Best Paper Award, IET 6th International Conference on Imaging for Crime Prevention and Detection, 2015

A Local Feature based on Lagrangian Measures for Violent Video Classification

Tobias Senst, Volker Eiselein, Thomas Sikora

Communication Systems Group, Technische Universität Berlin, Germany

Keywords: violent video detection, local feature, action recognition, lagrangian measures, lagrangian framework

Abstract

Lagrangian theory provides a diverse set of tools for continuous motion analysis. Existing work shows the applicability of Lagrangian method for video analysis in several aspects. In this paper we want to utilize the concept of Lagrangian measures to detect violent scenes. Therefore we propose a local feature based on the SIFT algorithm that incooperates appearance and Lagrangian based motion models. We will show that the temporal interval of the used motion information is a crucial aspect and study its influence on the classification performance. The proposed LaSIFT feature outperforms other state-of-the-art local features, in particular in uncontrolled realistic video data. We evaluate our algorithm with a bag-of-word approach. The experimental results show a significant improvement over the state-of-the-art on current violent detection datasets, i.e. Crowd Violence, Hockey Fight.

1 Introduction

The detection of violence in videos has been insufficiently studied for video surveillance. Violent detection on movie databases [19] could benefit from the fusion of multi modal data sources [1] e.g. video, audio and context data. The task becomes more challenging for surveillance videos which typically do not have audio tracks or other contextual sources such as subtitles. In addition, surveillance video data is often far below motion picture quality. The need to work in continuous operation usually makes it hard to guarantee good and constant image quality. The color cues are not reliable and neither are the details required for fine scale action recognition. Although high definition cameras are becoming cheaper and cheaper, most existing CCTV cameras still capture in low resolution e.g. VGA or CIF. Many sophisticated action recognition algorithms are specialized on a microscopic level [17]. That means they threat each individual in the scene separately and thus need an efficient person or face detection [4]. Robustly detecting peoples in surveillance video is still an open challenge as the algorithm has to deal with many difficulties, especially when groups or crowds of persons are involved.

Motion information has become a key aspect in violence and action recognition applications, not least thanks to the performance gain of recent optical flow-based motion estimation methods. Hassner *et al.* [8] introduced a global Violent Flows (ViF) descriptor, that considers statistics of flow vector magnitude dynamics over time. Thex showed that ViF video features classified with a linear support vector machine can achieve real time performance. Low run time is also the motivation of the Déniz *et al.*. In [7] they presented a method that implicit measures the acceleration of global motion by comparing the power spectrum of consecutive frames. In action recognition local features have become the most prominent technique to represent video data. Using a codebook based on bag-of-words or Fisher vectors the features are quantized by their dynamic components and accumulated into fixed dimension histograms over the duration of the video. Classification is done by support vector machines. As violent scene detection can be seen as a subdomain of action recognition it is obvious to consider local features for violent scene detection too. Hassner *et al.* [8] show that local feature such as the histogram of gradients (HoG) and histogram of flow (HoF) [6] which perform well on recent action recognition benchmarks fail on the Crowd Violence dataset. Nievas *et al.* [15] found the same conclusion by comparing the STIP feature [10] with the MoSIFT [5] feature showing that the MoSIFT outperforms the STIP for the Hockey Fight dataset. Xu *et al.* [22] further improved the performance of the MoSIFT by substituting the bag-of-words with a sparse coding scheme.

Although motion information is a key property in violence detection, the current methods [6, 7, 8, 15, 22] only consider two frame motion information of pixel correspondences estimated by optical flow. However, the most motion signatures of the subjects in a scene are not static and homogeneous in time, e.g. running and punching persons have a very dynamic motion signature that changes over time. To discover more complex temporal characteristic and correlations larger temporal periods need to be taken into account. On a microscopic level, there exist several approaches that use space time volumes of stacked silhouettes, Hidden Markov models or other temporal state space models. A detailed overview has been given by Poppe [17]. However, the requirement of a precise segmentation of the subject body parts or its silhouette which makes these approaches unapplicable for most of surveillance data. Wang *et al.* [20] proposed to use long term trajectories for action recognition. These trajectories are created by tracking either feature points or densely sampled points with the optical flow field. Trajectories are usually discontinuous in space and asynchronous, i.e. they start and end in different frames so that a complete description of a video can only be guaranteed if the distribution of the tracked points reflect the underlying moving objects. Wang *et al.*implemented densely sampled trajectories to construct video features shown to perform best on a variety of datasets. Dense trajectories are characterized by their good performance but also by a high complexity as they are structures in a 3D space.

The Lagrangian theory for time dependent vector field analysis provides rich set of tools for continuous motion analysis. The analysis of a dynamic flow field is based on integral field lines that are similar to trajectories. In the video analysis context the integral field lines are constructed by particle advection based on the optical flow field. Methods based on Lagrangian theory have gained significant attention in the video analysis context as they transform motion information about a given time interval into a 2D space. The existing work shows the applicability of Lagrangian methods for video applications in several aspects e.g. on a macroscopic level for crowd analysis [9, 14]. The macroscopic perspective focuses on a crowd as one entity [11]. In our previous work we have extended this field of applications and shown that Lagrangian methods are also an valuable tool on a microscopic level e.g for human action recognition [2] and people carrying baggage recognition [18].

In this paper we want to utilize the concept of Lagrangian measures for the violent video detection task. Following the Lagrangian framework proposed in [9] we make use of the path line concept as the integral field lines in the unsteady flow field. From the path lines we derive a Lagrangian measure, the directional Lagrangian fields. These fields are similar to the optical flow field as they contain also two directional motion like components that are integrated over its path lines. These fields are meant to represent the dynamic patterns in the scene related to a given time scale. To encode spatio temporal patterns, inspired by the MoSIFT [5] algorithm, we will extend the SIFT [12] method. Therefore we have to integrate the Lagrangian motion information in addition to the SIFT appearance model. The feature encoding will be based on a bag-of-words approach. The bag-of-words provides a finite set of descriptors by which the quantized dynamic components will be accumulated into a fixed size histogram for each video. Each video can then be classified by a support vector machine. In our experiments we will study the impact of the integration interval as it is a crucial aspect for the Lagrangian measure. The experiments will be performed on the Crowd Violence [8] and Hockey Fight [15] violence detection datasets.

2 Lagrangian Measures for Violent Video Detection

In this section we briefly review Lagrangian theory and deduce the direction Lagrangian measure. The origins of Lagrangian methods are in analysis of general dynamical systems. These methods are one aspect of computational fluid dynamics systems. Thereby the Lagrangian methods are able to capture the intrinsic characteristics of dynamic systems on variable temporal scales and are suitable to capture complex patterns that are hidden within the motion of such systems. Within the video analysis the Lagrangian methods are applied to characterize the motion dynamics in video data based on a sequence of optical flow fields to assemble a time dependent vector field. We explicitly do not average consecutive frames as proposed in [3] to retain all spatio temporal information. The integral field lines are the bases of the dynamic flow field analysis. In the context of optical flow they can be re-interpreted as the trace of motion of a specific piece of information within the spatio temporal domain. There can be found different types of integral field lines in unsteady flow fields.

Streak lines and path lines are the most prominent field lines in video surveillance. Streak lines are integral curves traced from a point source that does not move with the flow field and path lines trajectories traced from a specific starting point in space and time. Streak lines have gained significant attention for crowd analysis [13]. In this work we follow the Lagrangian framework proposed in [9] and utilize path lines. Under optimal boundary conditions path lines behaves like a trace of the objects of on which they were initialized. In contrast, streak lines captures the motion characteristics of a static position in the video over the time. For a large temporal scale one stream line contains the motion characteristic of all crossing objects.

Formally the estimation of path lines can be described as follows: Given the optical flow vector field $\mathbf{v} = (\mathbf{x}, t)$ we can start a path line that denotes a particle trajectory. This can be formulated as an autonomous system:

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

for a space time point $(\mathbf{x}_0, t_0)$. One core aspect of the implementation is the computation of the flow map $\phi(\mathbf{x}, t_0, \tau)$. The flow map defines a mapping of an initial point at time t_0 to its advected positions after an integration time τ. The path line is a polynomial curve that results from the combination of all positions in the flow map for a specific point over the interval $[t_0, t_0 + \tau]$

To analyze unsteady flow properties in a feature-oriented manner the concept of Lagrangian Coherent Structures has been proposed. These structures directly describes properties of neighboring trajectories in the space time domain [16]. The most prominent techniques to extract Lagrangian Coherent Structures is the Finite Time Lyapunov Exponent (FTLE). The FTLE scalar field has been successfully used in video analytics [13, 18] before. Beside FTLE, Lagrangian measures, i.e. path line attributes, can be used to describe aspects of motion by LCS. These measures present several advantages. The most notable one is the ability to transform the motion information about LCS of a given time scale τ into a 2D space. The resulting Lagrangian scalar field implicit describes properties of the flow map and characterizes the motion information within the time interval τ. It can be directly processed by common image processing techniques. In general a large number of Lagrangian measures can be defined. Kuhn *et al.* [9] provides an outlook of Lagrangian measures for crowd analysis.

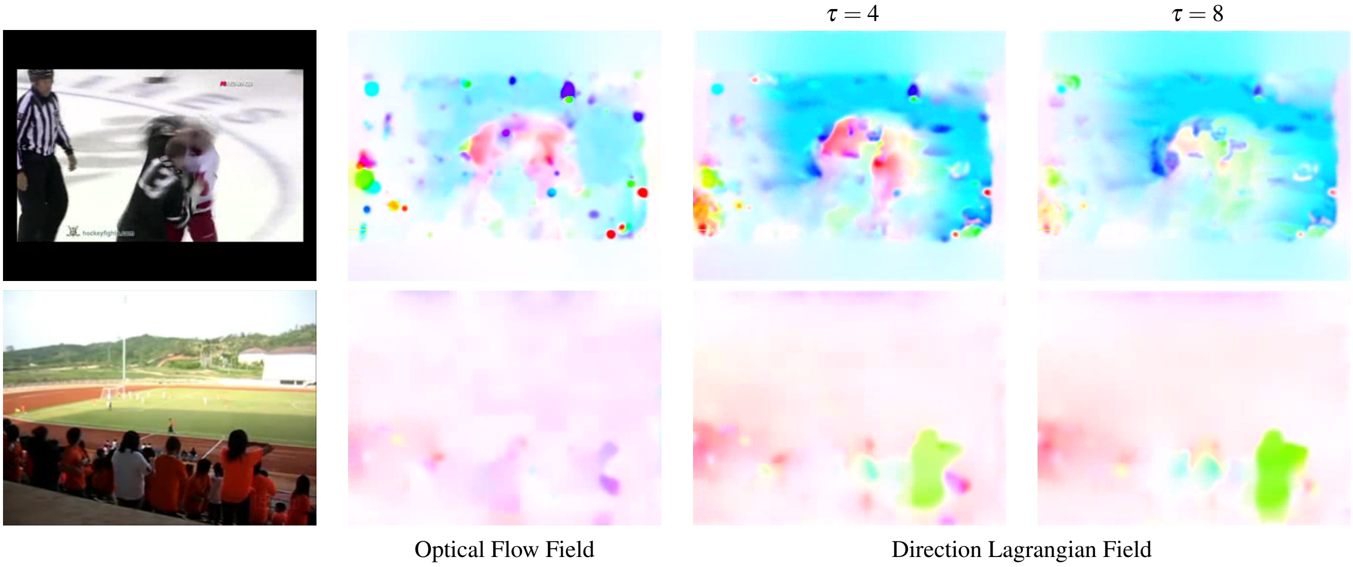

Figure 1: Optical flow and direction Lagrangian measures for different temporal scales applied to a fighting (top) and dancing crowd (bottom) sequence. Increasing τ allows to describe motion features on different temporal scales. While short term event such as boxing are present in short term integrated fields ($\tau = 4$), long term events such as the dancing person are present in integration over larger scales ($\tau = 8$).

In this work we will consider the x and y direction Lagrangian measure. The direction Lagrangian fields $\Lambda_{X/Y}(\mathbf{x}, t_0)$ can be obtained by estimating the average motion vector along the path line as follows:

$$\Lambda_{X/Y}(\mathbf{x}, t_0) = \frac{1}{\tau} \int v_{x/y}(\phi(\mathbf{x}, t_0, \tau)) \partial \tau, \quad (2)$$

where $v_{x/y}$ is a function providing the x or y motion components. The choice of the parameter τ is a crucial aspect, since it defines the temporal scale of the motion properties and the complexity of the Lagrangian feature we are interested in. The parameter defines also how many frames have been considered.

Figure 1 shows an example of the x and y direction Lagrangian measures for four and eight frames. As they are similar to the optical flow field they can be visualized in the same way. The example demonstrated that the motion properties of the Lagrangian field depend on the integration parameter. The parameter is a crucial choice and related to the duration of the expected event, e.g. small for the boxing. The integration over a larger time span reduces the noise of the optical flow estimation and allows to recognize long term motion structures such as the dancing person at the bottom sample.

3 Lagrangian Scale Invariant Feature Transform

Comparative studies done by Xu *et al.* [22] and Nievas [15] show that for the task of violence detection the MoSIFT [5] algorithm outperforms common local features e.g. HoG, HOG, STIP. The MoSIFT algorithm is a derivate of the SIFT algorithm and combines the appearance model of the SIFT descriptor with a motion model obtained by two frame optical flow. As the direction Lagrangian field has the same structure and a similar interpretation as the optical flow field it can be integrated by substituting the two frame optical flow field. The proposed Lagrangian Scale Invariant Feature Transform (LaSIFT) will be related to the MoSIFT but differs in following aspects:

Camera motion has a significant impact to the motion signature of the directional field. Therefore we apply a motion compensation with a global motion model. We assume a homography based background motion model that excludes independently moving objects. The compensated direction field can be found by subtracting a background direction field from the actual field. The direction field of the background motion depends on the propagation of the motion vectors for all homographies in that image stack over τ. Due to the linearity of the path line integration and the homography the background direction field can be directly estimated in the Lagrangian domain. Therefore we regularly subsample the direction Lagrangian field, which provides us a set of motion-like vectors. With this set we robustly estimate the eight parameter homography and rectify the current directional field. Compared to the current field the rectified directional field suppresses the background camera motion and enhances the foreground moving objects.

The SIFT interest point detection is salient at blob like structures at multiple scales. SIFT interest points are scale invariant and all scales of an image must be considered. To estimate the motion for each scale the MoSIFT applies optical flow computation for each level of the image pyramids. Since optical flow computation is quite expensive, we propose to build a flow pyramid of the optical flow estimate from the lowest level of the pyramid instead. This increases the speed of the scale space motion estimation significantly. Similar to Chen *et al.* [5]

Method	ACC ± SD	AUC
SIFT + BoW	89.93±2.47%	0.9636
LaSIFT($\tau = 1$) + BoW	91.82±2.67%	0.9678
LaSIFT($\tau = 2$) + BoW	91.72±1.69%	0.9684
LaSIFT($\tau = 3$) + BoW	91.72±2.07%	0.9691
LaSIFT($\tau = 4$) + BoW	92.42±2.57%	0.9682
LaSIFT($\tau = 5$) + BoW	91.72±2.00%	0.9703
LaSIFT($\tau = 6$) + BoW	91.72±3.01%	0.9693
LaSIFT($\tau = 8$) + BoW	92.52±2.99%	**0.9734**
LaSIFT(mix) + BoW	***93.32±2.24%***	0.9732
HOG + BoW [8, 15]	91.7	-
HOF + BoW [8, 15]	88.6	-
MoSIFT + BoW [22]	90.9	-
MoSIFT + KDE + SC [22]	**94.0±1.97**%	96.66

Table 1: Comparison of violence detection performance on Hockey Fight dataset between LaSIFT on different temporal scales (top) and state-of-the-art (bottom). The SIFT denotes the evaluation of the appearance histogram only but within the proposed feature encoding.

we want to select distinctive interest points with sufficient motion. If the direction field vector of a candidate contains a minimal length, the algorithm will extract a interest point at this position and scale.

The feature descriptor consists of two parts: the SIFT grid based aggregated histogram of gradients and the aggregated histogram of direction field components from surrounding regions. Each histogram is aggregated from the neighborhood grouped into 4 × 4 regions. For each region an oriented histogram of eight bins is formed. We do not concatenate both descriptors as we want to integrate the dependency of appearance and motion in a late fusion manner. Experiments have shown that this late fusion outperforms the early fusion approach.

4 Feature Encoding

To encode the LaSIFT features we use the bag-of-words approach which have become popular for image and video understanding. The approach applies a codebook to create a finite set of representative descriptors for the overall dataset. Typically these are the cluster centers obtained from k-means clustering. These clusters are the vocabulary and known as visual words. A video sequence can than be represented by a histogram over a set of these words to generate a fixed dimensional encoding. In a learning phase, we train the vocabulary for the appearance and the motion model separately. In this study we use the histogram intersection kernel based clustering to create to the two codebooks [21]. Due to the hard assignment of the visual words to histogram bins, very frequently appearing individual words could dominate the classification results. To restrict their influence and enhance less frequent words we limit the value range of the histogram by twice the standard deviation added to the mean of data values. For classification we use a non linear support vector machine with a multi channel χ^2 kernel [23] defined

Method	ACC ± SD	AUC
SIFT + BoW	86.43 ±7.32%	0.9390
LaSIFT($\tau = 1$) + BoW	89.27±8.14%	0.9605
LaSIFT($\tau = 2$) + BoW	89.99±8.64%	**0.9741**
LaSIFT($\tau = 3$) + BoW	92.01±8.01%	0.9729
LaSIFT($\tau = 4$) + BoW	90.85±7.53%	0.9638
LaSIFT($\tau = 5$) + BoW	90.41±9.01%	0.9561
LaSIFT($\tau = 6$) + BoW	90.04±8.24%	0.9542
LaSIFT($\tau = 8$) + BoW	87.51±9.06%	0.9454
LaSIFT(mix) + BoW	**92.42±6.73**%	0.9696
HOG + BoW [8, 15]	57.43±0.37%	0.6182
HOF + BoW [8, 15]	58.53±0.32%	0.5760
LTP [8]	71.53±0.17%	0.7986
VIF [8]	81.30±0.21%	0.8500
MoSIFT + BoW [22]	83.42±8.03%	0.8751
MoSIFT + KDE + SC [22]	89.05±3.26%	0.9357

Table 2: Comparison of violence detection performance on Crowd Violence dataset between LaSIFT on different temporal scales (top) and state-of-the-art (bottom). The SIFT denotes the evaluation of the appearance histogram only but within the proposed feature encoding

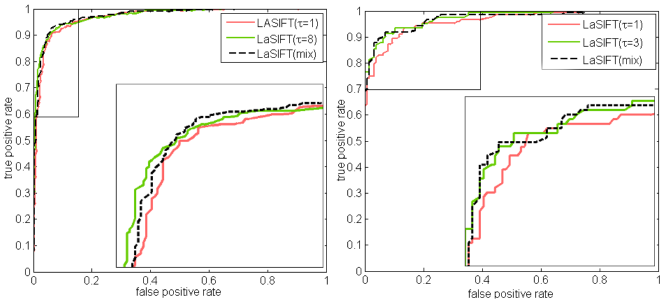

Figure 2: Violence classification ROC curves for the Hockey Fight (left) and Crowd Violence (right) dataset between LaSIFT short term ($\tau = 1$), the best long term and the combined temporal configuration.

by:

$$K(i,j) = \sum_{c \in C} \frac{1}{A^c} D_c(H_i^c, H_j^c) \quad (3)$$

where H_i^c and H_j^c are two histograms of the appearance or motion channel and $D_c(\ldots)$ is the χ^2 distance. The weighting parameter A^c is the mean value of the distances between all training samples of the channel c.

5 Experiments

We evaluated our method on two challenging datasets created for violent video detection:

Hockey Fight The Hockey Fight dataset created by Nievas *et al.* [15] contains 1000 short video clips from hockey games of the National Hockey League. The video sequences show fights of single persons mostly captured at a close distance. The dataset contains several difficulties such as camera motion, different point of views and the unknown number of involved subjects. Especially, Motion blur caused by the very fast body limbs is challenging for the motion estimation. Each video sequence consists of 50 frames with a resolution of 360×288 pixels and is manually labeled as fighting or not fighting.

Violent Crowd While the Hockey Fight dataset is constrained to an ice hockey arena, the violent crowd created by Hassner *et al.* [8] has various number of arenas. The dataset is collected from YouTube and presents a wide range of video qualities and surveillance scenarios. The dataset consists of 246 short video sequences with a resolution of 320 pixels and labeled manually as violent or not violent. The dataset contains static and non static video sequences that capture a wide variety of crowd fights in various contexts such as football stadium, bars or open areas. Compression artifacts, motion blur, text overlay, flash lights and different temporal resolution make it very challenging to extract accurate motion estimation.

In the following experiments our main focus will be the impact of the integration parameter τ. To construct the bag-of-words vocabulary we use 500 visual words. In contrast to [15], we found the χ^2 kernel and the late fusion to perform slightly better that the histogram intersection kernel and the early fusion [5]. To illustrate the impact of the proposed directional field to the overall descriptor, we evaluated the appearance component of the LaSIFT in addition. This experiment is labeled as (SIFT + BoW). For both datasets the benchmarks are a five fold cross validation classification test. While the Crowd Violence dataset still provides a splitting into mutually scene exclusive sets, the Hockey Fight dataset was divided into five sets each containing 100 consecutive clips.

The results for the Hockey Fight benchmark are shown in Table 1. We studied integration intervals from $\tau = 1$ up to $\tau = 8$ where $\tau = 1$ considers the two frame optical flow and we studied the combination the descriptors for each temporal scale labeled as mix. In general we can see an improved accuracy for the LaSIFT compared to the baseline MoSIFT approach. Comparing the integration intervals $\tau = 1$ with the MoSIFT shows a significant improvement of our method. The improvement is cause especially by the camera stabilization and the more robust histogram intersection kernel clustering. We observed that each temporal scale contains a specific motion component, which improves the final results. In the Hockey Fight dataset the area under the ROC curve (AUC) measure is increasing with larger temporal scales. That is surprising as we expected the punches of the players, that are short term events, to be significant in that dataset. It appears that the optical flow estimation and thereby an accurate path line advection of the punching is hard to assure. The significant motion characteristic seams to be the motion of the torso. The LaSIFT outperforms current state of the art methods in terms of AUC. However, the performance in terms of accuracy is less than the improved feature coding scheme proposed by Xu *et al.*.

Table 2 gives the results for the Violent Crowd dataset. Interestingly, there is a maximal AUC and accuracy value around $\tau = 3$ i.e. when around three optical flow fields are considering. Unfortunately, it is very hard to name a specific motion signature like punches or the wrangling of the crowd as the important motion feature as the data is very complex and the sequences are too divers. Figure 3 gives some examples of true and false classified sequences of the performed experiments for both datasets. These sequences share the same results for the different temporal scales τ.

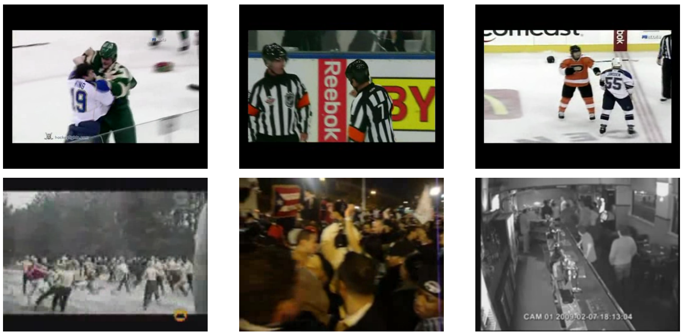
true positive false positive false negative

Figure 3: Examples of classification results that have the same results for different temporal scales τ.

But the comparison with the SIFT appearance component of the LaSIFT descriptor show that the video dataset contains common significant motion features that have been extracted by the LaSIFT descriptor. Furthermore, the proposed classification algorithm based on the LaSIFT feature outperforms state-of-the-art methods in terms of accuracy and AUC measures. Finally, the overall results of the evaluation have been shown that the LaSIFT feature is a valuable feature for violent video classification.

6 Conclusion

In this work we utilized Lagrangian measures to detect violent video footage. Lagrangian measures are a part of Lagrangian theory for time dependent vector fields. To describe the properties of moving structures we focus on Lagrangian Coherent Structures where we found the directional Lagrangian field a promising feature representation. This Lagrangian field characterizes the motion information over a time interval τ. We proposed a local feature that extends the SIFT algorithm and implements the Lagrangian field to encode the spatio temporal characteristic of a position in space time. In our experiments we studied the crucial aspect of the parameter τ. The parameter is strongly related to the duration of the motion event that should be encoded by the local features and determines the overall classification results. We evaluated our algorithm with a bag-of-words approach and showed significant improvements over the state-of-the-art on current violent detection datasets, i.e. Crowd Violence, Hockey Fight.

We have show that the LaSIFT algorithm outperforms current local features not only for the crowded scenes as seen in the Crowd Violence dataset but also for video sequences with individuals acting e.g. shown by the Hockey Fight dataset. As violent detection is a subclass of action recognition, in the future we want to extend the field of application and apply the LaSIFT algorithm to the more general field of action recognition.

Acknowledgements

The research leading to these results has received funding from the European Communitys FP7 under grant agreement number 607480 (LASIE).

References

[1] Esra Acar, Frank Hopfgartner, and Sahin Albayrak. Violence detection in hollywood movies by the fusion of visual and mid-level audio cues. In *International Conference on Multimedia*, pages 717–720, 2013.

[2] Esra Acar, Tobias Senst, Alexander Kuhn, Ivo Keller, Holger Theisel, Sahin Albayrak, and Thomas Sikora. Human action recognition using lagrangian descriptors. In *International Workshop on Multimedia Signal Processing*, pages 360–365, 2012.

[3] Saad Ali and Mubarak Shah. A Lagrangian Particle Dynamics Approach for Crowd Flow Segmentation and Stability Analysis. In *Computer Vision and Pattern Recognition*, pages 1–6, 2007.

[4] Liang-Hua Chen, Hsi-Wen Hsu, Li-Yun Wang, and Chih-Wen Su. Violence detection in movies. In *International Conference on Computer Graphics, Imaging and Visualization*, pages 119–124, Aug 2011.

[5] Ming-yu Chen and Alex Hauptmann. MoSIFT : Recognizing Human Actions in Surveillance Videos. *Technical Report CMU-CS-09-161*, pages 1–16, 2009.

[6] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *European Conference on Computer Vision*, volume 3952, pages 428–441, 2006.

[7] Oscar Déniz, Ismael Serrano, Gloria Bueno, and Tae-Kyun Kim. Fast Violence Detection in Video. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 478–485, 2014.

[8] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. *Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6, 2012.

[9] Alexander Kuhn, Tobias Senst, Ivo Keller, Thomas Sikora, and Holger Theisel. A lagrangian framework for video analytics. In *Workshop on Multimedia Signal Processing*, pages 387–392, 2012.

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

[11] Teng Li, Huan Chang, Meng Wang, Bingbing Ni, and Richang Hong. Crowded Scene Analysis : A Survey. *Transactions on Circuits and Systems for Video Technology*, 25(3):367–386, 2015.

[12] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[13] Ramin Mehran, Brian E. Moore, and Mubarak Shah. A streakline representation of flow in crowded scenes. In *European Conference on Computer Vision*, volume 6313, pages 439–452, 2010.

[14] Brian E. Moore, Saad Ali, Ramin Mehran, and Mubarak Shah. Visual crowd surveillance through a hydrodynamics lens. *Communications of the ACM*, 54:64–73, 2011.

[15] Enrique Bermejo Nievas, Oscar Déniz Suarez, Gloria Bueno García, and Rahul Sukthankar. Violence detection in video using computer vision techniques. *Computer Analysis of Images and Patterns*, pages 332–339, 2011.

[16] Thomas Peacock and John Dabiri. Introduction to focus issue: Lagrangian coherent structures. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 20(1):–, 2010.

[17] Ronald Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010.

[18] Tobias Senst, Alexander Kuhn, Holger Theisel, and Thomas Sikora. Detecting people carrying objects utilizing lagrangian dynamics. In *International Conference on Advanced Video and Signal-Based Surveillance*, pages 398–403, 2012.

[19] Mats Sjöberg, Bogdan Ionescu, Jiang Yu-Gang, Vu Lam Quang, Markus Schedl, and Claire-Hélène Demarty. The MediaEval 2014 Affect Task: Violent Scenes Detection. In *Working Notes Proceedings of the MediaEval 2014 Workshop*, 2014.

[20] Heng Wang and Cordelia Schmid. Action Recognition with Improved Trajectories. In *International Conference on Computer Vision*, pages 3551–3558, 2013.

[21] Jianxin Wu, Wc Tan, and Jm Rehg. Efficient and effective visual codebook generation using additive kernels. *The Journal of Machine Learning Research*, 12:3097–3118, 2011.

[22] Long Xu, Chen Gong, Jie Yang, Qiang Wu, and Lixiu Yao. Violent video detection based on MoSIFT feature and sparse coding. In *International Conference on Acoustics, Speech and Signal Processing*, pages 3562–3566, 2014.

[23] Jianguo Zhang, Marcin Marszałek, Svetlana Lazebnik, and Cordelia Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73:213–238, 2007.