Real Time Violence Detection in Video with ViF and Horn-Schunck

Vicente Machaca Arceda Universidad Nacional de San Agustín Arequipa, Perú
enriquefirst@gmail.com.
Karla Fernández Fabián Universidad Nacional de San Agustín Arequipa, Perú
karla.m.f.f@gmail.com
Juan Carlos Gutíerrez Universidad Nacional de San Agustín Arequipa, Perú
jcgutierrezc@gmail.com

Abstract—We evaluated the performance of ViF in violence datasets varying the optic flow algorithm, we used Iterative Reweighted Least Squares (IRLS), Horn-Schunck and Lucas-Kanade. As datasets, a Crowded and Hockey datasets were used, also we built a new dataset with videos taken from surveillance cameras, we named it Surveillance Videos (SV). In Hockey dataset Horn-Schunck had a better performance, but in SV and Crowded datasets IRLS was better. We also evaluated the computational cost processing two frames, Horn-Schunck had a very low cost of 0.25 seconds against 7.80 seconds of IRLS so we could use it in real time environments.

Keywords—SIFT, ViF, STIP, MoSIFT.

Digital Object Identifier
(DOI):http://dx.doi.org/10.18687/LACCEI2016.1.1.122
ISBN: 978-0-9822896-9-3
ISSN: 2414-6390

Real Time Violence Detection in Video with ViF and Horn-Schunck

Vicente Machaca Arceda
Universidad Nacional de San Agustín
Arequipa, Perú
enriquefirst@gmail.com

Karla Fernández Fabián
Universidad Nacional de San Agustín
Arequipa, Perú
karla.m.f.f@gmail.com

Juan Carlos Gutíerrez
Universidad Nacional de San Agustín
Arequipa, Perú
jcgutierrezc@gmail.com

***Abstract*—We evaluated the performance of ViF in violence datasets varying the optic flow algorithm, we used Iterative Reweighted Least Squares (IRLS), Horn-Schunck and Lucas-Kanade. As datasets, a Crowded and Hockey datasets were used, also we built a new dataset with videos taken from surveillance cameras, we named it Surveillance Videos (SV). In Hockey dataset Horn-Schunck had a better performance, but in SV and Crowded datasets IRLS was better. We also evaluated the computational cost processing two frames, Horn-Schunck had a very low cost of 0.25 seconds against 7.80 seconds of IRLS so we could use it in real time environments.**

***Keywords*—SIFT, ViF, STIP, MoSIFT.**

I. Introduction

The United Nations Office on Drugs and Crime (UNODC) has a site called Global Study on Homicide where they show the rate of homicides per 100,000 inhabitants, with 16.3 rate in America against 3.0 in Europe, in this context we have a big problem. Morover according to the Institute of Legal Defense (ILD-Peru) in 2015 the Peruvian people consider crime and insecurity as their mayor problem [37], the National Institute of Statistic and Informatic (NISI-Peru)'s technical report of security in Mar-2015 said that 30,5% of people was the victim of a criminal act[1]

For all these problems there are a lot of surveillance camera services, these systems can be easily implemented in order to monitor any stage, but it could be ineffective due to the lack of trained people who supervise the recording and the natural ability to pay attention [24].

Having support systems in real time to detect possible serious violent actions are very useful in controlling public safety. In addition, detecting a violent action is challenging due to the definition of "violence" and the high computational cost involved. The definition of "violence" varies among different researches in the state of art, ranging from the detection of fire, explosions, blood, fighting, etc. This work is based on statistics of change in the magnitude of the optical flow vectors[2] [19], this usually occurs when there is an abrupt change in a video sequence such as fighting, theft, accidents, etc. The detection will be also in real time, we hope to get a system, in future works, that support surveillance cameras controlling some criminal events. This work focus on getting a method with the minor computational cost and acceptable accuracy.

II. Related Work

Detection of violent actions is a particular problem within a larger that is the recognition of actions, these last are resolved using the same approach as visual categorization [8], they used a Harris detector [32] to get key points and Scale Invariant Feature Transform (SIFT) as descriptor, then they used Bag of Visual Words (BoVW) to get mid-level features. Space-time Interest Point (STIP) was used in [14] to recognize facial expressions, human activities and a mouse's behavior, getting 83%, 80% and 72% of accuracy respectively. In [45] Gaussian Difference [30] is used with Principal Component Analysis - Scale Invariant Feature Transform (PCA-SIFT) [23] and BoVW to classify video scenes, concluding that the size of the vocabulary used in BoVW depends heavily on the complexity of scenes classified. Most studies use BoVW, then [42] presented a BoVW comparison varying the descriptors. In [40] descriptors as Histogram of Optical Flow (HOF) and Histogram of Oriented Gradient (HOG) with variations in optical flow are evaluated using Lucas-Kanade [31], Horn-Schunck [21], and Farnebäck [15] as optical flow algorithms, they also evaluated the performance of BoVW comparing K-means against Random Forests [6] and Fisher kernel [36], they concluded that Lucas-Kanade and Horn-Schunck outperformed Farnebäck and Fisher kernel outperformed K-means.

One of the first works detecting violence is based on audio presented by [16] defined violence as those events containing shots, explosions, fights and screams, whereas nonviolent content corresponds to audio segments containing music and speech. The descriptors used were: energy entropy, short-time energy, zero crossing rate (ZCR), spectral flux, and roll-off with a polynomial Support Vector Machine (SVM) as the classifier getting 85.5% of accuracy. Bag of Audio Words (BoAW) also are used to get mid-level features, [13] used Mel-Frequency Cepstral Coefficients (MFCC) as audio descriptor and dynamic Bayesian networks. The main contribution of

[1] We consider the criminal act as an event that threatens the security, violates the rights of a person and leads to danger, harm or risk [22].

[2] The optical flow can be defined as the apparent movement of intensity patterns in an image. The word *apparent* indicates that the motion of objects in the space (range of motion) may coincide with the estimated flow. However, in situations in which the movement of objects implies a movement of intensity patterns in the image plane, the optical flow may be directly related to the movement of objects in the scene [33]

Digital Object Identifier (DOI): http://dx.doi.org/10.18687/LACCEI2016.1.1.122
ISBN: 978-0-9822896-9-3
ISSN: 2414-6390

this work is when using BoAW the noise produced by video segmentation is remove.

Another definition of violence as scenes those containing fights, regardless of the context and the number of people involved is used in the work of [9], they proposed Bag of Visual Words (BoVW) with Space-Time Interest Point (STIP), based on Laptev's research [26], as descriptor, they compared the performance of STIP-based BoVW with SIFT-based BoVW. Here STIP achieved a better result. A variation in STIP named Hue Space-Time Interest Points (HueSTIP) proposed by [39] take in count pixel colors, in this case they recognized general actions, for detecting fights HueSTIP outperforms STIP but with a higher computational cost.

Motion Scale-Invariant Feature Transform (MoSIFT) is used by [4] (it was proposed by [48]), to detect fights, they compared MoSIFT and STIP with BoVW and SVM as the classifier. In the experiments they used two datasets: Movies and Hockey games, in Hockey dataset STIP got a 91.7% of accuracy against 90.9% of MoSIFT, but in Movie dataset MoSIFT outperforms STIP with 89.5% of accuracy against 44.5% of STIP. In this context, we cannot decide which descriptor is better, but we can infer that both require a high computational cost doing it difficult to use in real time.

A real time model is presented in [19], here they detect violence in crowded scenes. They define "violence" as sudden changes in motion in a video footage. Their model basically considers statistics of magnitude changes of flow vectors over time, this es named Violent Flow (ViF). They also introduced a new dataset of crowded scenes. In the results ViF outperforms Local Trinary Patterns (LPT) [47], histogram of oriented gradient (HoG) [27], histogram of oriented optical flow (HoF) [27] and histogram of oriented gradient and optical flow(HNF)[27]. The model is also evaluated in other datasets, as Hockey [4] and ASLAN [25] to evaluated the ViF's performance in action recognition, here ViF outperforms STIP while with larger vocabularies, STIP outperforms ViF. The good thing to mention about this new descriptor is that it is one of the fastest enabling its use in real time.

MoSIFT is also used in [43] with characteristics based on Kernel Density Estimation (KDE) to improve efficiency, also instead of using BoVW they used Sparse coding, then they compared their proposal with HOG [27], HOF [27], HNF [27] and ViF [19] outperforming them in the Crowded and Hockey datasets.

Other work based in optical flow is presented in [46] where in addition to detect violent scenes it locates in what part of the scene occurred the violence, Gaussian Mixed Model is extended to the domain of optical flow to detect regions that may contain violent actions in each region, Histogram of Optical Flow Orientation HOFO is used as descriptor.

Recently [11] proposed a model inspired in psychology which suggests that the kinematic characteristics are discriminating for specific actions, they named it "Extreme Acceleration". In the work of [5], they concluded that the kinematic patterns are sufficient for the perception of actions, and this idea was validated in the research of [34], more specifically studies in this field show that simple kinematic characteristics like speed and acceleration are correlated to emotional attributes [20], thereby detecting the change in acceleration is based on the blur of the image when motion occurs, by calculating the spectral power as evidenced [3]. The results were evaluated in the Movies and Hockey [4] datasets. As a result, the new proposal outperformed STIP and MoSIFT as well as being 15 times faster. This new approach has a very low computational cost, enabling use in real time.

In the case of detecting horror in movies, [41] used Multiple Instance Learning (MIL; MI-SVM [2]) using color and texture and visual features and MFCC as audio features. From the results it is concluded that the audio features to this context, are most relevant.

In [17] the work of [16] is extended where they used a multimodal two-stage approach, in the first step, they perform audio and visual analysis of the segments of one-second duration. In the audio analysis part, audio features such as energy entropy, ZCR, and MFCC are extracted and the mean and standard deviation of these features are used to classify scenes into one of seven classes (shots, fights, screams, etc.) In the visual analysis part, average motion, motion variance, and average motion of individuals in a scene are used to classify segments as having either high or low activity. The results obtained in this first step are then used to train a k-NN classifier. This method was evaluated in a movie dataset where they concluded that audio features are more relevant.

A three-stage method is proposed in [18], they used a semi-supervised cross-feature learning algorithm [44], in the first stage they use audio-visual features such as motion activity, ZCR, MFCC, then in a second stage features as screams, shots and explosions are detected with a SVM as the classifier, in the last stage, the result of previous stages are lineally weighted for the classification. This work was evaluated only in action movies with probably a poor performance in other contexts.

In the work of [28] two classifiers are used in co-training. They used mid-level features with BoAW on MFCC, spectrum flux and ZCR, in the visual classification they detected motion intensity, the (non-)existence of flame, explosion, and blood. They considered fights, explosions, murders and shots as violence concept. As other multimodal methods they evaluated their results just in movies.

In [7] used the same concept of violence that [28] where violence is any action scene with blood, they used average motion, camera motion, and average shot length are used for scene representation and SVM as the classifier, then they used the "Viola-Jones" face detector, to detect faces and blood near. They outperform the work of [28] but they just used a movie dataset where we have good conditions.

The next paragraphs consider the same concept of violence adopted in "MediaEval 2013 VSD task" (objective and subjective definition [3]). [35] used temporal information and multimodal evaluating their results in Bayesian Networks, they also used the "MediaEval 2011 VSD task" dataset. They demonstrated that both multimodality and temporality add valuable information into the system and improve the performance in terms of MediaEval cost function [10], in addition, we have to mentions that the MediaEval 2013 dataset is a collection of movies where the conditions as illumination, resolution, etc. are ideal.

The dependencies between audio and visual features are studied in [12], They combined the audio and the visual features and then determined statistically joint multimodal patterns using audio-visual BoW, they also used the MediaEval 2013 dataset. They outperformed the majority of methods using the audio and visual features separately.

Recently [1] have proposed the use of audio and visual features also, as audio feature they use MFCC and for visual features they use HOF, ViF and color descriptors, they also evaluated their results in the MediaEval 2014 dataset. They concluded that the audio features are more relevant than the visual features, they also combined both features getting even betters results.

The used of Lagrangian theory show the applicability for video analysis in several aspects. In this context [38] utilized the concept of Lagrangian measures to detect violent scenes. They proposed a local feature based on the SIFT algorithm that incorporates appearance and Lagrangian based motion models, they named it as LaSIFT. They compared their results with HOG, HOF and MoSIFT in the Crowded and Hockey datasets. In the case of Hockey dataset, the LaSIFT feature 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 [43]. For Crowded dataset the LaSIFT feature outperforms state-of-the-art methods in terms of accuracy and AUC measures. LaSIFT seems to be very promising, but the authors didn't mention the computational cost, we could consider that by the used of BoVW it could have a high cost, a comparison of it with ViF in terms of accuracy and cost could be interesting.

III. The Violence Detection Method

A. Evaluating different optical flow algorithms

ViF consider the statistics of magnitude changes of flow vectors over time as we see in Figure 1, In order to get these vectors [19] used the optical flow algorithm proposed by [29] named Iterative Reweighted Least Squares (IRLS), but nowadays we have a lot of different optical flow algorithms, in this context, we propose to evaluate the ViF's performance with Lucas-Kanade [31] and Horn-Schunck [21] as optical flow algorithm in the same way as [40] did it, evaluating different optical flow algorithms in HOF to detect behaviors in video. We are going to evaluate the accuracy and the computational cost, so in the future, it will be used in real time. In this work we are not going to use any pre-processing step.

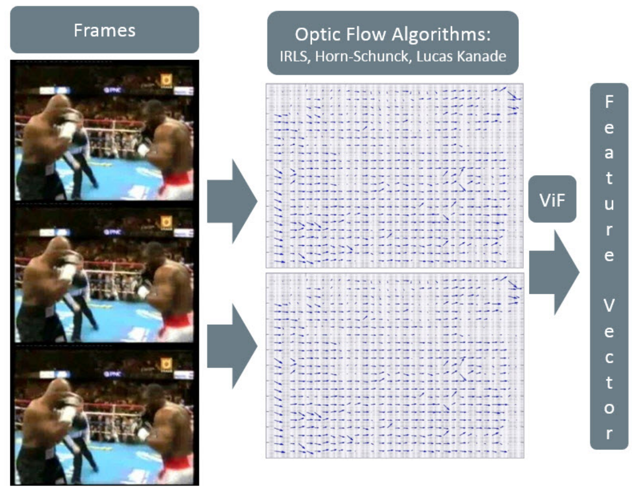

Figure 1: ViF descriptor in video.

B. Optical flow algorithm

ViF depends heavily on the magnitude of optical flow vectors, these vectors are calculated for each pixel in two consecutive frames as we see in Figure 1, these vectors could represent the motion of objects in a video scene, where the bigger vectors represent the objects with more movement, in Figure 2 we see two consecutive frames, and in Figure 3 we see the optical flow vectors computed. Actually there is a lot of different algorithms to get these vectors, in this work we evaluated the performance of Horn-Schunck, Lucas-Kanade and IRLS.

C. ViF descriptor

The ViF descriptor is presented in algorithm 1, here we get a binary, magnitude-change, significance map b_t for each frame f_t. Then we get a mean magnitude-change map, for each pixel, over all the frames with the equation 1:

$$b_{x,y} = (1/T) \sum_t b_{x,y,t} \tag{1}$$

Then the ViF descriptor is a vector of frequencies of quantized values $b_{x,y}$. For more details you could see the work of [19].

D. Subsampling video frames

Subsequent video frames could contain the same information. As the time for descriptor extraction is the largest bottleneck in this work, we sample every 3 frames the video.

[3]**Subtask 1**: objective definition The previous definition from 2012: Violence is defined as "physical violence or accident resulting in human injury or pain". **Subtask 2:** subjective definition For this subtask, the targeted violent segments are those "one would not let an 8 years old child see in a movie because they contain physical violence".

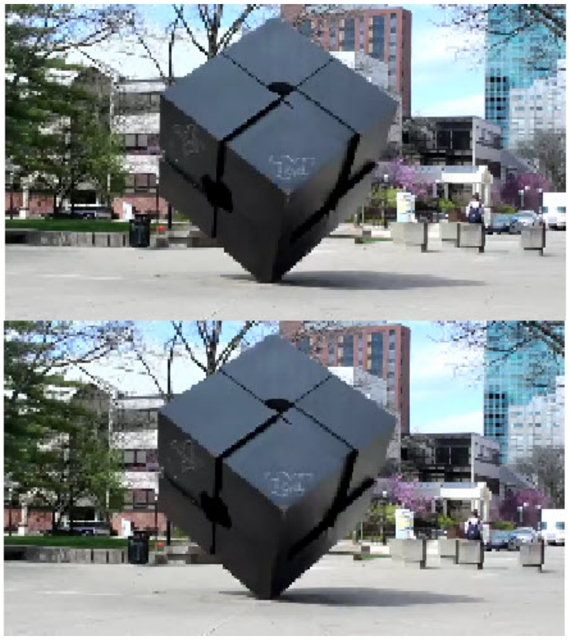

Figure 2: Two consecutive frames.
Source: Matlab examples.

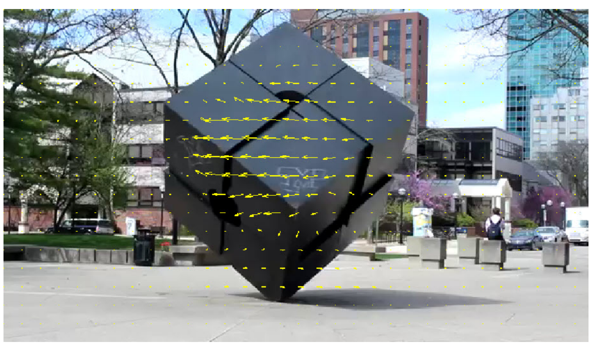

Figure 3: Optical flow vectors get by Lucas-Kanade algorithm.
Source: Matlab examples.

E. Classification

SVM is used as a classifier with a lineal kernel, taking as input the result of ViF descriptor (feature vector with 336 values). In the experiments we use cross-validation with k=10. In Figure 4 we can see the architect of the whole model.

IV. Experiment and Results

A. Datasets

We evaluated the performance in the Hockey [4] and Crowded [19] datasets, some frames are shown in Figures 5 and 6 respectively. In addition, we built a new dataset with videos containing fights from surveillance cameras, these videos are in real conditions as we can see in Figure 7, we named it Surveillance Videos (SV) dataset. In Table I we can see a comparison of the three datasets, we have to mention that actually the changeling dataset for violent detection is the Hockey, it's because it is so difficult to distinguish a fight in this game.

Data: S = Sequence of gray scale images.
Each image in S is denoted as $f_{x,y,t}$, where
$x = 1, 2, ..., N$, $y = 1, 2, ..., M$ and $t = 1, 2, ..., T$.
Result: Histogram($b_{x,y}$; n_bins = 336)
for $t = 1$ *to* T **do**
1. Get optical flow $(u_{x,y,t}, v_{x,y,t})$ of each pixel $p_{x,y,t}$ where t is the frame index.
2. Get magnitude vector: $m_{x,y,t} = \sqrt{u_{x,y,t}^2 + v_{x,y,t}^2}$
3. For each pixel we get:
$$b_{x,y,t} = \begin{cases} 1 & \text{if } |m_{x,y,t} - m_{x,y,t-1}| >= \theta \\ 0 & \text{other case} \end{cases}$$
where θ is a threshold adaptively set in each frame to the average value of $|m_{x,y,t} - m_{x,y,t-1}|$.

end

Algorithm 1: ViF descriptor

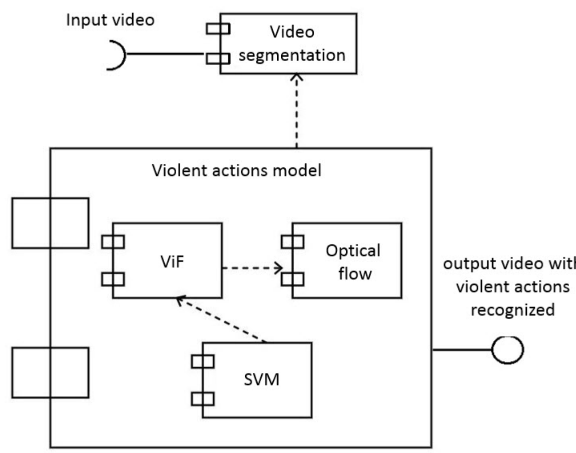

Figure 4: Model architect.

	Resolution	Framerate per second	Duration (seconds)	Number of videos
SV	480 x 360	25	2	100
Hockey	360 x 288	25	2	1000
Crowded	320 x 240	25	4	246

Table I: Datasets features.

B. Results

We evaluate the performance of ViF in a SVM classifier with a linear kernel and cross-validation (k=10). In Table II we can see the Accuracy (ACC) and Standard Deviation (SD) of the classifier, we also have included the Area Under the Curve (AUC) of the best model. As we can see for the SV and Hockey datasets, we get better results using the IRLS algorithm, but in the case of Hockey dataset we get better result with Horn-Schunck as optic flow algorithm.

The Receiver Operating Characteristic (ROC) of the classifier with ViF using IRLS as optic flow algorithm is shown

Figure 5: Some frames taken from the Hockey dataset.

Figure 6: Some frames taken from the Crowded dataset.

in Figure 8. Moreover the ROC curves for Lucas-Kanade and Horn-Schunck are shown in Figures 9 and 10 respectively.

We also evaluated the ViF performance with the three datasets together, in this case we take randomly 200 videos from the Hockey dataset, 200 from Crowded and 100 from

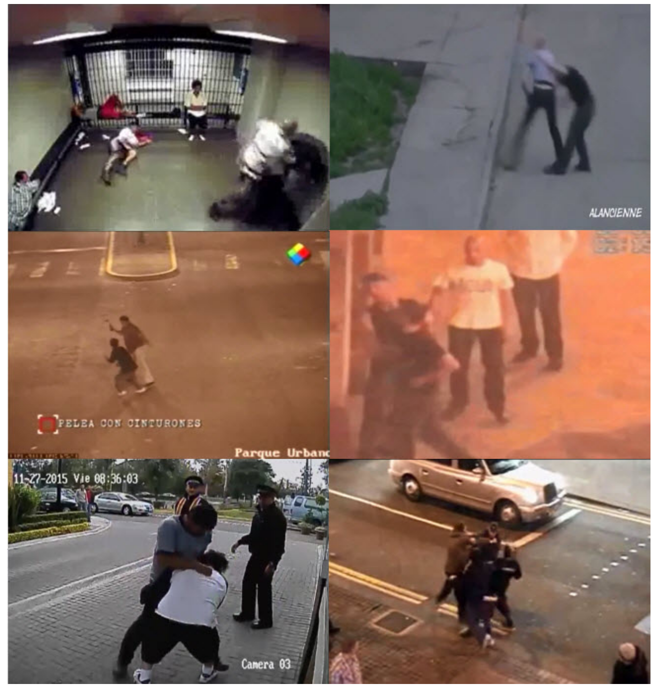

Figure 7: Some frames taken from the SV dataset.

ViF with IRLS		
Dataset	ACC ± SD	AUC
SV	0.7400 ± 0.1265	**0.9000**
Hockey	0.7190 ± 0.0848	0.8000
Crowded	0.7881 ± 0.1429	**0.9583**
ViF with Lucas-Kanade		
Dataset	ACC ± SD	AUC
SV	0.6300 ± 0.1494	0.8000
Hockey	0.6220 ± 0.0894	0.7100
Crowded	0.6614 ± 0.1022	0.8397
ViF with Horn-Schunck		
Dataset	ACC ± SD	AUC
SV	0.5900 ± 0.1524	0.8000
Hockey	0.7980 ± 0.0349	**0.8400**
Crowded	0.7375 ± 0.1092	0.8782

Table II: The performance of ViF with different optic flow algorithms. The Accuracy (ACC) and Standard Deviation (SD) of the classifier were evaluated by cross-validation (k=10) and also the Area Under the Curve (AUC) of the best model is included.

SV, the result is shown in Table III and the ROC curve in 11. In this case we see that the IRLS algorithm works well in these surveillance datasets, in second place is Horn-Schunck and then Lucas-Kanade. In addition, the accuracy could be improved by adjusting the SVM's kernel and parameters but we didn't focus on that. We have to mention that all the datasets are videos in real conditions with poor resolution and noisy, and also the videos from Crowded and SV datasets were taken from surveillance cameras with really poor conditions you could have ever seen. We focus on these videos because of the future applications of the method in security cameras.

A comparison of the computational cost of the different

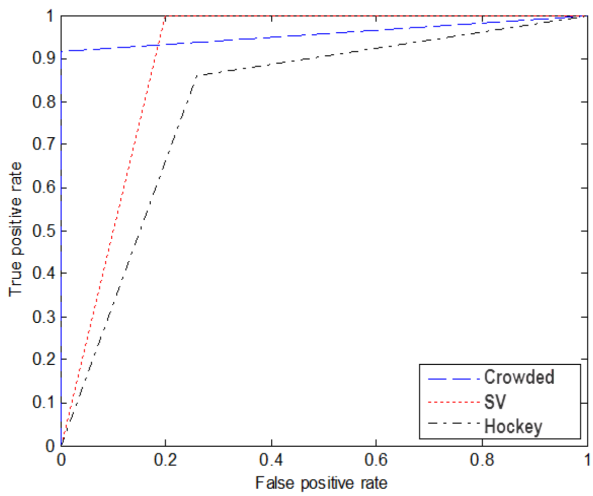

Figure 8: ROC curve of a SVM classifier with ViF and IRLS as optic flow algorithm in the SV, Hockey and Crowded datasets.

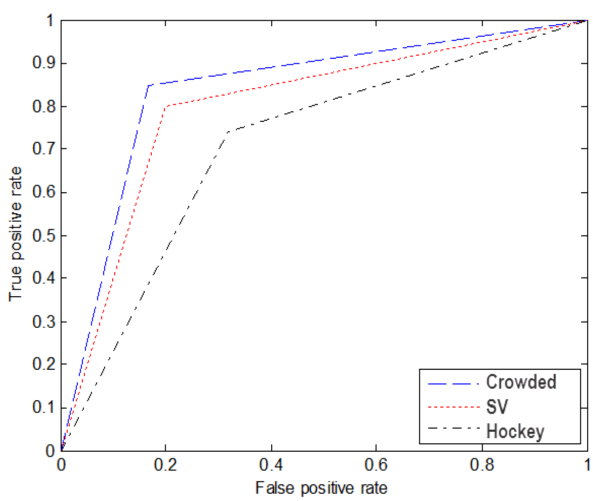

Figure 9: ROC curve of a SVM classifier with ViF and Lucas-Kanade as optic flow algorithm in the SV, Hockey and Crowded datasets.

Optic Flow	ACC $\pm$ SD	AUC
IRLS	0.7140 $\pm$ 0.0737	**0.8400**
Horn-Schunck	0.7120 $\pm$ 0.0391	0.7800
Lucas-Kanade	0.5680 $\pm$ 0.0444	0.6283

Table III: The performance of ViF with different optic flow algorithms with the three datasets together. The Accuracy (ACC) and Standard Deviation (SD) of the classifier were evaluated by cross-validation (k=10) and also the Area Under the Curve (AUC) of the best model is included.

optical flow algorithms evaluated by processing two frames is shown in Figure 12, unlike Lucas-Kanade and IRLS, Horn-

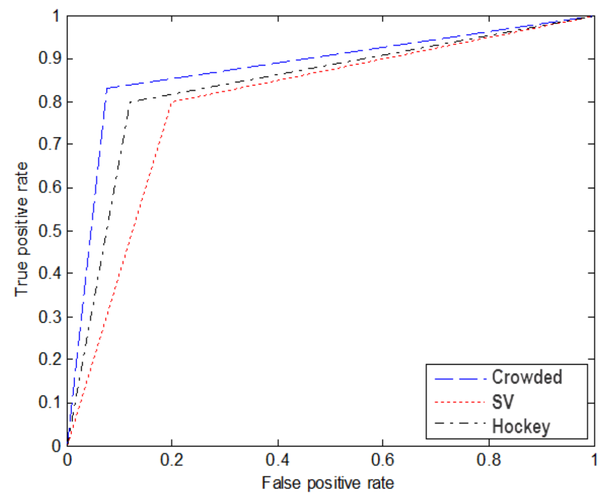

Figure 10: ROC curve of a SVM classifier with ViF and Horn-Schunck as optic flow algorithm in the SV, Hockey and Crowded datasets.

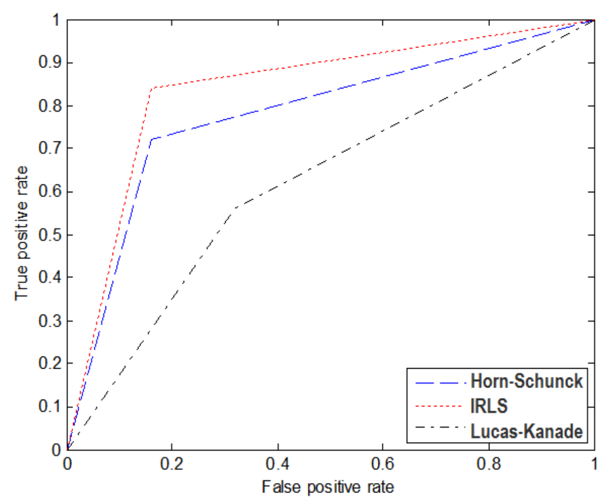

Figure 11: ROC curve of a SVM classifier with the joined dataset (Crowded, Hockey and SV). We compared the IRLS, Lucas-Kanade and Horn-Schunck optic flow algorithms in ViF descriptor.

Schunck presents a low cost, enabling its use in real time. The measurement was evaluated in a computer with a 1.8 GHz processor.

V. Conclusions

In this study, we sought to improve ViF using different optical flow algorithms as IRLS, Horn-Schunck and Lucas-Kanade, their performance in different datasets were evaluated. This evaluation concluded that the ViF's accuracy with the

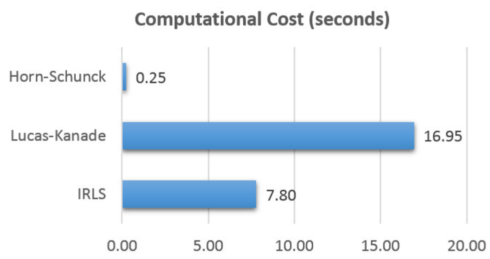

Figure 12: Comparison of computational cost of IRLS, Horn-Schunck and Lucas-Kanade.

IRLS optic flow algorithm had better results, but in the case of Hockey dataset ViF's with Horn-Schunck was better.

We also joined all the datasets and evaluates the ViF's performance, here IRLS outperformed the others. In this case we have to mention that we just took 200 videos of 1000 from Hockey dataset in order to have a balance dataset. In conclusion to have better results we need a Hockey-sized comparable dataset for a more accurate comparison.

On the other hand the computational cost of the optical flow algorithms was evaluated, the top performer was Horn-Schunck with only 0.25 seconds to process two frames, compared to 16.95 and 7.80 seconds of Lucas-Kanade and IRLS respectively.

Thus use ViF with Horn-Schunck is highly acceptable due to its low computational cost and have better results for certain datasets such as Hockey enabling its use in real time.

VI. Future Work

We planned to use the proposed method in surveillance cameras, the main goal is to have a method that work in real time, so we could alert the police officers if a criminal or violent act occurs, in this context we need a real surveillance videos that actually we are collecting in our SV dataset.

References

[1] E. Acar, M. Irrgang, D. Maniry, and F. Hopfgartner, "Detecting violent content in hollywood movies and user-generated videos," in *Smart Information Systems*, ser. Advances in Computer Vision and Pattern Recognition, F. Hopfgartner, Ed. Springer International Publishing, 2015, pp. 291–314. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-14178-7_11

[2] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in neural information processing systems*, 2002, pp. 561–568.

[3] H. B. Barlow and B. A. Olshausen, "Convergent evidence for the visual analysis of optic flow through anisotropic attenuation of high spatial frequencies," *Journal of Vision*, vol. 4, no. 6, p. 1, 2004.

[4] E. Bermejo, O. Deniz, G. Bueno, and R. Sukthankar, "Violence detection in video using computer vision techniques," *14th International Conference, CAIP 2011*, pp. 332–339, Ago 2011.

[5] R. Blake and M. Shiffrar, "Perception of human motion," *Annu. Rev. Psychol.*, vol. 58, pp. 47–73, 2007.

[6] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

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

[8] G. Csurkaa, L. F. C. R. Dance, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," *Workshop on Statistical Learning in Computer Vision*, pp. 1–22, 2004.

[9] F. D. M. de Souza, G. C. Chávez, E. A. do Valle Jr., and A. de A. Araújo, "Violence detection in video using spatio-temporal features," *Conference on Graphics, Patterns and Images (SIBGRAPI), 2010 23rd SIBGRAPI*, pp. 224 – 230, Ago 2010.

[10] C.-H. Demarty, C. Penet, M. Schedl, I. Bogdan, V. L. Quang, and Y.-G. Jiang, "The mediaeval 2013 affect task: violent scenes detection," in *MediaEval 2013 Working Notes*, 2013, p. 2.

[11] O. Deniz, I. Serrano, G. Bueno1, and T.-K. Kim, "Fast violence detection in video," *VISAPP 2014 - Proceedings of the 9th International Conference on Computer Vision Theory and Applications, Volume 2, Lisbon, Portugal, 5-8 January, 2014*, pp. 478–485, Oct 2014.

[12] N. Derbas and G. Quénot, "Joint audio-visual words for violent scenes detection in movies," in *Proceedings of International Conference on Multimedia Retrieval*. ACM, 2014, p. 483.

[13] B. do Nascimento Teixeira, "Mtm at mediaeval 2014 violence detection task," *MediaEval 2014, Multimedia Benchmark Workshop*, Oct 2014.

[14] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint*, pp. 65–72, Oct 2005.

[15] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*, ser. Lecture Notes in Computer Science, J. Bigun and T. Gustavsson, Eds. Springer Berlin Heidelberg, 2003, vol. 2749, pp. 363–370. [Online]. Available: http://dx.doi.org/10.1007/3-540-45103-X_50

[16] T. Giannakopoulos, D. I. Kosmopoulos, A. Aristidou, and S. Theodoridis, "Violence content classification using audio features," *ECCV 2004 workshop Applications of Computer Vision*, p. 502–507, May 2006.

[17] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis, "Audio-visual fusion for detecting violent scenes in videos," in *Artificial Intelligence: Theories, Models and Applications*, ser. Lecture Notes in Computer Science, S. Konstantopoulos, S. Perantonis, V. Karkaletsis, C. Spyropoulos, and G. Vouros, Eds. Springer Berlin Heidelberg, 2010, vol. 6040, pp. 91–100. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-12842-4_13

[18] Y. Gong, W. Wang, S. Jiang, Q. Huang, and W. Gao, "Detecting violent scenes in movies by auditory and visual cues," in *Proceedings of the 9th Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, ser. PCM '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 317–326. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-89796-5_33

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

[20] S. Hidaka, "Identifying kinematic cues for action style recognition." Cognitive Science Society, 2012.

[21] B. K. Horn and B. G. Schunck, "Determining optical flow," in *1981 Technical symposium east*. International Society for Optics and Photonics, 1981, pp. 319–331.

[22] INEI, "Informe técnico - estadísticas de seguridad ciudadana," INEI, Tech. Rep., Mar 2015.

[23] Y. Ke and R. Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors," *Proceedings of the IEEE Computer Society Conference*, 2004.

[24] H. U. Keval, "Effective, design, configuration, and use of digital cctv," Ph.D. dissertation, University College London, 2008.

[25] O. Kliper-Gross, T. Hassner, and L. Wolf, "The action similarity labeling challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 615–621, Mar. 2012. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2011.209

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

[27] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, June 2008, pp. 1–8.

[28] J. Lin and W. Wang, “Weakly-supervised violence detection in movies with audio and video based co-training,” in *Advances in Multimedia Information Processing - PCM 2009*, ser. Lecture Notes in Computer Science, P. Muneesawang, F. Wu, I. Kumazawa, A. Roeksabutr, M. Liao, and X. Tang, Eds. Springer Berlin Heidelberg, 2009, vol. 5879, pp. 930–935. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-10467-1_84

[29] C. Liu, “Beyond pixels: exploring new representations and applications for motion analysis,” Ph.D. dissertation, Citeseer, 2009.

[30] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Journal of Machine Learning Research*, vol. 60, no. 2, pp. 91–110, Jan 2004.

[31] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI’81. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1981, pp. 674–679. [Online]. Available: http://dl.acm.org/citation.cfm?id=1623264.1623280

[32] K. Mikolajczyk and C. Schmid, “Scale and affine invariant interest point detectors,” *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, Oct 2004.

[33] D. Mora and A. Páez, “Detección de objetos móviles en una escena utilizando flujo Óptico,” Ph.D. dissertation, Pontificia Universidad Javeriana, Jun 2010.

[34] O. Oshin, A. Gilbert, and R. Bowden, “Capturing the relative distribution of features for action recognition,” in *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, March 2011, pp. 111–116.

[35] C. Penet, C.-H. Demarty, G. Gravier, and P. Gros, “Multimodal information fusion and temporal integration for violence detection in movies,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 2393–2396.

[36] F. Perronnin, J. Sánchez, and T. Mensink, “Improving the fisher kernel for large-scale image classification,” in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 143–156.

[37] C. B. Seminario, N. M. Huisa, J. L. Tapia, and I. U. Villanueva, “Seguridad ciudadana informe anual,” Instituto de Defensa Legal - IDL, Tech. Rep., 2014.

[38] T. Senst, V. Eiselein, and T. Sikora, “A local feature based on lagrangian measures for violent video classification,” in *Imaging for Crime Prevention and Detection (ICDP-15), 6th International Conference on*. IET, 2015, pp. 1–6.

[39] F. Souza, E. Valle, G. Chávez, and A. de A. Araújo, “Color-aware local spatiotemporal features for action recognition,” *16th Iberoamerican Congress, CIARP 2011*, pp. 248–255, Nov 2011.

[40] J. R. R. Uijlings, I. C. Duta, N. Rostamzadeh, and N. Sebe, “Realtime video classification using dense hof/hog,” in *Proceedings of International Conference on Multimedia Retrieval*, ser. ICMR ’14. New York, NY, USA: ACM, 2014, pp. 145:145–145:152. [Online]. Available: http://doi.acm.org/10.1145/2578726.2578744

[41] J. Wang, B. Li, W. Hu, and O. Wu, “Horror video scene recognition via multiple-instance learning,” in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 1325–1328.

[42] J. Wu, Z. Cui, V. S. Sheng, P. Zhao, D. Su, and S. Gong, “A comparative study of sift and its variants,” *Measurement science review*, 2013.

[43] L. Xu, C. Gong, J. Yang, Q. Wu, and L. Yao, “Violent video detection based on mosift feature and sparse coding,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, May 2014, pp. 3538–3542.

[44] R. Yan and M. Naphade, “Semi-supervised cross feature learning for semantic concept detection in videos,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, June 2005, pp. 657–663 vol. 1.

[45] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, “Evaluating bag-of-visual-words representations in scene classification,” in *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval*, ser. MIR ’07. New York, NY, USA: ACM, 2007, pp. 197–206. [Online]. Available: http://doi.acm.org/10.1145/1290082.1290111

[46] Z. Yang, T. Zhang, J. Yang, Q. Wu, L. Bai, and L. Yao, “Violence detection based on histogram of optical flow orientation,” pp. 906 718–906 718–4, 2013. [Online]. Available: http://dx.doi.org/10.1117/12.2051390

[47] L. Yeffet and L. Wolf, “Local trinary patterns for human action recognition,” in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 492–497.

[48] M. yu Chen and A. Hauptmann, “Mosift: Recognizing human actions in surveillance videos,” 2009.