both in training and prediction steps.

For **violence detection**, we use the Fast Sliding Window-based framework, explained in Sec. 2.3.

4. Experiments

4.1. Datasets

We use 4 benchmark datasets for evaluation and we follow the recommended evaluation protocols provided by the authors of the datasets. We use Violent-Flows dataset [9], Hockey Fight dataset [23] and Movies dataset [23] for violence recognition task. We use Violence-Flows 21 dataset [9] for violence detection task. Sample video frames from the datasets are presented in Fig. 2.

The **Violent-Flows** (Crowd Violence \ Non-violence) dataset [9] contains 246 videos with real-world footage of crowd violence. Videos are collected from YouTube and

contain a variety of scenes, *e.g*. streets, football stadiums, volleyball and ice hockey arenas, and schools. The dataset is divided into 5 folds and we follow the recommended 5-folds cross-validation to report the performance.

The **Hockey Fight** dataset [23] contains 1000 real-world videos: 500 violent scenes (between two and more participants) and 500 non-violent scenes. Videos are divided into 5 folds, where each fold contains 50% violent and 50% non-violent videos, and we follow the recommended 5-folds cross-validation to report the performance.

The **Movies** dataset [23] contains 200 video clips: 100 videos with a person-on-person fight (collected from action movies) and 100 videos with non-fight scenarios (collected from various action recognition datasets). This dataset contains a wider variety of scenes than the Hockey Fight dataset, and scenes are captured at different resolutions [23]. Videos are divided into 5 folds and we follow the recommended 5-folds cross-validation to report the performance. Although this dataset does not contain surveillance videos, it has been widely used in the past for violence recognition task.

The main differences between above datasets are: various scenarios and scenes, violence/fight and non-violence/non-fight classes variations, number of training and testing samples, pose and camera view point variations, motion blur, background clutter, occlusions, and illumination conditions.

The **Violent-Flows 21** dataset (Crowd Violence $\setminus$ Non-violence 21 Database) [9] contains 21 videos with real-world video footage of crowd violence. Videos are collected from YouTube, they are of spatial size 320×240 pixels, and they begin with non-violent behavior, which turns to violent mid-way through the video. The training is performed using 227 out of 246 videos from the Violent-Flows dataset; 19 videos are removed as they are included in the detection set. The original annotations are not available. Therefore, as proposed in the original paper [9], we manually mark the frame in each video where the transition happens from non-violent to violent behavior [1].

4.2. Implementation Details

We use the GMM with $K = 128$ and $K = 256$ to compute the IFV / Spatio-Temporal IFV, and we set the number of Gaussians using 5-folds cross-validation. To increase clustering precision, we initialize the GMM 10 times and we keep the codebook with the lowest error. To limit the complexity, we cluster a subset of $100,000$ randomly selected training features. To report recognition results, we use the Mean Class Accuracy (MCA) metric. For violence detection, we use six temporal windows of length $\{5i\}_{i=1}^{6}$ frames and the window stride equal to 1 frame. To report detection results, we use the Receiver Operating Characteristic (ROC) curve and the Area Under Curve (AUC) metrics.

[1] Differences can exist between our and [9] annotations.

Approach	Size	Violent-F.	Hockey F.	Movies
Baseline	1	93.5	93.2	97
Ours: STIFV	$\sim$1	**96.4**	**93.4**	**99**
IFV 1x1x2	2	94.0	93.3	98.0
IFV 1x2x1	2	94.3	93.6	97.5
IFV 2x1x1	2	94.3	93.2	97.5
IFV 1x1x3	3	93.5	93.1	98.5
IFV 1x3x1	3	94.3	93.2	97.0
IFV 3x1x1	3	93.5	93.2	97.5
IFV 2x2x2	8	93.5	93.4	97.0
IFV 2x2x3	12	93.1	93.4	97.0
IFV 2x2x1	4	93.9	**93.8**	97.5
IFV 2x1x2	4	93.5	92.9	98.0
IFV 1x2x2	4	93.9	93.5	97.5

Table 1. Evaluation results: the baseline (IFV with 1x1x1) approach, our IFV with spatio-temporal information (STIFV), and the IFV with various spatio-temporal grids on the Violent-Flows, Hockey Fight, and Movies datasets. Second column presents the size of the video representation relatively to the size of the video representation of the baseline approach.

4.3. Results: Violence Recognition

For violence recognition, we evaluate the standard IFV approach (baseline approach) and our IFV with spatio-temporal information (STIFV, Sec. 2.2). Moreover, we evaluate the IFV with 11 various spatio-temporal grids (1x1x2, 1x2x1, 2x1x1, 1x1x3, 1x3x1, 3x1x1, 2x2x2, 2x2x3, 2x2x1, 2x1x2, and 1x2x2). The evaluations are performed on 3 datasets: Violent-Flows, Hockey Fight and Movies datasets. The results are presented in Table 1. In all cases, our STIFV approach outperforms the IFV method, and achieves better or similar performance as compared to the IFV with spatio-temporal grid. Note that finding an appropriate size of the spatio-temporal grid is time consuming (there are 3 additional parameters to learn). Moreover, a spatio-temporal grid-based representation requires significantly more amount of memory (up to 12 times in our experiments, see Table 1).

Then, we compare our approach with the state-of-the-art. The comparison on the Violent-Flows, Hockey Fight, and Movies datasets is presented in Table 2. Note that our approach significantly outperforms remaining techniques, achieving even up to 11% better results (on the Violent-Flows dataset).

In summary, for violence recognition, the proposed improvement (IFV with spatio-temporal information) boosts the state-of-the-art IFV, and achieves better or similar ac-

Violent-Flows Dataset

Approach	Acc. (%)
HNF [16]	56.5
HOG [16]	57.4
HOF [16]	58.3
LTP [30]	71.5
Jerk [6]	74.2
Interaction Force [20]	74.5
ViF [9]	81.3
HOT [22]	82.3
$F^L \vert F^{Cv}$ [21]	85.4
Our Approach	**96.4**

Hockey Fight Dataset

Approach	Acc. (%)
LTP [30]	71.9
ViF [9]	82.9
STIP-HOF + HIK [23]	88.6
Extreme Accelerations [5]	90.1
MoSIFT + HIK [23]	90.9
BoW-MoSIFT [5]	91.2
STIP-HOG + HIK [23]	91.7
Our Approach	**93.7**

Movies Dataset

Approach	Acc. (%)
STIP-HOG + HIK [23]	49
STIP-HOF + HIK [23]	59
BoW-MoSIFT [5]	86.5
MoSIFT + HIK [23]	89.5
ViF [9]	91.3
Jerk [6]	95.0
Interaction Force [20]	95.5
$F^L \vert F^{Cv}$ [21]	96.9
Extreme Accelerations [5]	98.9
Our Approach	**99.5**

Table 2. Comparison with the state-of-the-art on the Violent-Flows (left table), Hockey Fight (middle), and Movies (right) datasets.

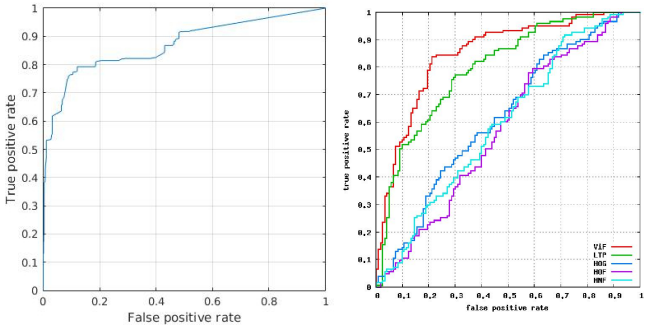

Figure 3. ROC curves: our approach (on the left) *vs*. the state-of-the-art (on the right) on the Violent-Flows 21 dataset.

Approach	LTP	HOG	HOF	HNF	VIF	**Ours**
AUC	79.9	61.8	57.6	59.9	85.0	**87.0**

Table 3. AUC metric on the Violent-Flows 21 dataset [9].

Process	Processing Time (fps)
Feature Extraction (IDT)	5.7
Sliding Window	9.28
Ours: Fast Sliding Window	**99.21**

Table 4. Average processing time on the Violent-Flows 21 dataset using a single Intel(R) Xeon(R) CPU E5-1630 v3 @ 3.70GHz.

curacy (keeping the representation more compact) as compared to the IFV with spatio-temporal grids. Moreover, our approach significantly outperforms the existing techniques on all three violence recognition datasets.

4.4. Results: Violence Detection

We evaluate our Fast Sliding Window-based approach on the Violence-Flows 21 dataset.

Firstly, we evaluate the accuracy of the sliding window / Fast Sliding Window approach (both techniques achieve the same results). The results and comparison with the state-of-the-art are presented in Figure 3 (using the ROC curves) and in Table 3 (using the AUC metric).

Then, we evaluate the speed of the Improved Dense Trajectories (IDT), and we compare the speed of the standard sliding window approach with the speed of our Fast Sliding Window technique (Sec. 2.3). The results are presented in Table 4. We observe that the proposed Fast Sliding Window technique is more than 10 times faster than the standard sliding window approach.

5. Conclusions

We have proposed an extension of the Improved Fisher Vectors (IFV) for violence recognition in videos, which allows to represent a video using both local features and their spatio-temporal positions. The proposed extension has shown to boost the IFV achieving better or similar accuracy (and keeping the representation more compact) as compared to the IFV with spatio-temporal grid. Moreover, our approach has shown to significantly outperform the existing techniques on three violence recognition datasets. Then, we have studied the popular sliding window approach for violence detection. We have re-formulated the IFV and have used the summed area table data structure to significantly speed up the violence detection framework. The evaluations have been performed on 4 state-of-the-art datasets.

Acknowledgements.

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n°[324359]. However, the views and opinions expressed herein do not necessarily reflect those of the financing institutions.

References

[1] P. Bilinski and F. Bremond. Video Covariance Matrix Logarithm for Human Action Recognition in Videos. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

[2] P. Bilinski, M. Koperski, S. Bak, and F. Bremond. Representing Visual Appearance by Video Brownian Covariance Descriptor for Human Action Recognition. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2014.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[4] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas. Abnormal Detection Using Interaction Energy Potentials. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[5] O. Daniz, I. Serrano, G. Bueno, and T.-K. Kim. Fast Violence Detection in Video. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2014.

[6] A. Datta, M. Shah, and N. D. V. Lobo. Person-on-Person Violence Detection in Video Data. In *International Conference on Pattern Recognition (ICPR)*, 2002.

[7] F. D. de Souza, G. C. Chavez, E. A. do Valle, and A. de A Araujo. Violence Detection in Video Using Spatio-Temporal Features. In *SIBGRAPI Conference on Graphics, Patterns and Images*, 2010.

[8] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior Recognition via Sparse Spatio-Temporal Features. In *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005.

[9] T. Hassner, Y. Itcher, and O. Kliper-Gross. Violent Flows: Real-Time Detection of Violent Crowd Behavior. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2012.

[10] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A Practical Guide to Support Vector Classification. Technical report, Department of Computer Science, National Taiwan University, 2003.

[11] V. Kantorov and I. Laptev. Efficient feature extraction, encoding and classification for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[12] A. Klaser, M. Marszalek, C. Schmid, and A. Zisserman. Human Focused Action Localization in Video. In *Trends and Topics in Computer Vision*, pages 219–233. Springer, 2010.

[13] J. Krapac, J. Verbeek, and F. Jurie. Modeling Spatial Layout with Fisher Vectors for Image Categorization. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[14] J. Krapac, J. Verbeek, and F. Jurie. Spatial Fisher Vectors for Image Categorization. Research Report RR-7680, INRIA, 2011.

[15] I. Laptev. On Space-Time Interest Points. *International Journal of Computer Vision (IJCV)*, 64(2-3):107–123, 2005.

[16] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[17] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[18] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly Detection in Crowded Scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[19] S. McCann and D. G. Lowe. Spatially Local Coding for Object Recognition. In *Asian Conference on Computer Vision (ACCV)*, 2012.

[20] R. Mehran, A. Oyama, and M. Shah. Abnormal Crowd Behavior Detection using Social Force Model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[21] S. Mohammadi, H. Kiani, A. Perina, and V. Murino. Violence detection in crowded scenes using substantial derivative. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2015.

[22] H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, and V. Murino. Analyzing Tracklets for the Detection of Abnormal Crowd Behavior. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2015.

[23] E. B. Nievas, O. D. Suarez, G. B. Garcia, and R. Sukthankar. Violence Detection in Video Using Computer Vision Techniques. In *International Conference on Computer Analysis of Images and Patterns (CAIP)*, 2011.

[24] F. Perronnin, J. Sanchez, and T. Mensink. Improving the Fisher Kernel for Large-Scale Image Classification. In *European Conference on Computer Vision (ECCV)*, 2010.

[25] J. Sanchez, F. Perronnin, and T. De Campos. Modeling the Spatial Layout of Images Beyond Spatial Pyramids. *Pattern Recognition Letters*, 33(16):2216–2223, 2012.

[26] O. Tuzel, F. Porikli, and P. Meer. Region Covariance: A Fast Descriptor for Detection And Classification. In *European Conference on Computer Vision (ECCV)*, 2006.

[27] H. Wang and C. Schmid. Action Recognition with Improved Trajectories. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[28] L. Wang, Y. Qiao, and X. Tang. Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[29] G. Willems, T. Tuytelaars, and L. Van Gool. An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. In *European Conference on Computer Vision (ECCV)*, 2008.

[30] L. Yeffet and L. Wolf. Local Trinary Patterns for human action recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.