www.doi.org/10.1109/TIFS.2017.2725820

Crowd Violence Detection Using Global Motion-Compensated Lagrangian Features and Scale-Sensitive Video-Level Representation

Tobias Senst, Volker Eiselein, Alexander Kuhn, and Thomas Sikora

***Abstract*—Lagrangian theory provides a rich set of tools for analyzing non-local, long-term motion information in computer vision applications. Based on this theory, we present a specialized Lagrangian technique for the automated detection of violent scenes in video footage. We present a novel feature using Lagrangian direction fields that is based on a spatio-temporal model and uses appearance, background motion compensation, and long-term motion information. To ensure appropriate spatial and temporal feature scales, we apply an extended bag-of-words procedure in a late-fusion manner as classification scheme on a per-video basis. We demonstrate that the temporal scale, captured by the Lagrangian integration time parameter, is crucial for violence detection and show how it correlates to the spatial scale of characteristic events in the scene. The proposed system is validated on multiple public benchmarks and non-public, real-world data from the London Metropolitan Police. Our experiments confirm that the inclusion of Lagrangian measures is a valuable cue for automated violence detection and increases the classification performance considerably compared to state-of-the-art methods.**

***Index Terms*—violence detection, lagrangian theory, lagrangian measures, crowd analysis, local feature, action recognition, long-term motion**

I. Introduction

THE rapid increase in the number of deployed video surveillance cameras fosters both an improvement of the analytical methods used with these cameras and the research of upcoming analysis techniques. A major research focus lies on the development of intelligent systems supporting the analysis of a huge amount of closed-circuit television (CCTV) footage in order to disburden the operator of the need to view all the data manually.

As an important example, the analysis of specific human actions, such as violence in crowds, has recently attracted a lot of attention in the computer vision community. However, while the task of automated detection of violence in movie databases [1] has inspired many works in this field, the area of video surveillance has not yet been studied sufficiently. Motion picture footage usually provides cues for multi-modal data analysis (e.g., fusion of audio, video, and contextual data [2]), leading to high-performing algorithms. In contrast, video surveillance data poses a number of difficulties: At the one hand, in most cases no audio information is available. On the other hand, video information quality is usually far below movie standards. Due to the need to operate continuously over months or years, a constant image quality in terms of contrast or color cues can often not be expected. Despite the rising number of high-resolution cameras, most existing CCTV cameras record in lower resolutions, such as VGA (640×480) or CIF (352×288). In addition, it is often unclear where events will occur in a scene in which an operator is interested. Consequently, CCTV cameras tend to show more overview with fewer details, cover longer periods of time, and they may not always be focused appropriately.

First approaches for violence detection (e.g., [3], [4]) focus on the microscopic level, meaning that a crowd is modeled by individuals and their specific behavior can be analyzed. Such methods usually have difficulties with CCTV footage of crowded scenes since individual persons need to be detected and tracked in the crowd robustly over a longer time. As a remedy to these problems, the macroscopic point of view has been proposed for the detection of violent crowd behavior [5]. Macroscopic methods do not consider individual pedestrian behaviors, but treat the crowd as an entity and perform their analysis based on the properties extracted from the whole scene. Especially motion information has been shown to carry important cues for detection of violent human actions in videos and crowds. To exploit motion information, different feature descriptors have been proposed which often benefit from the performance increase of recent optical flow-based motion estimation techniques.

From the survey of recent violence detection methods it can be concluded that motion information is a key property for detecting violence in videos. Most current methods (i.e. [6]–[10]) only consider two-frame (local or short-term) motion retrieved by optical flow between two consecutive frames. However, most characteristic motion signatures in a video are inhomogeneous over time and are potentially non-local in time. For instance, the process of kicking or punching has several phases composed by multiple individual long-term motion pattern sequences. It would thus be advantageous to encode these patterns into a representation that comprises the motion signature of multiple frames. In the field of action recognition, descriptors based on long-term trajectories, e.g. the improved dense trajectories [11], retrieved by computing dense optical flow fields or tracking feature points, constitute a significant step towards better performances. Wang *et al.* showed in [11] that densely sampled trajectories have the advantage of high discriminative power on a variety of datasets but also have a high computational complexity and high memory demand.

In our previous work, we have proposed a framework on Lagrangian methods for video analytics [12] as an generic

concept for integrating motion information over multiple temporal scales: The concept is based on the numerical integration of fieldlines that denote trajectories (or virtual particle traces) in the time-dependent optical flow field sequence. The computation of field lines is based on standard integration schemes (like Runge-Kutta) and does not require an explicit tracking or object identification step. These integral field lines and their properties gained significant attention in the field of video analytics: Similar frameworks have been used successfully in the field of video-based crowd analysis [13], [14] and crowd motion segmentation [15], [16]. With our previous work on person-oriented human action recognition [17] and people carrying baggage recognition [18], we showed the efficiency of the proposed framework on Lagrangian methods to describe long-term motion features for a variety of video-based surveillance applications.

In this paper we propose a violent video detection method, based on the Lagrangian methodology, with focus on a robust performance for challenging video surveillance data. This work is based on Lagrangian local features [19] and the bag-of-word model video-level representation. We extended our bag-of-word models in order to take the scale information of the local features into account. This will allow to generate distinct Lagrangian-visual vocabularies for motion patterns of different spatial sizes. In addition we apply a background motion compensation scheme to take account for dynamic camera motion within the scene. As a proof-of-concept and to substantiate the performance of the system, we tested our approach using real-world data from the London Metropolitan Police (London Riots 2011) and common violence detection benchmarks for comparison.

The remainder of this paper is organized as follows: In section II we review the current state-of-the-art and and relevant work for violent video classification. Section III briefly reviews the theory behind Lagrangian measures for video analytics, section IV presents the proposed Lagrangian Scale Invariant Feature Transform (LaSIFT) descriptor, based on a specialized direction measure. In section V, we follow the bag-of-words paradigm to encode a word frequency histogram for each video and classify each video into 'violent' or 'non-violent' using a support vector machine. Experimental results are presented in section VI, which covers suitable parameter setups, an benchmark on datasets for automated violence detection, and real-world datasets from London Metropolitan Police. Section VII concludes the paper and provides an outlook to possible future work.

II. Related Work

In general, two major concepts can be found for the classification of violent videos: *global* descriptors or *local* feature representations. Hassner *et al.* introduced the global Violent Flows (ViF) descriptor based on statistics of flow vector magnitude dynamics over time. They showed that such features classified with linear support vector machines are able to achieve real-time performance [9]. Real-time performance was one goal of Déniz *et al.* [7], who proposed a global descriptor that implicitly measures the acceleration of the global motion by comparing the power spectrum of consecutive video frames.

Local features have been first developed for the task of action recognition, where the common state-of-the-art methods are based on space-time interest point (STIP) [20] detections, histograms of oriented gradients (HoG) [21] or histogram of flow (HoF) [6] descriptors. De Souza *et al.* [22] presented an approach for violence detection based on local features. They compared STIP with the scale invariant feature transform (SIFT) [23] and showed that spatio-temporal features improve the detection performance compared to pure spatial SIFT. Hassner *et al.* showed in [9] that such feature representations fail on a newly proposed Crowd Violence dataset. They found that STIP is better suited for so-called "structured video" instead of CCTV footage which they consider "more textural" videos. Similar results are given by Nievas *et al.* in [8]. In their work, they compare the performance of the generalization capacity between STIP and Motion SIFT (MoSIFT) [24] features by using the Hockey Fight dataset [8] for training and the action movie database [8] for the testing phase and found that MoSIFT outperforms STIP. The MoSIFT feature has been proposed by Chen *et al.* [24] as an extension of the SIFT feature containing additional motion information. Further improvements were proposed by Xu *et al.* [10] who substituted the bag-of-words step with a sparse coding scheme to encode MoSIFT features for violent video detection. A different approach has been proposed by Mohammadi *et al.* with the Visual Information Processing Signature (VIPS) [25] feature. The VIPS is based on heuristic motion based rules that are related to acceleration, body compression and the aggressive drive in the video.

Apart from MoSIFT, the most related approaches to ours are the tracklet-based descriptor proposed by Mousavi *et al.* in [26], the Substantial Derivative proposed by Mohammadi *et al.* in [27] and the local mid-level visual description (MLV) proposed by Fradi *et al.* in [28]. The relation with [26] and [28] is that they provide a description for long-range motion patterns based on motion-trajectories. In addition, the tracklet-based descriptor [26] has shown to outperform dense trajectories for the Crowd Violence dataset. The relation with [27] is that it utilizes a concept from fluid dynamics and shows to be an appropriate feature for the classification of violence in videos.

III. Lagrangian Measures for Violent Video Detection

Lagrangian methods are commonly used to describe non-linear dynamic systems that can be described by a series of time-dependent fields. Commonly, properties of such systems are characterized by motion vector fields describing its dynamics, e.g. the physical motion of particles in fluid flows. Lagrangian methods quantify properties of each particle while moving within the flow field and reveal intrinsic motion patterns that are governed by the temporal evolution of the system. In our case, we adopt those concepts to a sequence of optical flow fields, which characterize the transport of information within an image sequence over time. Lagrangian fields can be derived by computing different types of field lines (or integral curves) within such a time-dependent flow

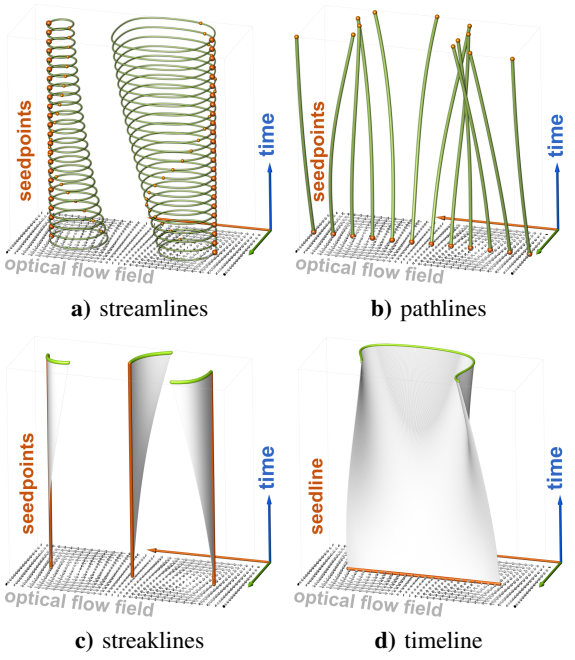


Fig. 1. Overview of different field line types (green color) that can be derived from a time-dependent optical flow field.

field. An overview about existing field line types are shown in figure 1. Streamlines (Fig. 1 a) are computed by integrating a single time step and consequently only describe a single state of the dynamic system. Streamlines represent the properties of that time step and closely relate to its topological features (e.g., points of zero-velocity or noise). Alternatively, streamlines can be used to analyze time-averaged flow fields (e.g. see Ali *et al.* [15]) at the cost of losing or displacing non-stationary motion features. Pathlines (Fig. 1 b) directly characterize the transport within the underlying flow field over time. They map a single seed point to a new position at each point in time, while this mapping for all pathlines at a specific time interval is denoted as the *flow map*. For optical flow applications, each trajectory point ideally corresponds to the same piece of information at each frame, while this might by violated during integration, e.g., due to occlusion or optical flow artifacts. Streaklines (Fig. 1 c) map a stationary point to its integrated positions and describe the evolution of pathlines that pass this point over time. They are commonly used to observe physical phenomena in numerical simulations, since they are easy to reproduce in real-world setting (e.g. by continuously injecting smoke into a flow field at a fixed position). Timelines (Fig. 1 d) capture the progression of a given seed structure over time and capture its deformation during integration in the flow field. To accurately compute streaklines and timelines, additional refinement procedures are required (i.e., inserting and integrating new seed points) to reproduce their geometry over longer integration intervals. In the literature, it has been shown that prominent Lagrangian features, can be extracted using pathlines [29], streaklines [30], and timelines [31]. In addition, time-dependent trajectories can be reformulated as streamlines in a higher-dimensional domain. The optimal choice of the field type typically depends on the characteristics of the flow field, computational overhead, and analysis goals at hand. For applications using optical flow fields, specifically video surveillance, the most prominent line types are pathlines [12], [18] and streaklines [16].

A. Formal Description of Lagrangian Fields

The Lagrangian measure used in this work is based on the notion of pathlines because in application to video analytics pathline are most related to object trajectories (e.g., as result of an object tracking). In contrast to streaklines, pathlines ideally remain on the objects during integration, while streakline seed points are fixed to an static location in the scene. Since we expect dynamic camera setups (relative motion in the frame of reference) we focus on the concept of pathlines, since they directly reflect the observed motion, use less samples, and do not require additional refinement schemes. However, our concept can be adopted for all above-mentioned line types. Note that the underlying optical flow methodology strongly influences of the accuracy and performance of those integration-based approaches [18].

Formally, pathlines can be computed as follows: Given a vector field $\mathbf{v}(\mathbf{x},t)$ defined on $D \in \mathbb{R}^n$ we can start a pathline that denotes a single trajectory. This can be formulated as an autonomous system:

$$\frac{d}{dt}\begin{bmatrix}\mathbf{x}\\ t\end{bmatrix} = \begin{bmatrix}\mathbf{v}(\mathbf{x}(t),t)\\ 1\end{bmatrix}, \begin{bmatrix}\mathbf{x}\\ t\end{bmatrix}(0) = \begin{bmatrix}\mathbf{x}_0\\ t_0\end{bmatrix}, \quad (1)$$

for a space-time point $[\mathbf{x}_0, t_0]$, with $\mathbf{x} \in \mathbb{R}^n$. This standard domain lifting technique allows to obtain two-dimensional pathlines in terms of three-dimensional streamlines in the space-time domain by interpreting time as additional dimension. In general, the concepts of the Lagrangian theory holds for any dimension $n \in \mathbb{N}$.

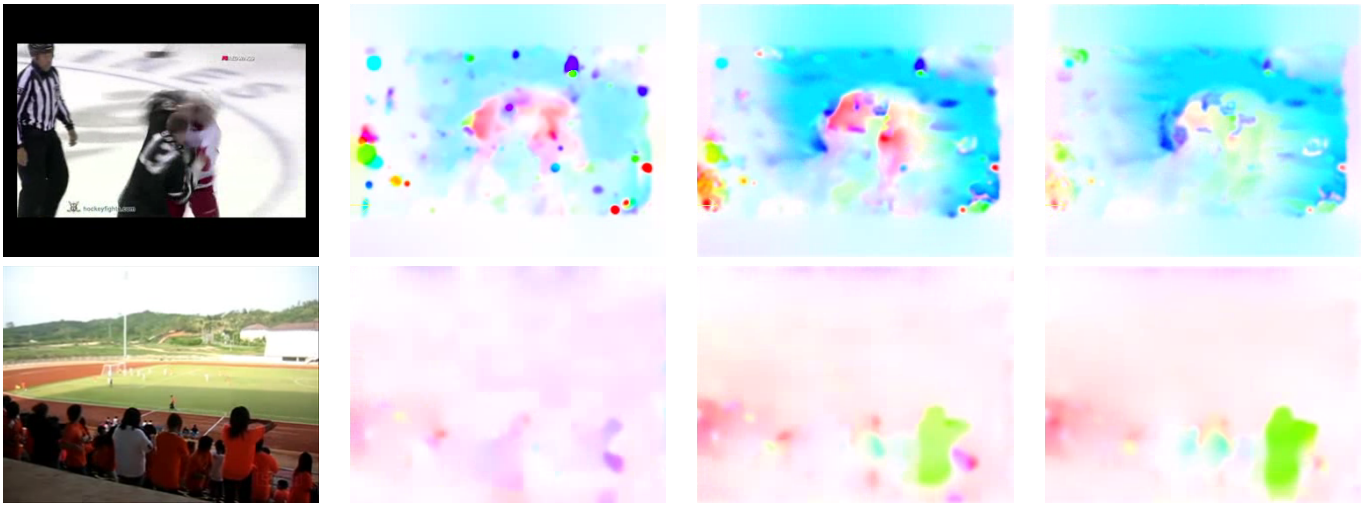
In this work we treat a series of optical flow fields as time-dependent vector fields and maintaining the notation for the time dependent vector field $\mathbf{v}(\mathbf{x},t) \in \mathbb{R}^2$ and the position $\mathbf{x} \in \mathbb{R}^2$. A trajectory $\mathbf{x}(t : t_0, \mathbf{x}_0)$ of a particle in that space depends on the initial position $\mathbf{x}_0$ and the initial time t_0 and can be estimated by the integration of Eq. 1 over t. Note, referring to Eq. 1 the pathline evolution is always forward in time.

One core aspect of Lagrangian methods is the computation of the flow map $\phi^\tau_{t_0}(\mathbf{x}) = \phi(\mathbf{x}, t_0, \tau)$, with $\phi^\tau_{t_0}(\mathbf{x}) \in \mathbb{R}^3$ which defines the mapping of all points at time t_0 to their corresponding positions after an integration time τ:

$$\phi^\tau_{t_0} : D \to D : \mathbf{x}_0 \mapsto \phi^\tau_{t_0}(\mathbf{x}_0) = \mathbf{x}(t : t_0, \mathbf{x}_0) \quad (2)$$

The flow map $\phi^\tau_{t_0}$ is constructed by integrating pathlines in a series of optical flow fields $\mathbf{v}(\mathbf{x},t)$ following Eq. 1 over τ. Since the optical flow fields are discrete in space and time, trilinear interpolation is applied to estimate $\mathbf{v}(\mathbf{x}(t),t)$.

Note that τ directly controls the size of the temporal interval (or temporal scale) and the complexity of the resulting mapping function. One of the most prominent Lagrangian fields used in this context is the finite-time Lyapunov exponent (FTLE) which is derived from the spatial gradient of the flow map and measures the amount of separation over a



a) frame **b)** optical flow field **c)** direction field ($\tau = 4$) **d)** direction field ($\tau = 8$)

Fig. 2. Optical flow and direction Lagrangian measures for different temporal scales applied to a fighting (top) and dancing crowd (bottom) sequence. Increasing τ allows to describe motion features on different temporal scales. While short-term events such as boxing are present in short-term integrated fields ($\tau = 4$), long-term events such as the dancing person are present in integration over larger scales ($\tau = 8$).

limited time interval (see overview in [29]). FTLE is often used to approximate features such as Lagrangian coherent structures (LCS) [32] and has been successfully applied in video analytics applications [16], [18]. As outlined before, in CCTV applications low resolutions and encoding artifacts are common, while it has been shown that FTLE fields are especially sensitive to such effects (c.f. [18]). Considering these specific challenges, we propose the use of an simpler, but more robust Lagrangian measure that directly encodes areas of coherent motion direction and magnitude over time (section III-C).

B. Properties of Lagrangian Fields

In comparison to local and time-independent feature descriptors, Lagrangian measures offer several distinct advantages: Lagrangian measures map information about the long-term dynamic motion behavior to a single frame. The resulting field compactly represents several aspects of the temporal evolution, while Lagrangian measures can be adopted towards specific analysis scenarios [12]. The parameters used for integration (spatial density of seed points, integration time τ) can be chosen independently of the underlying image sequence to capture dynamic features on varying spatial and temporal scales. Finally, the resulting fields spatially correspond to features in the underlying video sequence (e.g. boundary silhouettes [18]) and can be processed using standard image processing techniques. Any features extracted in this frame (e.g. regions of similar values) naturally translate into groups of pathline segments that are consistently defined for the corresponding subsequent frames. Recent research highlights strong relations between Lagrangian advection measures in the vicinity of LCS (e.g., see [33]–[35]). One specific property of gradient-based measures, such as FTLE, is Galilean invariance, which guarantees independence against global translations of the frame of reference. In optical flow applications, this implies that (if boundary effects and projective distortion are ignored) the FTLE field for a scene with camera motion is identical to the same scene observed by a static camera.

C. Lagrangian Fields for Optical Flow Analysis

Classic Lagrangian approaches (such as FTLE) use the flow map to derive the Cauchy-Green deformation tensor and its eigenvalues to quantify separation and stretching across neighboring pathlines [29]. In contrast to high-resolution simulation data, approximated optical flow fields contain a significantly higher amount of artificial discontinuities (due to noise, approximation errors, projective distortion, and occlusion). In our previous works [17], [18] we have shown that separation-based and fused measures are able to obtain robust results, but also introduce inaccuracies due to those artifacts. To reduce the influence of those effects, we focus our evaluation on metrics that do not require flow map gradient information, but emphasize characteristics that are specifically discriminative for (violent) motion detection and classification.

As a result of our previous experiments [12], we opted for a simple measure that: i) is less prone to motion estimation errors ii) provides direction and velocity information over a given time span and thus allows to distinguish objects by their motion. We found the *Lagrangian direction field* to fulfill these requirements and to offer a good tradeoff between discriminative efficiency and computational simplicity. The direction field $\Lambda_{X/Y}(\mathbf{x}, t_0) = [\Lambda_X(\mathbf{x}, t_0)\ \Lambda_Y(\mathbf{x}, t_0)]^T$, with $\Lambda_X, \Lambda_Y \in \mathbb{R}^1$ can be obtained by estimating the integral motion of the vectors along the path line as follows:

$$\Lambda_X(\mathbf{x}, t_0) = \frac{1}{\tau} \int u(\phi(\mathbf{x}, t_0, \tau))\partial\tau$$
$$\Lambda_Y(\mathbf{x}, t_0) = \frac{1}{\tau} \int v(\phi(\mathbf{x}, t_0, \tau))\partial\tau \qquad (3)$$

where u and v are the vertical and horizontal motion components of the optical flow field $\mathbf{v}(\mathbf{x}, t) = [u(\mathbf{x}, t)\ v(\mathbf{x}, t)]^T$ at time t. As with all Lagrangian measures, t_0 marks the

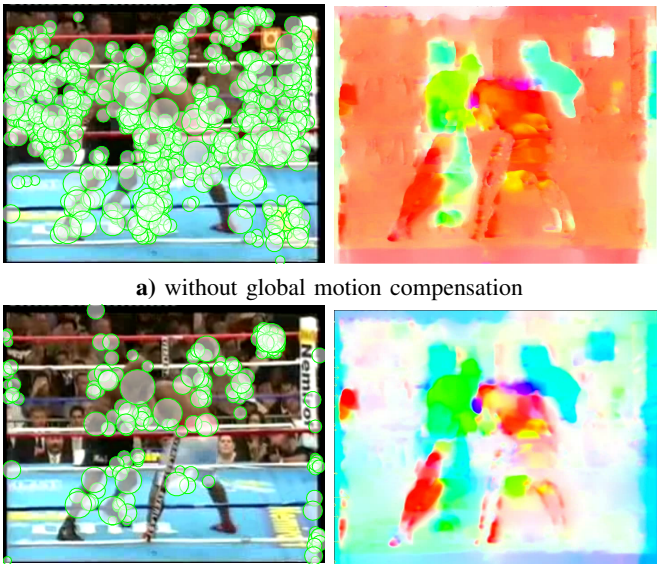
corresponding starting frame, while τ defines the complexity and temporal range of the resulting field. Since the optical flow field is discrete in time and space, trilinear interpolation and a lower-order integration scheme (i.e., RK2) is used to obtain values in the subpixel domain. The direction field is the mean motion direction and velocity information estimated for a temporal range τ of the trajectories starting from each point in the image at time t_0. In general, the starting points $[x_0, t_0]$ are independent of the original video resolution, i.e. the discrete sequence of optical flow fields can be sub- or oversampled, thus leading e.g. to a motion description with an implicit super-resolution. If $\tau = 1$ and the direction field is estimated with the original resolution, the direction field is the optical flow field of time t_0. For $\tau > 1$, the integration is continued over the next frames according to Eq. 3. Note that compared to FTLE this measure is *not* Galilean invariant, i.e. translating the frame of reference of the optical flow fields (e.g., due to camera motion) will influence the resulting direction fields.

Figure 2 shows an example of the direction Lagrangian measures for four and eight frames. For visualization, we transform the resulting flow map direction values into the HSV color space and project the resulting color back to the original starting frame (in analogy to common local optical flow depictions). The resulting hue value (H) represents the direction (or angle) of the flow map displacement, while the saturation (S) indicates the magnitude of the displacement, and V is kept constant. The direction fields estimated for different values of τ capture events of different time durations in a video. A punch, as shown in the top row, is a short-time event and best visible at short integration times. The chearleader dancing shown in the second row is captured best at longer time scales because its motion is much slower than the punch and can be perceived for a longer time. Figure 2 shows that the direction measure allows to distinguish motion patterns on different temporal scales in a video sequence.

IV. Point Detection and Feature Description

For the task of automated detection of violence in videos, representations using local features have been established. Comparative studies [8], [10] show that in this task the MoSIFT algorithm outperforms common local features such as HoG, HoF and STIP. The MoSIFT descriptor combines the histogram of oriented gradients appearance model from the SIFT descriptor and the histogram of flow motion model obtained by two-frame optical flow fields. We propose a Lagrangian-based local feature based on the Lagrangian direction field. This field has the same structure and a similar interpretation as the optical flow. Thus it can be integrated into the MoSIFT structure by substituting the optical flow field by the Lagrangian measure. The proposed Lagrangian Scale Invariant Feature Transform (LaSIFT) is related to the MoSIFT but differs in the following aspects:

Motion estimation proposed by Chen *et al.* [24] applies the pyramidal Lucas-Kanade (PLK) [36] method on each level of the scale-space image pyramids to estimate an optical flow field for each scale. Since the optical flow computation can



a) without global motion compensation

b) using global motion compensation

Fig. 3. Comparison between interest point detections and the direction field when applying global motion compensation.

be a bottleneck, we propose to estimate the optical flow and direction Lagrangian fields at original resolution and build the corresponding scale-space pyramid from these fields. This increases the speed of the scale-space motion estimation significantly. In our experiments, we were able to confirm the findings from common benchmarks (e.g. [37]) that the DualTVL1 method [38] is a both more accurate and faster optical flow method than PLK.

Global motion compensation has been implemented in order to compensate for camera motion which has a significant impact to the motion signature of the direction fields. As proposed in [11], we assume a homography-based background global motion model that excludes independently moving objects. The compensated direction field can be found by subtracting the background direction field from the actual one by:

$$\tilde{\Lambda}_{X/Y}(\mathbf{x}, t_0) = \underbrace{\frac{1}{\tau}\int \mathbf{v}(\phi(\mathbf{x}, t_0, \tau))\partial\tau}_{\Lambda_{X/Y}} - \underbrace{\frac{1}{\tau}\int \mathbf{v}^{GM}(\phi(\mathbf{x}, t_0, \tau))\partial\tau}_{\Lambda^{GM}_{X/Y}} \quad (4)$$

where $\tilde{\Lambda}_{X/Y}$ denotes the global motion compensated direction field, with $\Lambda^{GM}_{X/Y}$ the direction field of the background and $\mathbf{v}^{GM}$ the background optical flow. This field of the background is given by particle advection described by the concatenated homographies that are estimated for the consecutive images for the time-span τ. Due to the linearity of the pathline integration and the homography estimation, the background direction field of a certain position $\mathbf{x}$ at time t_0 can be directly estimated in the Lagrangian domain with:

$$\Lambda^{GM}_{X/Y}(\mathbf{x}, t_0) = \begin{bmatrix} \frac{m^{\Lambda}_0(t_0)\cdot x + m^{\Lambda}_1(t_0)\cdot y + m^{\Lambda}_2(t_0)}{m^{\Lambda}_6(t_0)\cdot x + m^{\Lambda}_7(t_0)\cdot y + 1} - x \\ \frac{m^{\Lambda}_3(t_0)\cdot x + m^{\Lambda}_4(t_0)\cdot y + m^{\Lambda}_5(t_0)}{m^{\Lambda}_6(t_0)\cdot x + m^{\Lambda}_7(t_0)\cdot y + 1} - y \end{bmatrix} \quad (5)$$

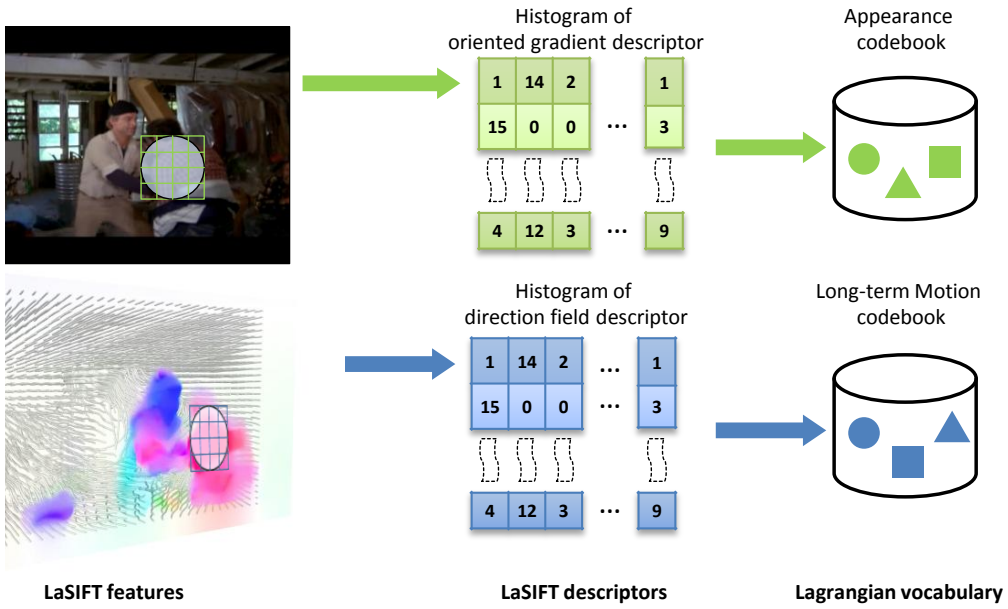


Fig. 4. LaSIFT feature extraction and Lagrangian vocabulary estimation. Left images shows the reference frame and the volume rending of the pathlines and direction field. After the feature detection, appearance descriptors (i.e histogram of oriented gradients, top) and long-term motion descriptors (i.e. histogram of direction field, bottom) are estimated on the whole dataset. With the total number of descriptors the appearance and long-term motion codebook are generated separately.

where $\mathbf{m}^{\Lambda}$ are the parameters of an eight-parameter homography global model estimated using the robust RANSAC method [39]. In this step, a regularly subsampled Lagrangian field provides a set of motion-like vectors and the point correspondence input. Compared to the current field the resulting rectified direction field suppresses the background camera motion and represents the moving foreground objects better. Thereby the motion compensation is only applied to the Lagrangian measure estimation and not to the particle advection itself, thus in contrast to [11] errors caused by the global motion estimation only affect the Lagrangian measure temporally but not the motion trajectory estimation, i.e. flow map estimation, itself.

Note that in general, motion compensation mainly affects the scaling of the Lagrangian field (including sign changes), while topological properties, in terms of minimum and maximum extremal regions, are generally preserved.

Interest Point Detection is based on the SIFT detector. This algorithm implements a difference-of-Gaussian scale-space detector which is salient at blob-like structures in multiple scales. Similar to [24], distinctive interest points with sufficient motion will be extracted. If a candidate point's compensated direction field vector is too small, the feature is considered to be too similar to camera motion and thus will be removed. This step allows to extract only features related to human action even under camera motion. Fig. 3 gives an example of feature detection with and without global motion compensation.

Feature description contains two parts: i) the SIFT appearance descriptor, which is a grid-based aggregated histogram of oriented gradients and ii) the long-term motion descriptor, which is a grid-based aggregated histogram of the direction field vectors from the same surrounding regions. The region is split into 4×4 cells and for each cell a histogram of eight bins is formed. Therefore for each pixel in the cell the orientation bins are voted with corresponding magnitudes. For the appearance descriptor the magnitudes, i.e. $||[I_x(\mathbf{x})\ I_y(\mathbf{x})]^T||$ and orientation, i.e. $arctan(I_y(\mathbf{x})/I_x(\mathbf{x}))$ are estimated with the spatial gradients $I_x(\mathbf{x})$ and $I_y(\mathbf{x})$, and for the long-term descriptor magnitudes, i.e $||[\Lambda_X(\mathbf{x})\ \Lambda_Y(\mathbf{x})]^T||$ and orientation, i.e. $arctan(\Lambda_Y(\mathbf{x})/\Lambda_X(\mathbf{x}))$ are estimated with the directional fields. Contrary to MoSIFT, we do not concatenate both descriptors as we want to integrate the dependency of appearance and motion in a so-called late fusion manner. Experiments have shown that this strategy outperforms the early fusion proposed in [24]. In the following paragraph, we will show how the proposed LaSIFT feature can be used to describe complex video content and how it is integrated into a framework for detection of violence in videos.

V. Scale-sensitive Video Representation

In order to efficiently exploit local features for violence detection, many authors proposed using a bag-of-words approach [4], [8], [22]. In these methods, codebooks are used to quantize features based on their components and accumulate them into fixed-dimensional histograms. Codebooks are typically built from cluster centers obtained from k-means clustering. These cluster centers can be interpreted as vocabulary and are also known as visual words. In the proposed framework, we use a histogram-intersection-based clustering method proposed by Wu *et al.* [40] who showed that this method substantially improves the overall accuracy of the system while the computational complexity remains almost as low as for k-means. Experimental results in [8] confirm this finding also for the special case of violence detection in videos.

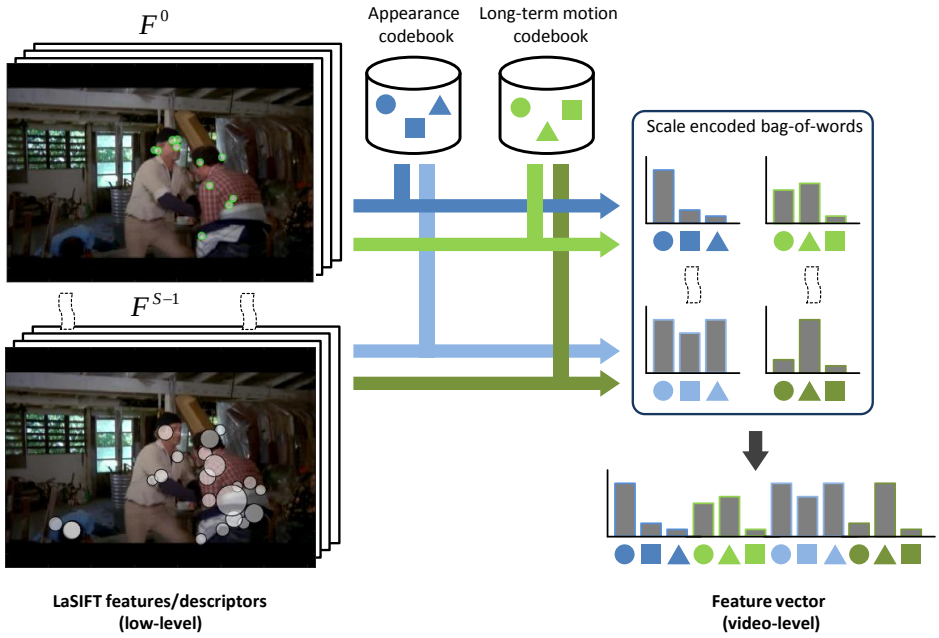


Fig. 5. Scale-sensitive video representation. The set of LaSIFT descriptors of a video sequence are partitioned into $S-1$ subsets F^s with respect to LaSIFT feature scale. The circular label in the left images visualizes the scale of the detected LaSIFT features. For each subset F^s an appearance and a long-term motion bag-of-word histogram is generated. The final video-level feature is a result of the concatenation of appearance and long-term motion histograms for each subset F^s.

In our method, appearance- and long-term motion-based descriptors are considered separately. For both modalities separate codebooks are trained, resulting in a Lagrangian vocabulary (Figure 4). Training separately reduces the sparsity in the vocabulary and thus enhances the generalization capacity. This strategy is motivated by properties of the datasets used for evaluation: The number of feature points over the whole dataset (typically between 300.000 and 1.000.000) is rather small in relation to the dimensionality of the LaSIFT descriptor. If joint descriptors were to be used, the number of usable codewords would have to be reduced due to sparsity in the data. Therefore, two codebooks for appearance and long-term motion information are created using the training data.

An important information provided by many feature point detectors is the spatial scale information which maps the size of the extracted feature to the image. In current approaches, including our previous work [19], this information remains unused. However, considering the presence of macroscopic as well as microscopic events in the scene, the spatial scale is a valuable cue for the characterization of the complex motion patterns in a video scene. For instance, the rather large motion signature of a person's torso should not be confused with smaller features captured on a hand or foot. Consequently, we propose a video-level feature which takes the scale of the LaSIFT feature into consideration. An overview of the scale sensitive video representation scheme is given in Figure 5. We partition the LaSIFT features and the corresponding appearance and long-term motion descriptors by their respective scale into several subgroups F^s with $s = 0 \ldots S-1$ of equally-sized intervals. The maximal interval bound is defined by the maximal scale which has been observed in the dataset.

For each LaSIFT descriptor set extracted from the video and selected for a certain scale interval, two separate bag-of-words histogram descriptors are estimated containing the appearance- and long-term motion visual and Lagrangian word frequencies. The scale-sensitive video-level descriptor is built by combining the concatenated appearance and long-term motion histograms of each scale. Consequently, the resulting scale-sensitive video-level descriptor is a $2 \times S$-dimensional vector. The final classification is obtained by using a support vector machine with a nonlinear χ^2-kernel [41].

In contrast to [19] no thresholding and channel-based normalization of the video-level descriptor is required, since the scale-sensitive descriptor is a higher dimensional vector and the visual and Lagrangian word frequencies have been distributed more evenly along the scales. As an optional adjustment we added an offset $\epsilon = 1$ to the scale-sensitive descriptor, which improves the overall stability of the support vector estimation, since the descriptor may contain a large number of zeros.

VI. Experiments

We evaluated our approach on three common benchmarks created for violent video detection (Hockey Fight, Violence in Movies and Violent Crowd) and performed tests on a proprietary real-world dataset (London Riots 2011), a subset of video footage captured at the London Riots in 2011. The latter dataset is used to demonstrate the performance of the system for an actual use-case. A particular focus will be on the Violent Crowd and London Riots 2011 dataset. In contrast to the Hockey Fight and Violence in Movies datasets, which contain only close-ups of person-on-person fights, Violent Crowd and London Riots 2011 contain crowded indoor and outdoor scenes and thus are a more realistic benchmark for

video-surveillance scenarios. Figure 6 gives a brief overview of typical 'violent' and 'non-violent' events in the datasets used whereby each dataset reflects different conditions in terms of number of people involved, camera motion, location and view-point. Note that the information about 'violence' or 'non-violence' is implicitly contained in the the datasets by the baseline classification of the underlying training datasets. Our system does not represent actual context-dependent violence against humans or groups of humans, but rather classifies human-action and background motion signatures that discriminate the predefined training videos.

The **Hockey Fight** dataset has been presented by Nievas *et al.* [8]. The footage contains 1000 short video clips from hockey games of the National Hockey League. Each video clip is about 50 frames long and has a resolution of 360×288. The clips contain a number of person-on-person fights mostly captured from a close distance. The dataset has several difficulties: different point of views, camera motion and a unknown number of involved actors. Especially, the motion blur around the very fast moving arms and legs is challenging for the optical flow-based motion estimation.

The **Violence in Movies** dataset has been introduced by Nievas *et al.* [8], too. In 200 short video clips 100 person-on-person fights are shown. The collection contains 100 non-fight scenarios containing various sport events and samples from the Weizmann dataset for action recognition [42]. Each sequence contains about 50 frames and has a resolution of 720×480 except some sequences have a resolution of 720×576. Compared to the Hockey Fight dataset, this dataset has a higher variety of the scenes and suffers from the interlacing artifacts. However, the detection of violent and non-violent videos is simplified because the fight scenes have similar structure and backgrounds differ a lot from non-violent scenes.

The **Violent Crowd** dataset has been published by Hassner *et al.* [9]. It comprises a collection of YouTube videos and includes 246 short video sequences which have been captured in a variety of arenas (as opposed to the Hockey Fight and Violence in Movies datasets). The scenarios in this dataset are manifold and include, for instance, football stadia, bars, and demonstrations. Both indoor and open areas are covered using static and non-static cameras. The image resolution of this dataset is 320×240 and the video length is varying from around 50 to 150 frames. Major difficulties on this dataset arise due to the image quality which is affected by compression artifacts, motion blur, text overlay, flash lights, and varying temporal resolutions. All these factors make the extraction of accurate motion information very challenging.

The **London Riots 2011** is a non-public dataset which has been composed in order to assess our system's accuracy using real-world data. The videos have been captured by the London Metropolitan Police during the disturbances across England in 2011 and show lootings, violent rallies, vandalism, and other violent scenes by a variety of actors. Videos used from this dataset typically have a resolution of 704×625 and show footage from non-static CCTV cameras with both overviews and heavy zoom-ins. Overall image quality is poor: low contrast, reflections, and motion blur are frequent. Videos show scenarios with both crowds and single actors and have been annotated manually. The footage has been divided into 50 videos containing violence and 50 videos capturing normal activities.

To assess the performance of our system we use the 5-fold cross validation as proposed by Hassner *et al.* [9]. For each of the five runs, the training set is used to generate the codebooks with 500 words. We chose this number in order to be comparable with [8]–[10]. The Lagrangian fields are estimated based on the Dual TV-L1 optical flow [38].

In our experiments, we evaluated the influence of the integration time τ and the spatial scale-sensitivity S. Performance comparison is done in terms of accuracy and area-under-curve of receiver-operating-characteristic (ROC-AUC). These measures are commonly used in the literature, but are not equivalent. The results indicate that, depending on the measure chosen, the optimal system configuration can be different. We report both measures in order to provide comparison with other state-of-the-art methods.

The classification metrics for different τ and S for the Hockey Fight, Violent Crowd, Violence in Movies and London Riots 2011 datasets are shown in Fig. 7. It can be seen that both parameters have a significant influence on the accuracy and ROC-AUC measure. Except for the Riot dataset, the performance decreases if τ is chosen too small or too large. The effect is with a minimal standard deviation for accuracy of 2.1% most significant on the Violent Crowd dataset and with 0.41% least significant on the Violence in Movies dataset. This supports the assumption that there is an optimal integration time related to the violent events occurring in the datasets.

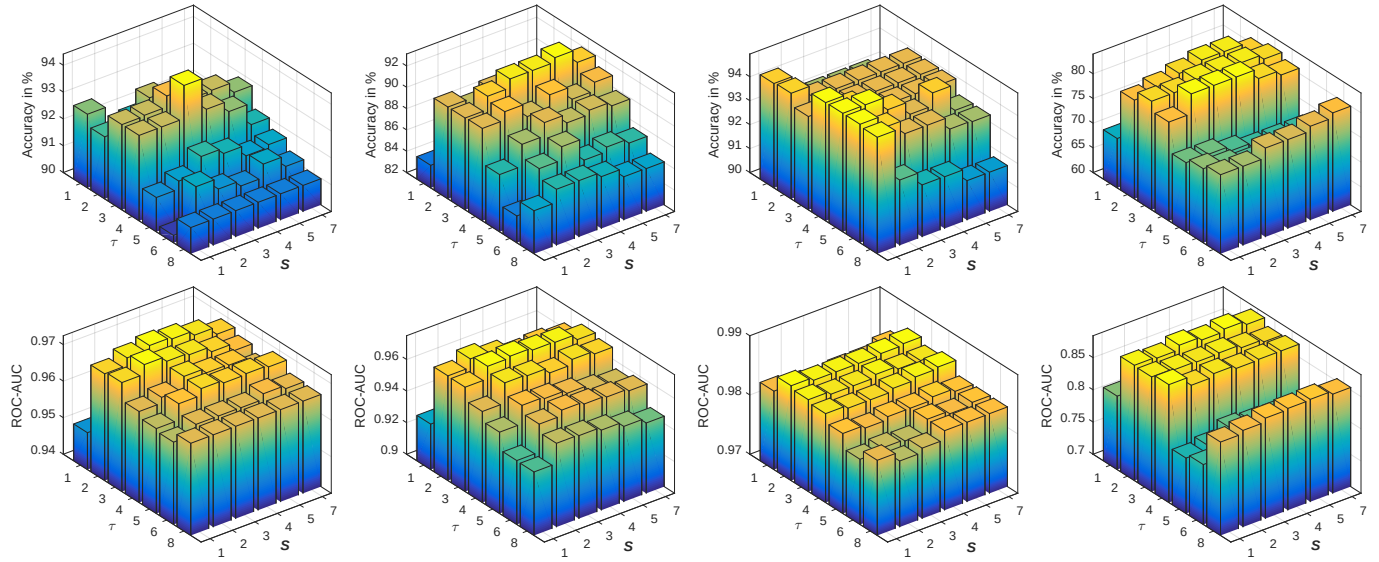
Similar observations can be found for the changes in S expect for the Violence in Movies data. It can be further concluded that there is an optimal partition of the local feature related to a specific motion structure size. Both findings quantify appropriate spatio-temporal scales of characteristic motion signatures for the considered scenarios.

We further compare the performance our system with recent state-of-the-art methods by selecting the configuration with the optimal accuracy values. The numerical results are shown in the related Tables I, II, III, IV. Apart from MoSIFT which denotes the baseline, HOT [26] and the well-established Dense Trajectories [11] based on long-term motion information are considered for comparison. Furthermore, a two-stream CNN proposed by Feichtenhofer *et al.* [43] has been adjusted to perform violence detection for comparison. The modified two-stream CNN has been initialized with temporal and spatial VGG-16 networks provided by Feichtenhofer, pre-trained with UCF101 action classification dataset and refined on the violence dataset. In addition, Substantial Derivatives [27] and Interaction Force [16], which are both based on fluid dynamical concepts, will be of special interest as they are most related to the proposed Lagrangian approach.

The entry 'SIFT' in the tables denotes the performance of our system, i.e. with scale-sensitive coding and motion-based feature selection, when using only appearance description and $\tilde{\Lambda}_{X/Y}$ will denote the performance of our system when using only long-term motion description. This allows to distinguish the influence of the scale-sensitive classification framework proposed in Section V and the Lagrangian descriptor pro-

a) Hockey Fight **b)** Violent Crowd **c)** Violence in Movies **d)** London Riots 2011

Fig. 6. Sample frames from video sequences of violent (top) and non-violent (bottom) content for each of the datasets considered in our study. Due to privacy regulations, real-world content from London riots has been anonymized before publishing.



a) Hockey Fight **b)** Violent Crowd **c)** Violence in Movies **d)** London Riots 2011

Fig. 7. Violence detection performed with LaSIFT configurations varying the integration times τ and the number of applied scale intervals S. Comparison of the mean accuracy and the area under the ROC curve (AUC) with 5-fold cross validation on Hockey Fight, Crowd Violence, Violence in Movies and London Riots 2011 dataset.

posed in Section IV. In addition, for the Hockey Fight and Violent Crowds datasets we provide numerical results of our baseline work proposed in [19] denoted with LaSIFT$^{\propto}$(mix) and LaSIFT$^{\propto}$, where (mix) denoted the combination of video-level descriptors from different integration intervals. This comparison underlines the advantage of the novel scale-sensitive video representation proposed in Section V leading to further performance improvements of the LaSIFT violent video classification. For the Violent Crowds dataset the proposed video-level description based on early fusion (LaSIFT^{early}) has been compared. This experiment demonstrates the superior performance of the late fusion applied in the remaining experiments.

Comparing the numerical results of the Hockey Fight and Violence in Movies benchmarks (see Table I and II) shows that our SIFT implementation outperforms most well-established features such as MoSIFT [24] or VIF [9]. This indicates that for person-on-person actions the global motion compensation, which inhibits placement of feature points in the background, significantly improves the performance of the classification. The accuracy was then further improved by integrating the Lagrangian motion model. Consequently, the proposed method sets the state-of-the-art for the Hockey Fight benchmark.

Table III shows a state-of-the-art comparison for the Violent Crowd dataset. This dataset reflects best the challenges of modern violence classification systems processing CCTV data. For this dataset, an accuracy improvement of around

TABLE I
COMPARISON OF VIOLENCE DETECTION PERFORMANCE ON HOCKEY FIGHT DATASET BETWEEN LASIFT AND STATE-OF-THE-ART METHODS. SIFT DENOTES AN EVALUATION OF THE PROPOSED SYSTEM, I.E. WITH THE PROPOSED FEATURE ENCODING SCHEME, BUT BASED ON APPEARANCE MODEL AND $\tilde{\Lambda}_{X/Y}$ ON THE LAGRANGIAN MODEL ONLY.

Method	ACC ± SD	ROC-AUC
STIP(HoG) + BoW [8], [9]	91.7	-
STIP(HoF) + BoW [8], [9]	88.6	-
MoSIFT + BoW [10]	90.9	-
MoSIFT + KDE + SC [10]	94.0±1.97%	0.9666
LaSIFT$^{\propto}(\tau = 4)$	92.42±2.57%	0.9682
LaSIFT$^{\propto}$(mix)	93.32±2.24%	**0.9732**
SIFT + BoW($S = 3$)	91.51±4.83%	0.9563
$\tilde{\Lambda}_{X/Y}(\tau = 4)$ + BoW($S = 3$)	81.54±10.03%	0.8996
LaSIFT($\tau = 4$) + BoW($S = 3$)	**94.42±2.82%**	0.9699

TABLE II
COMPARISON OF VIOLENCE DETECTION PERFORMANCE ON VIOLENCE IN MOVIES DATASET BETWEEN LASIFT AND STATE-OF-THE-ART METHODS ($^{\#}$RESULTS ARE TAKEN FROM [27]).SIFT DENOTES AN EVALUATION OF THE PROPOSED SYSTEM, I.E. WITH THE PROPOSED FEATURE ENCODING SCHEME, BUT BASED ON APPEARANCE MODEL AND $\tilde{\Lambda}_{X/Y}$ ON THE LAGRANGIAN MODEL ONLY.

Method	ACC ± SD	ROC-AUC
Jerk$^{\#}$ [3]	95.02±0.56%	-
STIP (HoG) + BoW [8]	44.5%	-
STIP (HoF) + BoW [8]	50.5%	-
MoSIFT + BoW [8]	89.5%	-
VIF$^{\#}$ [9]	91.31±1.06%	-
Interaction Force$^{\#}$ [16]	95.51±0.79%	-
$F^L \vert F^{Cv}$ [27]	96.89±0.21%	-
VIPS [25]	**96.91%**	-
SIFT($S = 2$) + BoW	93.33±6.99%	0.9807
$\tilde{\Lambda}_{X/Y}(\tau = 2)$ + BoW($S = 5$)	93.40±4.90%	0.986
LaSIFT($\tau = 2$) + BoW($S = 5$)	94.95±4.57%	**0.9830**

TABLE III
COMPARISON OF VIOLENCE DETECTION PERFORMANCE ON VIOLENT CROWD DATASET BETWEEN LASIFT AND STATE-OF-THE-ART METHODS ($^{\#}$RESULTS ARE TAKEN FROM [27]).SIFT DENOTES AN EVALUATION OF THE PROPOSED SYSTEM, I.E. WITH THE PROPOSED FEATURE ENCODING SCHEME, BUT BASED ON APPEARANCE MODEL AND $\tilde{\Lambda}_{X/Y}$ ON THE LAGRANGIAN MODEL ONLY.

Method	ACC ± SD	ROC-AUC
Jerk$^{\#}$ [3]	74.18±0.85%	-
LTP [9]	71.53±0.17%	0.7986
VIF [9]	81.30±0.21%	0.8500
HoG + BoW [8], [9]	57.43±0.37%	0.6182
HoF + BoW [8], [9]	58.53±0.32%	0.5760
MoSIFT + BoW [10]	83.42±8.03%	0.8751
MoSIFT + KDE + SC [10]	89.05±3.26%	0.9357
MLV [28]	84.44%	0.8800
Dense Trajectories$^{\#}$ [11]	79.38±0.14%	-
Interaction Force$^{\#}$ [16]	74.50±0.65%	-
HOT [26]	82.30%	-
$F^L \vert F^{Cv}$ [27]	85.43±0.21%	-
VIPS [25]	86.61%	-
Two-Stream CNN (VGG-16) [43]	91.83±3.34	-
LaSIFT$^{\propto}(\tau = 3)$	92.01±8.01%	**0.9741**
LaSIFT$^{\propto}$(mix)	92.01±8.01%	0.9729
LaSIFT$^{early}(\tau = 3)$ + BoW($S = 5$)	87.82±8.70%	0.9306
SIFT + BoW($S = 5$)	73.13±4.39%	0.8521
$\tilde{\Lambda}_{X/Y}(\tau = 3)$ + BoW($S = 5$)	81.37±5.13%	0.8948
LaSIFT($\tau = 3$) + BoW($S = 5$)	**93.12±8.77%**	0.9731

TABLE IV
COMPARISON OF VIOLENCE DETECTION PERFORMANCE ON LONDON RIOT 2011 DATASET BETWEEN LASIFT AND MOSIFT.SIFT DENOTES AN EVALUATION OF THE PROPOSED SYSTEM, I.E. WITH THE PROPOSED FEATURE ENCODING SCHEME, BUT BASED ON APPEARANCE MODEL AND $\tilde{\Lambda}_{X/Y}$ ON THE LAGRANGIAN MODEL ONLY.

Method	ACC ± SD	ROC-AUC
MoSIFT + BoW	72.38± 11.86	0.790
SIFT + BoW($S = 4$)	78.00± 8.24	0.810
$\tilde{\Lambda}_{X/Y}(\tau = 4)$ + BoW($S = 4$)	74.00±9.62%	0.785
LaSIFT($\tau = 4$) + BoW($S = 4$)	**84.00± 7.42**	**0.874**

20% has been obtained. The classification result significantly benefits from the integration of the Lagrangian descriptors. The application of the long-term motion descriptor $\tilde{\Lambda}_{X/Y}$ achieves competitive results to most of the state-of-the-art methods. For the Violent Crowd dataset the combination of the appearance and long-term motion descriptor results into the highest performance gain and outperforms comparative long-term motion based methods such as HOT and Dense Trajectories. Whereas $F^L|F^{Cv}$ [27] with a difference of 1.94% slightly outperforms our approach at the Violence in Movies dataset, our approach achieves an about 7.69% better accuracy on the Violent Crowd dataset. Despite the accuracy being reduced on the high quality Violence in Movies data, this shows that the Lagrangian-based video-level representation is more robust on video footage with low visual quality data and also robust to the application of a larger variety of scenarios. Especially the optical flow estimation suffers from the low video quality, bad contrast and lots of block-like coding artifacts contained in the Violent Crowd data but this effect is alleviated for direction fields because the flow map integration contains an implicit denoising.

Finally, the numerical results for London Riots 2011 dataset (non-public) are given in Table IV. The LaSIFT classification framework has been compared against the scale-sensitive video-level descriptor with the single appearance (SIFT) and Lagrangian model ($\tilde{\Lambda}_{X/Y}$), and the MoSIFT feature with the baseline bag-of-word video-level representation. The Lagrangian approach outperforms significantly the baseline MoSIFT. Our proposed system using $\tau = 4$ and $S = 4$ achieves both high accuracy and ROC values. For the London Riots 2011 with a resolution of 704×625 the proposed system operates with 0.62 fps on a Intel i7 with 3.4 GHz.

Figure 8 shows exemplary failure cases of the proposed systems. The causes for misclassification can be manifold, e.g in Fig. 8(a) the boxing event only takes place in the last frames of the sequence, Fig. 8(b) is affected by very strong motion blur. Fig. 8(c,d) shows rare movements such as a runner's arm movements or the bending down of a person which can have similar motion characteristics as violence, e.g. boxing, and are thus difficult to discriminate.

In summary, the experiments have shown that with the proposed LaSIFT descriptor the complementary appearance and long-term motion information have been successfully combined. This indicates the importance of long-term motion cues in this data where the discrimination power of the appearance model is low due the low contrast or visual quality of real-world video footage as in the London Riots 2011 or

Fig. 8. Samples frames of misclassified video sequences.

Violent Crowd datasets. The good performance for each of the datasets underlines that our system is capable of dealing with a variety of scenarios including real-world footage from police sources.

VII. Conclusion

In this paper we presented a novel approach for violence detection in videos which is based on Lagrangian measures. Lagrangian measures, as a tool from Lagrangian theory describing non-linear dynamic systems, have been revised and adopted for video analytics. Dynamic characteristics of moving objects in video footage can efficiently be described using a direction-based Lagrangian field measure that offers a appropriate trade-off between discriminative efficiency and computational complexity. The proposed measure comprises salient motion information over multiple time scales τ and represents a more robust alternative to gradient-based measures, such as FTLE, in low-quality CCTV scenarios. In order to develop a feature for violence analysis in videos, we integrated this concept into the LaSIFT method which includes both appearance information and long-term motion cues. We further proposed a framework for violence detection based on a bag-of-words approach including a scale-sensitive feature encoding scheme and a late-fusion approach.

The proposed framework has been extensively tested on various challenging datasets and on non-public, real-world data obtained by the London Metropolitan Police. Our method shows good accuracy and improves upon multiple state-of-the-art algorithms. As violence detection can be see as a subclass of context-based video classification and action recognition, in the future we want to extend the application of our approach to these more general fields and show its suitability for other areas in computer vision.

Acknowledgment

The research leading to these results has received funding from the European Community's FP7 and BMBF-VIP+ under grant agreement number 607480 (LASIE) and 03VP01940 (SiGroViD). The images in this paper are partially created using the Amira visualization software.

References

[1] M. Sjöberg, B. Ionescu, J. Yu-Gang, V. L. Quang, M. Schedl, and C.-H. Demarty, "The MediaEval 2014 Affect Task: Violent Scenes Detection," in *Working Notes Proceedings of the MediaEval 2014 Workshop*, 2014.

[2] E. Acar, F. Hopfgartner, and S. Albayrak, "Violence detection in hollywood movies by the fusion of visual and mid-level audio cues," in *International Conference on Multimedia*, 2013, pp. 717–720.

[3] A. Datta, M. Shah, and N. D. V. Lobo, "Person-on-person violence detection in video data," in *International Conference on Pattern Recognition*, vol. 1, 2002, pp. 433–438.

[4] L.-H. Chen, H.-W. Hsu, L.-Y. Wang, and C.-W. Su, "Violence detection in movies," in *International Conference on Computer Graphics, Imaging and Visualization*, Aug 2011, pp. 119–124.

[5] B. Krausz and C. Bauckhage, "Automatic detection of dangerous motion behavior in human crowds," in *International Conference on Advanced Video and Signal Based Surveillance*, 2011, pp. 224–229.

[6] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European Conference on Computer Vision*, 2006, pp. 428–441.

[7] O. Déniz, I. Serrano, G. Bueno, and T.-K. Kim, "Fast Violence Detection in Video," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2014, pp. 478–485.

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

[9] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," *Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–6, 2012.

[10] L. Xu, C. Gong, J. Yang, Q. Wu, and L. Yao, "Violent video detection based on MoSIFT feature and sparse coding," in *International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 3562–3566.

[11] H. Wang and C. Schmid, "Action Recognition with Improved Trajectories," in *International Conference on Computer Vision*, 2013, pp. 3551–3558.

[12] A. Kuhn, T. Senst, I. Keller, T. Sikora, and H. Theisel, "A lagrangian framework for video analytics," in *Workshop on Multimedia Signal Processing*, 2012, pp. 387–392.

[13] B. E. Moore, S. Ali, R. Mehran, and M. Shah, "Visual crowd surveillance through a hydrodynamics lens," *Communications of the ACM*, vol. 54, pp. 64–73, 2011.

[14] T. Li, H. Chang, M. Wang, B. Ni, and R. Hong, "Crowded Scene Analysis : A Survey," *Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 3, pp. 367–386, 2015.

[15] S. Ali and M. Shah, "A Lagrangian Particle Dynamics Approach for Crowd Flow Segmentation and Stability Analysis," in *International Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–6.

[16] R. Mehran, B. E. Moore, and M. Shah, "A streakline representation of flow in crowded scenes," in *European Conference on Computer Vision*, vol. 6313, 2010, pp. 439–452.

[17] E. Acar, T. Senst, A. Kuhn, I. Keller, H. Theisel, S. Albayrak, and T. Sikora, "Human action recognition using lagrangian descriptors," in *International Workshop on Multimedia Signal Processing*, 2012, pp. 360–365.

[18] T. Senst, A. Kuhn, H. Theisel, and T. Sikora, "Detecting people carrying objects utilizing lagrangian dynamics," in *International Conference on Advanced Video and Signal-Based Surveillance*, 2012, pp. 398–403.

[19] T. Senst, V. Eiselein, and T. Sikora, "A Local Feature based on Lagrangian Measures for Violent Video Classification," in *International Conference on Imaging for Crime Prevention and Detection*, 2015, pp. 1–6.

[20] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[21] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[22] F. D. M. de Souza, G. C. Chávez, E. A. do Valle, and A. de Albuquerque Araújo, "Violence Detection in Video Using Spatio-Temporal Features," in *Conference on Graphics, Patterns and Images*, 2010, pp. 224–230.

[23] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[24] M.-y. Chen and A. Hauptmann, "MoSIFT : Recognizing Human Actions in Surveillance Videos," *Technical Report CMU-CS-09-161*, pp. 1–16, 2009.

[25] S. Mohammadi, A. Perina, H. Kiani, and V. Murino, "Angry Crowds: Detecting Violent Events in Videos," in *European Conference on Computer Vision*, 2016, pp. 3–18.

[26] H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, and V. Murino, "Analyzing Tracklets for the Detection of Abnormal Crowd Behavior," in *Winter Conference on Applications of Computer Vision*, 2015, pp. 148–155.

[27] S. Mohammadi, A. Perina, and I. Italiano, "Violence Detection in Crowded Scenes using Substantial Derivative," in *Conference on Advanced Video and Signal Based Surveillance*, 2015.

[28] H. Fradi, B. Luvison, and Q. C. Pham, "Crowd Behavior Analysis Using Local Mid-Level Visual Descriptors," *Transaction on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 589–602, 2017.

[29] A. Pobitzer, R. Peikert, R. Fuchs, B. Schindler, A. Kuhn, H. Theisel, K. Matkovic, and H. Hauser, "On the way towards topology-based visualization of unsteady flow-the state of the art," *H. und E. Reinhard (Hrsg.), Eurographics*, 2010.

[30] M. Uffinger, F. Sadlo, and T. Ertl, "A time-dependent vector field topology based on streak surfaces," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, no. 3, pp. 379–392, 2013.

[31] A. Kuhn, W. Engelke, C. Rössl, M. Hadwiger, and H. Theisel, "Time line cell tracking for the approximation of lagrangian coherent structures with subgrid accuracy," *Computer Graphics Forum*, vol. 33, no. 1, pp. 222–234, 2014.

[32] T. Peacock and J. Dabiri, "Introduction to focus issue: Lagrangian coherent structures," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 20, no. 1, p. 017501, 2010. [Online]. Available: http://link.aip.org/link/?CHA/20/017501/1

[33] S. Ameli, Y. Desai, and S. Shadden, "Developing flexible but efficient software for dynamical systems analysis of fluid flow," in *APS Meeting Abstracts*, Nov. 2012, p. 29010.

[34] A. Pobitzer, A. Lež, K. Matkovic, and H. Hauser, "A statistics-based dimension reduction of the space of path line attributes for interactive visual flow analysis," *Proceedings of the IEEE Pacific Visualization Symposium (PacificVis 2012*, pp. 113–120, 2012.

[35] L. Zhang, R. S. Laramee, D. Thompson, A. Sescu, and G. Chen, "Compute and visualize discontinuity among neighboring integral curves of 2d vector fields," *(to appear)*, 2015.

[36] J.-Y. Bouguet, "Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm," Intel Corporation Microprocessor Research Lab, Technical Report, 2000.

[37] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A Database and Evaluation Methodology for Optical Flow," Microsoft Research, Technical Report {MSR-TR-2009-179}, 2009.

[38] C. Zach, T. Pock, and H. Bischof, "A Duality Based Approach for Realtime TV-L1 Optical Flow," in *Proceedings of Pattern Recognition*, 2007, pp. 214–223.

[39] M. Tok, A. Glantz, A. Krutz, and T. Sikora, "Monte-Carlo-based Parametric Motion Estimation using a Hybrid Model Approach," *Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 607–620, 2013.

[40] J. Wu, W. Tan, and J. Rehg, "Efficient and effective visual codebook generation using additive kernels," *The Journal of Machine Learning Research*, vol. 12, pp. 3097–3118, 2011.

[41] S. Maji, A. C. Berg, and J. Malik, "Classification using Intersection Kernel SVMs is efficient," in *Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[42] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[43] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1933–1941.

Tobias Senst received the Dipl.-Ing. degree in computer engineering from Otto von Guericke Universität Magdeburg, Germany and is currently working towards the Ph.D. degree at the Communication Systems Group at Technische Universität Berlin, Germany. His main research interests include image and video processing, optical flow, feature tracking and video surveillance.

Volker Eiselein received the Dipl.-Ing. degree in computer engineering from Technische Universität Berlin, Germany and is currently working towards the Ph.D. degree at the Communication Systems Group at Technische Universität Berlin, Germany. His main research interests include image and video processing, video surveillance, object detection and tracking.

Alexander Kuhn is a Postdoctoral Research Fellow at the Department of Visualization and Data Analysis at the Zuse Insitute Berlin (ZIB), Germany. In 2009, he received the M.Sc. in Computational Visualistics and in 2012, a Ph.D. in Computer Science from the University of Magdeburg. His research interests are in the visualization, segmentation, and analysis of flow data.

Thomas Sikora Thomas Sikora is professor and director of the Communication Systems Group at Technische Universität Berlin, Germany. He received the Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering from Bremen University, Germany, in 1985 and 1989, respectively. In 1990, he joined Siemens Ltd. and Monash University, Melbourne, Australia, as a Project Leader responsible for video compression research activities in the Australian Universal Broadband Video Codec consortium. Between 1994 and 2001, he was the Director of the Interactive Media Department, Heinrich Hertz Institute (HHI) Berlin GmbH, Germany. Prof. Sikora is co-founder of Imcube GmbH and Vis-a-Pix GmbH, two Berlin-based start-up companies involved in research and development of audio and video signal processing and compression technology. Prof. Sikora has been involved in international ITU and ISO standardization activities as well as in several European research activities for a number of years. As the Chairman of the ISOMPEG (Moving Picture Experts Group) video group, he was responsible for the development and standardization of the MPEG-4 and MPEG-7 video algorithms. He also served as the chairman of the European COST 211ter video compression research group. He was appointed as Research Chair for the VISNET and 3DTV European Networks of Excellence. He is an Appointed Member of the Advisory and Supervisory board of a number of German companies and international research organizations. He frequently works as an industry consultant on issues related to interactive digital audio and video. Prof. Sikora is a Member of the German Society for Information Technology (ITG). He is a recipient of the 1996 ITG Award, co-receipient of the 1996 Engineering Grammy Award and the 2016 Google Faculty Research Award in Machine Perception. He has published more than 150 papers related to audio and video processing. He was the Editor-in-Chief of the IEEE Transactions on Circuits and Systems for Video Technology. From 1998 to 2002, he was an Associate Editor of the IEEE Signal Processing Magazine. He is an Advisory Editor for the EURASIP Signal Processing: Image Communication journal and was an Associate Editor of the EURASIP Signal Processing journal.