Available online at www.sciencedirect.com

ScienceDirect

Electronic Notes in Theoretical Computer Science 329 (2016) 5–26

Electronic Notes in Theoretical Computer Science

www.elsevier.com/locate/entcs

Fast Face Detection in Violent Video Scenes

V. E. Machaca Arceda[1], K. M. Fernández Fabián[2], P. C. Laguna Laura[3], J. J. Rivera Tito[4] and J. C. Gutiérrez Cáceres[5]

University of Saint Augustine
Arequipa, Perú

Abstract

In this work we aim to detect faces in violence scenes, in order to help the security control. We used the Violent Flow (ViF) descriptor with Horn-Schunck proposed in [50] for violence scenes detection at first stage. Then we applied the non-adaptive interpolation super resolution algorithm to improve the video quality and finally we fire a Kanade-Lucas-Tomasi (KLT) face detector. In order to get a very low time processing, we paralleled the super resolution and face detector algorithms with CUDA. For the experiments we used the Boss Dataset and also we built a violence dataset, taking scenes from surveillance cameras. We have promising results detecting faces in this environment, because of the benefits of our proposal.

Keywords: ViF, Horn-Schunck, Optic Flow, Resolution enhancement, Video, Super-Resolution, Face detection, GPU parallel computing, Lower computational time.

1 Introduction

In recent years the fear and insecurity have been increasing considerably. Most of news in newspapers, television, radio or internet shows acts of vandalism, theft, murder, etc. For example: The United Nations Office on Drugs and Crime (UNODC) in its site *Global Study on Homicide*, they show the rate of homicides per 100, 000 inhabitants, with 16.3 rate in America against 3.0 in Europe. Moreover according to the Institute of Legal Defense (ILD-Peru) in 2015 the Peruvian people consider crime and insecurity as their mayor problem [9], the National Institute of Statistic and Informatic (NISI-Peru)'s technical report of security in Mar-2015 said

[1] Email: enriquefirst@gmail.com
[2] Email: karla.m.f.f@gmail.com
[3] Email: pamela.c.laguna@gmail.com
[4] Email: titex777@gmail.com
[5] Email: jcgutierrezc@gmail.com

http://dx.doi.org/10.1016/j.entcs.2016.12.002

that $30,5\%$ of people was the victim of a criminal act [6]

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 [22]. Because of this, we are motivate to seek technological solutions that help us feel safer. Having support systems to detect possible serious violent actions are very useful in controlling public safety. Moreover, it is important for us to detect the violent people in the scene.

A normal surveillance scene have poor conditions and it is very difficult to detect the people involve. In this context, we propose the use of super resolution algorithms to improve the results of a face detector algorithm in order to recognize de people in a violent scene. We aim 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.

2 Related Work

In the following paragraphs we show you the most relevant works in detection of violent actions as well as face detection in video. It is also important to mention the definition of *violence* taken by each author.

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 [17], they used a Harris detector [30] 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 [42] to recognize facial expressions, human activities and a mouse's behavior, getting 83%, 80% and 72% of accuracy respectively. In [55] Gaussian Difference [10] is used with Principal Component Analysis - Scale Invariant Feature Transform (PCA-SIFT) [54] 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 [28] presented a BoVW comparison varying the descriptors. In [49] descriptors as Histogram of Optical Flow (HOF) and Histogram of Oriented Gradient (HOG) with variations in optical flow are evaluated using Lucas-Kanade [36], Horn-Schunck [23], and Farnebäck [14] as optical flow algorithms, they also evaluated the performance of BoVW comparing K-means against Random Forests [6] and Fisher kernel [44], they concluded that Lucas-Kanade and Horn-Schunck outperformed Farnebäck and Fisher kernel outperformed K-means.

[6] 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 [27].

One of the first works detecting violence is based on audio presented by [48] 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, [7] used Mel-Frequency Cepstral Coefficients (MFCC) as audio descriptor and dynamic Bayesian networks. The main contribution of 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 [15], they proposed Bag of Visual Words (BoVW) with Space-Time Interest Point (STIP), based on Laptev's research [25], 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 [16] 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 [12] (it was proposed by [38]), 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 [20], 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) [57], histogram of oriented gradient (HoG) [32], histogram of oriented optical flow (HoF) [32] and histogram of oriented gradient and optical flow(HNF)[32]. The model is also evaluated in other datasets, as Hockey [12] and ASLAN [31] 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 [35] 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 [32], HOF [32], HNF [32] and ViF [20] outperforming them in the Crowded and Hockey datasets.

Other work based in optical flow is presented in [56] 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 [39] 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 [41], more specifically studies in this field show that simple kinematic characteristics like speed and acceleration are correlated to emotional attributes [21], thereby detecting the change in acceleration is based on the blur of the image when motion occurs, by calculating the spectral power as evidenced [4]. The results were evaluated in the Movies and Hockey [12] 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.

The used of Lagrangian theory show the applicability for video analysis in several aspects. In this context [47] 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 [35]. 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.

All works mentioned above focus on detecting violence related to fights, explosions, gunfire violence, etc. But there are also many works that focus to detect violent content in movies, specifically to have better control for children. For example [29] detect horror in movies using Multiple Instance Learning (MIL; MI-SVM [2]). In [18] the work of [48] is extended where they used a multimodal two-stage approach, they performed audio and visual analysis. A three-stage method is proposed in [19], where they used a semi-supervised cross-feature learning algorithm [45]. In the work of [33] two classifiers are used in co-training, they considered fights, explosions, murders and shots as violence concept. [43] 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 perfor-

mance in terms of MediaEval cost function [11], 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. 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.

In the case of face detection in video, it is one of the new fields where it is dabbling biometrics, where the properties of the video allow some considerations with the movement sequence of images that make up the video, allowing more feasible way the location of moving objects in the image, thanks to frame the difference that exists between the images. For video processing, there are various techniques, including one that measures variations in vertical and horizontal to find the eyes [26].

Some problems related to face detection are:

- In most cameras video surveillance, by the very fact of its location, it is very difficult to face focus directly [53], since different positions, occlusion, movement and also size variation arise due the distance from the optical axis [37].
- The most frequent problems in video, is the presence of shade due to different illumination conditions [13], hiding part of his face difficult to detect. In addition, some algorithms detect the face based on color, but generates confusion in the skin color because hair tones very similar color between neighboring pixels, thus affecting the rate of sharpness of the face.
- With respect to video surveillance, variation in video quality is defined by the resolution, it is very marked. For example, in Peru most video-recording systems have very low quality [3].
- In addition, due to the amount of information they have videos, calculations performed computationally take longer, therefore, the execution time is longer, avoiding obtain results in real time [8].
- Owing the videos are in uncontrolled conditions, the sum of all these problems, negatively affecting most methods and techniques of face detection, as the percentage of accuracy in the results decreases confusion in the calculation, increasing number of false positives, because confuses another portion of pixels and displays it as a false positive.

3 Proposal

The proposal is divided in three stages: A Violent scene detector, a normalization algorithm and a face detector, we could see it in Figure 1.

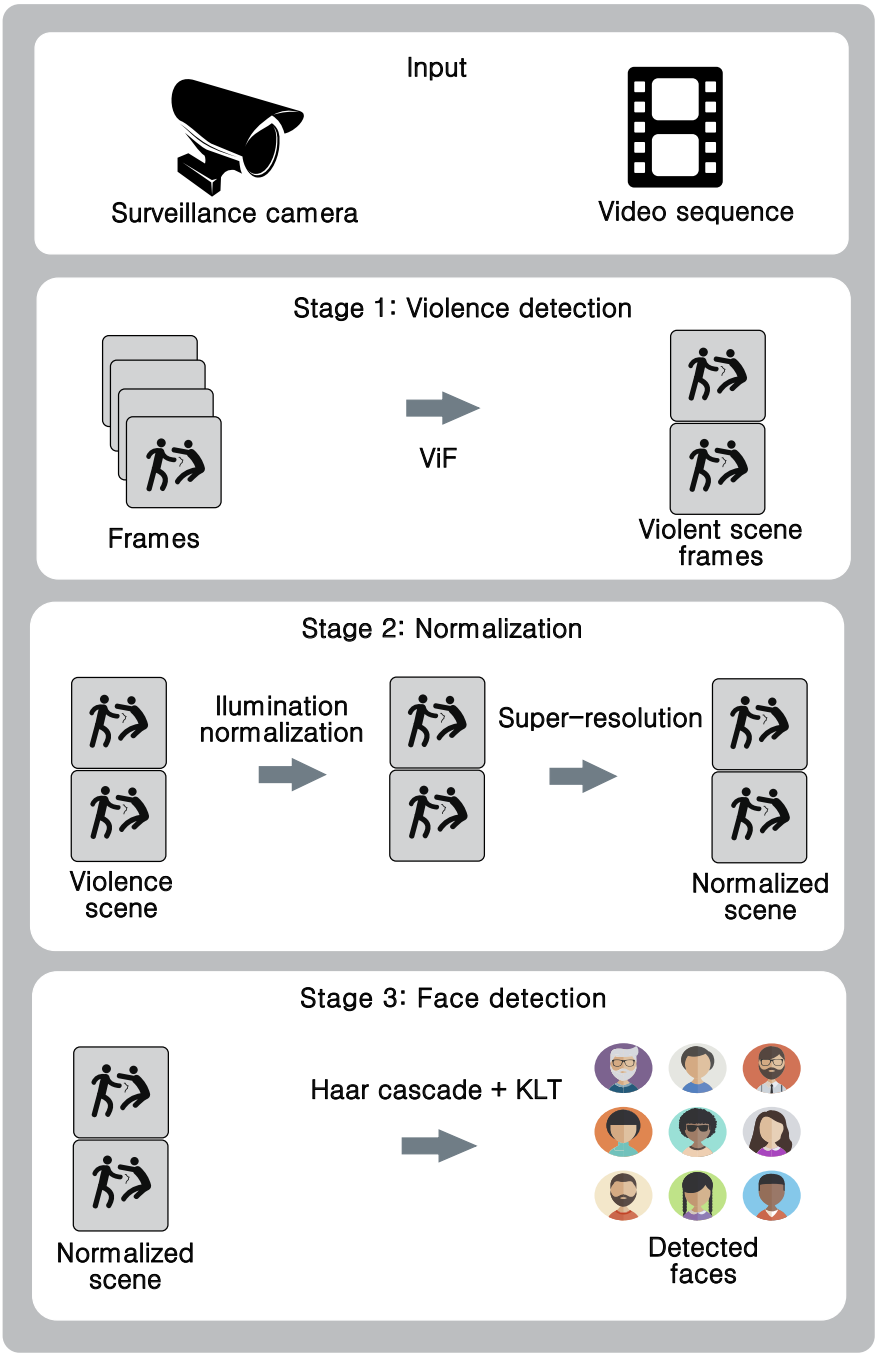

Fig. 1. The three stages of face detection in violence video scenes.

3.1 The Violence Detection Method

In order to detect a violence scene, we are going to use the ViF descriptor because of its very low cost and acceptable accuracy, the ViF descriptor consider the statistics

of magnitude changes of flow vectors over time as we see in Figure 2, In order to get these vectors [20] used the optical flow algorithm proposed by [34] named Iterative Reweighted Least Squares (IRLS), but nowadays we have a lot of different optical flow algorithms, in this context, we used the ViF descriptor with Horn-Schunck [23] as optical flow algorithm proposed by [50].

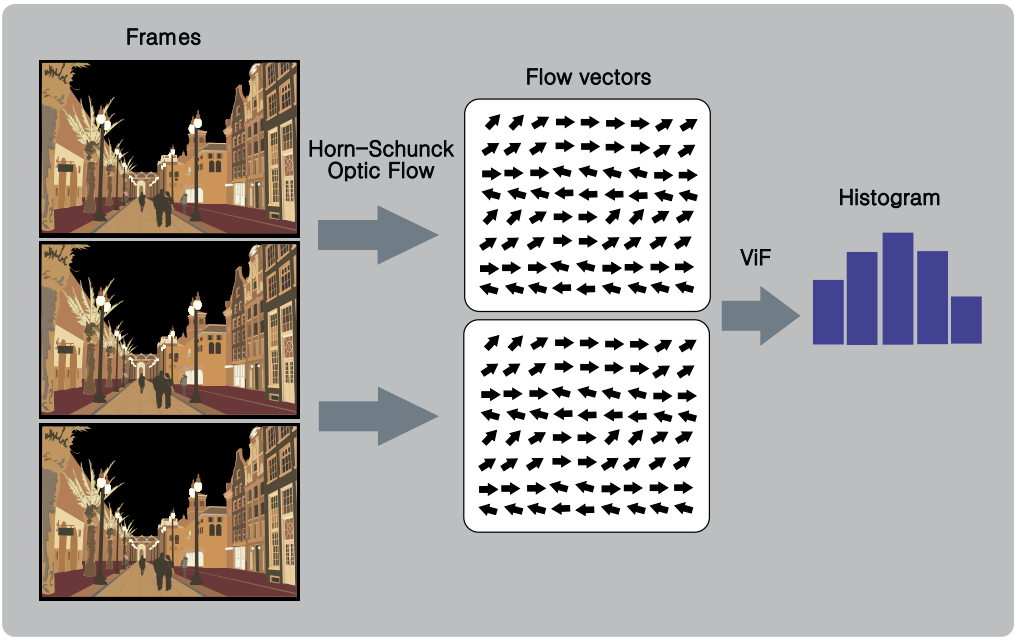

Fig. 2. ViF descriptor in video.

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 [20].

Then the violence detector is trained using a SVM classifier with a polynomial kernel, taking as input the result of ViF descriptor. In the experiments we use cross-validation with k=10. The whole method for detect violence in video is shown in Figure 3

3.2 Super-Resolution and standardization of video sequences

In order to perform real-time processing, taking into account the current problems of poor resolution, we performed a super resolution algorithm and face detector using the GPU. For better understanding, the flow of each component is shown in Figure 4.

Algorithm 1 ViF descriptor

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)

1: **for** $t = 1$ to T **do**
2: Get optical flow $(u_{x,y,t}, v_{x,y,t})$ of each pixel $p_{x,y,t}$ where t is the frame index.
3: Get magnitude vector: $m_{x,y,t} = \sqrt{u_{x,y,t}^2 + v_{x,y,t}^2}$
4: 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 $|m_{x,y,t} - m_{x,y,t-1}|$.

Starting with the brightness compensation, it consists of: Gamma Intensity Correction (GIC), Difference Gauss (DG), Local Histogram Coincidence (LHC) and Local Normal Distribution (LND). Where gamma correction of an image is a non-linear transformation, grayscale, which replaces the input image g with exponent g^{γ}.

The default image g_0 has normal illumination conditions. Given an input image $g(x, y)$, its GIC corrected image is $g'(x, y)$, and it is calculated by transforming the input image (x, y) pixel by pixel with optimal gamma coefficient γ^*.

As a sign of the whole process that involves illumination normalization, the Algorithm 2 is presented.

Algorithm 2 Illumination normalization

Data: Image $g(x, y)$
Result: Image $c(x, y)$

1: Gamma Intensity Correction GIC: $\gamma^* = arg \ \ min \sum_{x,y} [G(g(x, y); \gamma) - g_0(x, y)]^2$
2: Difference Gauss DG: $G_\sigma(x, y) = \frac{1}{2\pi F_n^2} \ \ e^{-(x^2+y^2)/2F_n^2} - N$
3: Apply local histogram coincidence: $c(x, y) = \frac{E^{-1}(s) - \mu_i}{\sigma_i}$

This illumination compensation procedure can work on various lighting conditions, shadows and other conditions of the original image, which can preserve the essential elements for the detection of visual appearances.

A classic form of super resolution is resizing the kernel, using the non-adaptive interpolation [51], as this approach treats all pixels in an image alike, where normally the location of the pixel is multiplied with the function kernel and then multiplied by the known discrete sample; the same process is repeated for all pixel locations. Then adjacent pixels ranging from 0 to 256 are used for bilinear interpolation which determines the value of the gray level of the weighted average of the four nearest pixels and assigns that value to the output coordinate.

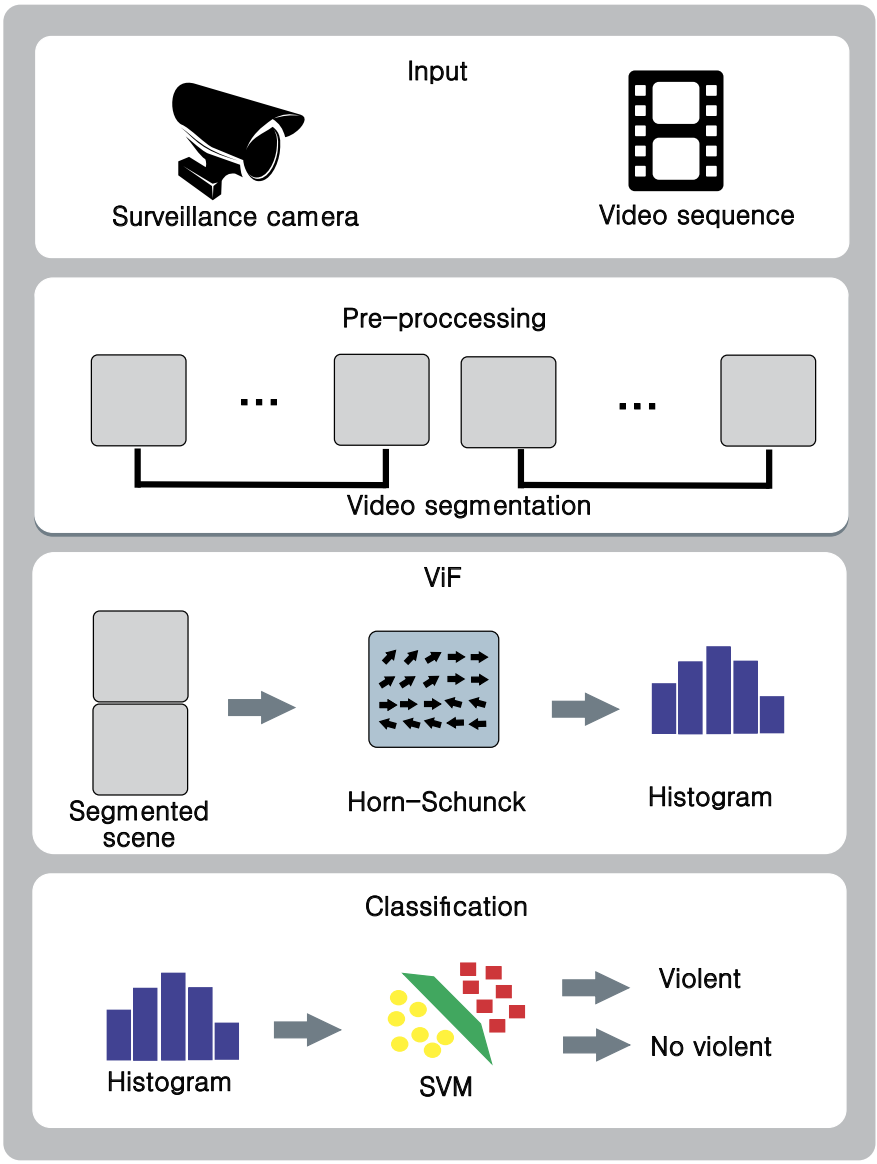

Fig. 3. The violence detection method.

The Super Resolution (SR) reconstruction is represented by equation 2.

$$Y = DBM + n \tag{2}$$

Where: D = sampling operators, B = blur, M = deformation, n = noise (Gaussian white noise commonly additive).

Then follows the similarity nonlocal descriptor where the auto nonlocal similarity information is useful for improving the quality of the reconstructed images. [24] proposes a method that builds the nonlocal similarity descriptor to present the structural characteristics of the image and helps to predict global and local

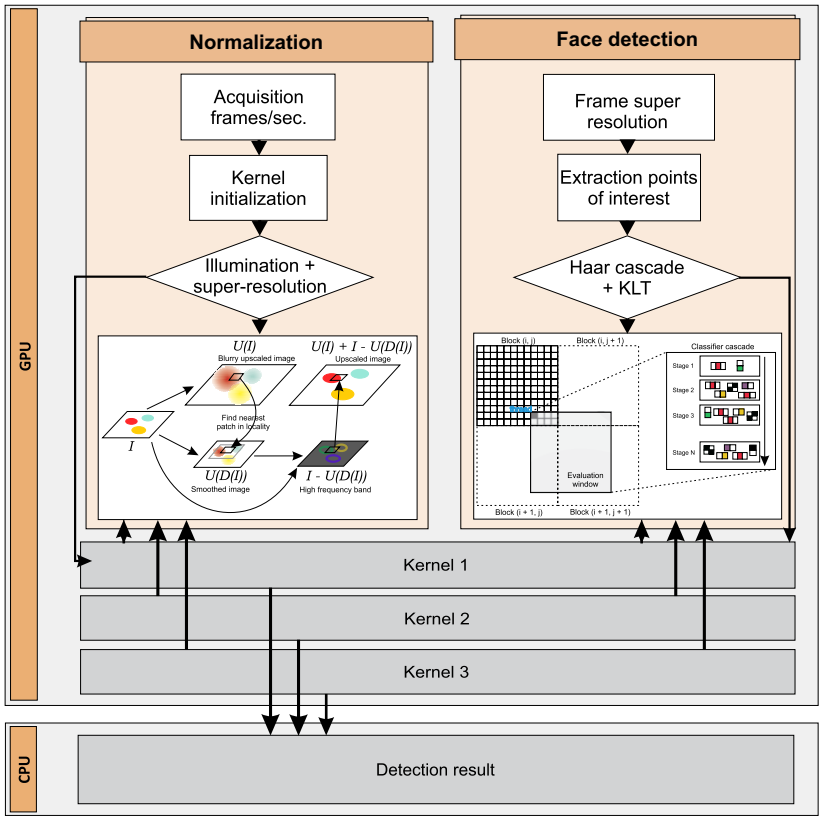

Fig. 4. Communication diagram of video normalization and face detection.

information containing movement between continuous frames.

The local movement is used to represent the motion. In Figure 5 (a) a red and a green squares are shown, then they are transferred to new destinations with a single motion vector from a time t, to time $t+1$, and then from time $t+1$ to $t+2$, the red, green and blue squares move to nearby places with three different movement vectors, which are given from local movements of different objects in the image. This behavior often occurs irregularly throughout the video sequence for that reason it is called *nonlocal similarity*, note that is very relevant the estimate of movement, as well as shown in part (b) of Figure 5.

The frame super-resolution rebuilt, will be under the observation model of low resolution, consistent with the input frame. For this process it will use the iterative back projection method, which ensures the convergence of the iterative process and makes the reconstructed image close to the original high resolution image.

The mathematical description of Iterative Back Projection (IBP) algorithm is:

$$\hat{f}^{k+1} = \hat{f}^{k} - \lambda \sum_{i=1}^{P} H_i^{BP}(\hat{y}_i^k - y_i) \tag{3}$$

Where:

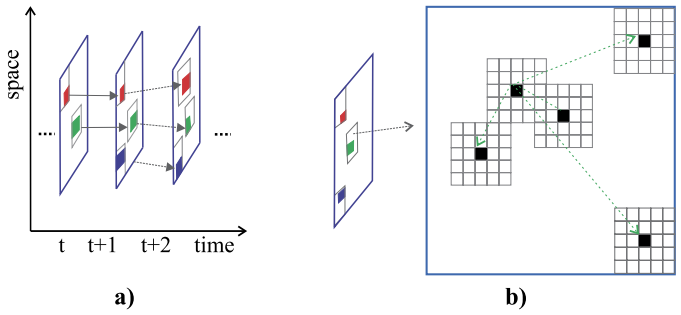

Fig. 5. (a) Global and local movement simultaneously. (b) Motion vectors. Extracted from [24].

- k = Iteration number
- $\hat{f}^{k+1}$ y $\hat{f}^{k}$ the frames obtained in super-resolution, acquired in $(k+1)'$ and k' iterations.
- $\hat{y}_i^k$ is the i' low resolution frame $\hat{f}^k$ under the model of observation low-resolution frame.
- λ = Step gradient.

Estimated by projecting the image in high resolution the low resolution model degraded H, where P simulated the low-resolution frames generated. Then, the error simulated re-projected onto the high resolution image. Finally, according to the sum of rear projection error, the estimated high resolution image is renewed. The above process is repeated until the number of iterations reaches its maximum value.

3.3 Face detection

Face detection is given according to the location of interest points, basically it consists of the following steps:

- Detection and segmenting regions of the skin.
- Check each component of the face (mouth, nose, eyes, and eyebrows). Where a bounding box of the face detected depending on the anthropometric form of a human face is marked.

Then a deep cascade evaluation is made with a group of classifiers organized in stages containing filters [52], in order to detect faces in multiple locations and different sizes. Also the kernel not only have to test all pixels, but also all the possible dimensions, as we see in Figure 6. Moreover the data structure cascade is constrained by dependencies between stages, and should be evaluated sequentially for each candidate [40].

It is understood that the normalization (super resolution and illumination) will work in a block (0.0), while it in parallel will work in the block (1.0) with face detection. As CUDA consists of functions running simultaneously in several very lightweight threads on the GPU, these threads have a hierarchy: the blocks, which

are a set of threads in 1, 2 or 3 dimensions, then this one grid which is a set of blocks that can be 1 or 2 dimensional [46]

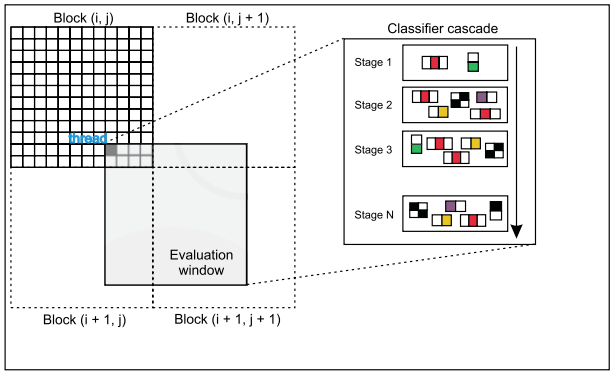

Fig. 6. Evaluation cascade classifier block. Extracted from [40].

An example of the classifier cascade's operation over each video frame, is shown in Figure 7, which affords better method used in the proposal.

Fig. 7. Example of Haar classifier's features in the face.

4 Experiment and Results

4.1 Datasets

We evaluated the performance in the Boss dataset, which is a collection of videos from different cameras inside a train, in Figure 8, we can see some frames. Also we built a new dataset with violent and no violent videos taken from surveillance cameras, we named it as Surveillance Videos (SV), we can see some frames in Figures 9. Moreover in Table 1 we can see a comparison of the datasets. We have to mention that in the SV dataset it is very difficult to detect faces because of the poor illumination, resolution, framerate, etc.

	Resolution	Framerate per second	Duration (seconds)	Number of videos
Boss	720 x 576	25	200	70
SV	480 x 360	25	2	200

Table 1
Datasets features.

Fig. 8. Some frames taken from the Boss dataset.

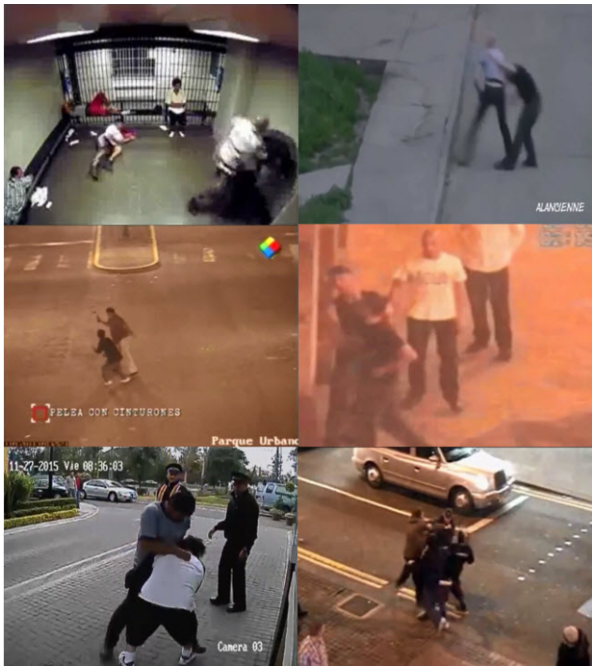

Fig. 9. Some frames taken from the SV dataset.

4.2 Results in violent scenes detection

We evaluate the performance of ViF with Horn-Schunck in a SVM classifier with a polynomial kernel and cross-validation (k=10). In Table 2 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. Moreover, for this experiment we put together the SV and Boss datasets. We know that these datasets contains videos in real situations, with very poor quality and that is the reason why we got a low accuracy, but with an acceptable AUC. We have to mention that we chose this method because of the low cost against others in the literature.

Surveillance Videos and Boss dataset		
Algorithm	ACC ± SD	AUC
ViF (Horn-Schunck)	0.6600 ± 0.1174	0.8500

Table 2
The performance of ViF with Horn-Schunck as optic flow algorithm. 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.

The Receiver Operating Characteristic (ROC) of the classifier with ViF in the SV dataset is shown in Figure 10.

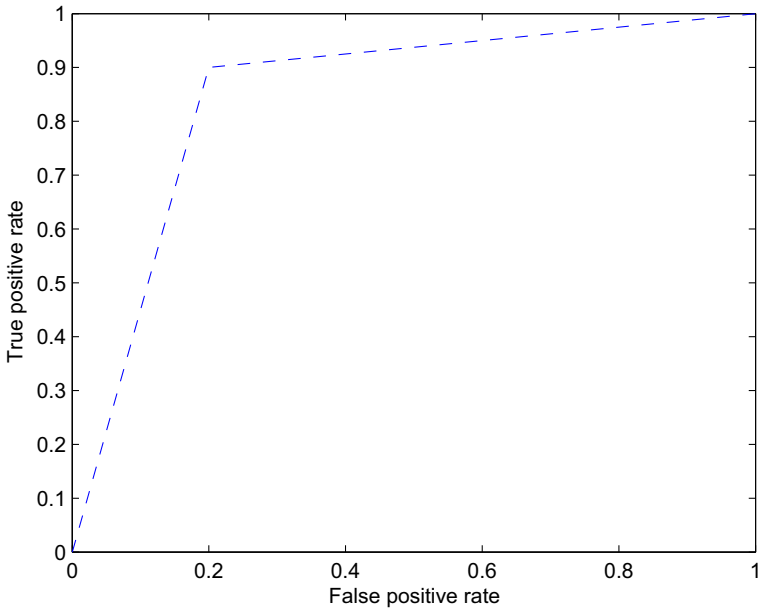

Fig. 10. ROC curve of a SVM classifier with ViF in the SV dataset.

An evaluation of time processing of ViF with Horn-Schunck is show in Table 3, here we got an almost real time system. Also we have consider that this is just the

first stage of our proposal, so we have to save time processing for the next stages. The measurement was evaluated in a computer with a 1.8 GHz processor.

	Video duration (seg.)	Time processing (seg.)
ViF with Horn-Schunck	2	2.1563

Table 3
Time processing of ViF with Horn-Schunck for violence detection.

4.3 Results in super resolution and face detection

As a sample of improved super resolution, the Table 4 is shown in binary codes, for better visualization of the frame iteration, which also tries to minimize the margin error based on the adjacent pixels. After Gaussian filter the pixels which were filter and assigned to their coordinates, have a value of 1.

	VIDEO SUPER-RESOLUTION																											
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	1	0	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	0	0	1	0	1
2	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	0	0	1	1	1	0	0	1	0	0	1
3	1	0	0	1	0	1	0	1	1	1	0	0	1	1	1	1	0	1	1	1	0	0	1	0	1	1	0	0
4	1	0	1	1	0	0	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	0	1	1	0
5	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	0	1	0	1	1	0	1
6	0	0	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	1	1	1	0	0	1	1	1
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	0
8	1	1	1	1	1	1	1	0	1	0	0	1	1	1	1	1	0	1	1	1	1	1	0	1	0	1	1	1
9	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1
10	0	1	1	1	1	1	0	0	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1
11	1	1	0	0	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1
12	1	1	1	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0
13	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	1	1	1	0	1	1
14	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	0	0	1	0	1
15	1	1	1	1	1	1	0	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	0	1
16	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	1	0	0	1	1	1
17	0	1	1	1	0	0	1	1	1	1	1	0	1	1	1	1	0	1	1	0	1	1	1	1	1	1	0	1
18	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1
19	1	1	0	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0
20	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	0

Table 4
Matrix showing the change in video resolution. After Gaussian filter the pixels which were filter and assigned to their coordinates, have a value of 1.

A example of the face detection and super resolution is shown in Figure 11. The Lucas Kanade and Tomasi (KLT) face detector cannot detect any face at the beginning but after a super resolution algorithm (the video's resolution was increased to 700 x 497 pixels) our detector found two faces.

Tests with different video features were made. A low-resolution video (535 kbps with 10 frames per second) is shown in Figure 12, here we didn't detect faces but we could see a improvement in video quality.

Fig. 11. Improved image resolution and face detection.

Fig. 12. Improved video framerate of 10 fps and 535 kbps.

There is also false positives in our results, as we can see in Figure 13, here we worked with a video of 987 kbps with 20 frames per second.

Fig. 13. Improved video framerate of 20 fps and 987 kbps.

In Figure 14 we have a 8 seconds video, here the camera recorded a person getting into a laboratory. As we see, at the beginning we had good results until we got a face occlusion by hair and illumination.

The results of face detection within videos with different frame rates are summarized in Table 5. Also we included the time processing in CPU and GPU and moreover the accuracy.

In Table 6 we have face detection accuracy based on various video features. The measure is based on framerate and the data flow (kbits per second). It is observed that while the data flow is lower, it is more difficult to detect faces. Moreover we saw that the super resolution improve the quality and face detection but when the video have a very poor resolution and data flow the results were bad.

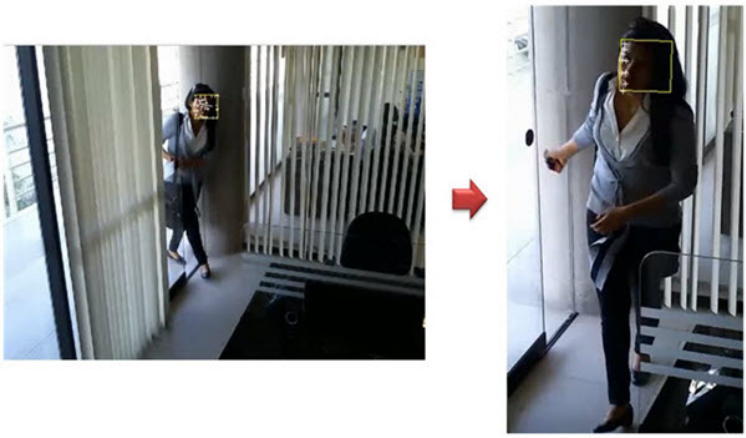

Fig. 14. Detection and tracking of points of interest in faces of a person.

Face detect with KLT	**Time process/s**		**%**
	CPU	**GPU**	**Accuracy**
Video than 8 seconds with 30 f/s.	858	283	94%
Video than 7 seconds with 30 f/s.	735	243	94%
Video than 5 seconds with 29 f/s.	506	167	92.5%
Video than 249 seconds with 25 f/s.	19423	6474	89%
Video than 78 seconds with 25 f/s.	6085	2028	88%
Video than 206 seconds with 20 f/s.	12972	4326	74%
Video than 5 seconds with 15 f/s.	251	85	62%
Video than 4 seconds with 10 f/s.	141	49	51%
Video than 2 seconds with 10 f/s.	68	26	48.5%
Video than 2 seconds with 7 f/s.	53	21	29%

Table 5
Results of processing time with videos of different duration and amount of f/s, in this case random videos of the various databases used.

For testing, a computer with GPU is used and the power and the tools incorporated in its latest version, Matlab R2016, with the toolbox for parallel computing and computer vision.

Then in Figure 15 efficiency has GPU compared CPU to perform extensive mathematical calculations and especially, the use of massive data, which in this case becomes the sum of matrices generated for each frame of video, and the sum of the processes to be performed on an instance of time.

Proving that CPU does not work well with large amounts of data; in contrast

Videos with variation in frames	Average speed(s)		Quantity detected face	% Accuracy
	CPU	GPU		
Videos with lower framerate 10 fs/s	170	57	1-	14%
Videos with lower framerate 15 fs/s	295	98	2-	37%
Videos with framerate near to 25 fs/s	320	105	4±	78%
Videos with framerate equal to 30 fs/s	365	121	5±	91%
Videos with, higher framerate 35 fs/s	390	130	8+	97%

Table 6
Results average processing time in seconds and percentage of successes in the face detection optimization with super-resolution.

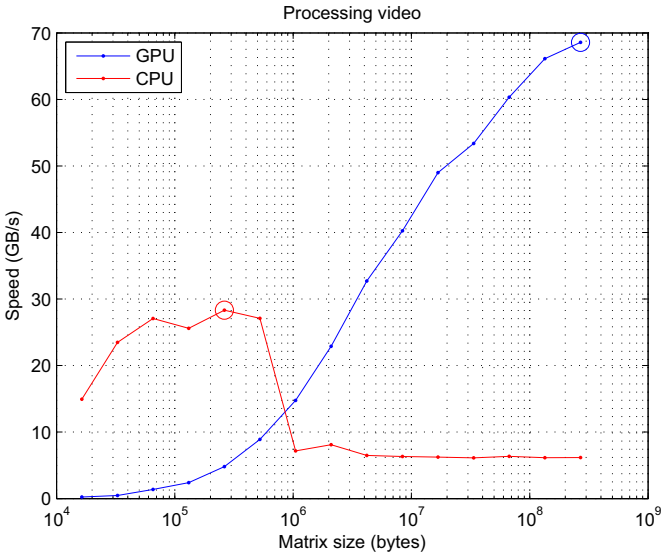

Fig. 15. GPU and CPU comparison with the use of bandwidth for processing large arrays based on time and size.

GPU works well with large amounts of data without diminishing over time due to work in multi-hiloque that works inside the kernel.

5 Conclusions

The use of ViF with Horn-Schunck is highly acceptable due to its low computational cost and good results. The time processing of the descriptor for a 2 seconds video was just 2.15 seconds, this is almost a real time system consider also that we performed the experiments in a slow computer.

It is shown that the implementation of super-resolution improves the quality of videos, and also generates more pixels per frame, as well as higher data flow, which contributes to process as many images that together help to get one frame of higher resolution.

For some very low resolution videos, the SR does not meet expectations, because having too little of pixels per frame, interpolation is difficult without being able to generate higher resolution video.

The use of GPU reduce considerably the computational cost of super resolution and face detection. We expect to use it in real time applications.

References

[1] Acar, Esra and Irrgang, Melanie and Maniry, Dominique and Hopfgartner, Frank, *Detecting Violent Content in Hollywood Movies and User-Generated Videos*, Springer International Publishing, book "Smart Information Systems" Advances in Computer Vision and Pattern Recognition, (2015), 291–314.

[2] Andrews, Stuart and Tsochantaridis, Ioannis and Hofmann, Thomas, *Support vector machines for multiple-instance learning*, book "Advances in neural information processing systems", (2002), 561–568.

[3] Arashloo, Shervin Rahimzadeh and Kittler, Josef, *Fast pose invariant face recognition using super coupled multiresolution Markov Random Fields on a GPU*, "Pattern Recognition Letters". **48** (2014). 49–59.

[4] Barlow, Horace B and Olshausen, Bruno A, *Convergent evidence for the visual analysis of optic flow through anisotropic attenuation of high spatial frequencies*, "Journal of Vision", **4**, number 6, (2004).

[5] Blake, Randolph and Shiffrar, Maggie, *Perception of human motion*, "Annu. Rev. Psychol", **58**, (2007). 47–73.

[6] Breiman, Leo, *Random forests*, "Machine learning", **45** (2001). 5–32.

[7] Bruno do Nascimento Teixeira, *MTM at MediaEval 2014 Violence Detection Task*, "MediaEval 2014, Multimedia Benchmark Workshop", (2014).

[8] Castrillón, M and Déniz, Oscar and Guerra, Cayetano and Hernández, Mario , *ENCARA2: Real-time detection of multiple faces at different resolutions in video streams*, "Journal of Visual Communication and Image Representation", **18**, number 2, (2007). 130–140.

[9] César Bazán Seminario and Nancy Mejía Huisa and Jorge Levaggi Tapia and Isabel Urrutia Villanueva, *Seguridad Ciudadana Informe Anual*, Instituto de Defensa Legal - IDL, (2014).

[10] David G. Lowe, *Distinctive Image Features from Scale-Invariant Keypoints*, "Journal of Machine Learning Research", **60**, (2004), 91–110.

[11] Demarty, Claire-Hélène and Penet, Cédric and Schedl, Markus and Bogdan, Ionescu and Quang, Vu Lam and Jiang, Yu-Gang, *The MediaEval 2013 affect task: violent scenes detection*, (2013).

[12] Enrique Bermejo and Oscar Deniz and Gloria Bueno and Rahul Sukthankar, *Violence Detection in Video Using Computer Vision Techniques*, "14th International Conference", (2011). 332–339.

[13] Enrique G. Ortiz and Brian C. Becker, *Face recognition for web-scale datasets*, "Computer Vision and Image Understanding", **118**, (2014), 153–170, URL: http://www.sciencedirect.com/science/article/pii/S1077314213001744

[14] Farnebck, Gunnar, *Two-Frame Motion Estimation Based on Polynomial Expansion*, book Springer Berlin Heidelberg "Image Analysis", Lecture Notes in Computer Science. **2749**, (2003), 363–370.

[15] Fillipe D. M. de Souza and Guillermo C. Chávez and Eduardo A. do Valle Jr. and Arnaldo de A. Araújo, *Violence Detection in Video Using Spatio-Temporal Features*, "Conference on Graphics, Patterns and Images (SIBGRAPI)", **23**, (2010). 224–230.

[16] Fillipe Souza and Eduardo Valle and Guillermo Chávez and Arnaldo de A. Araújo, *Color-Aware Local Spatiotemporal Features for Action Recognition*, "16th Iberoamerican Congress, CIARP 2011", (2011), 248–255.

[17] G. Csurkaa and C. R. Dance, L. Fan and J. Willamowski and C. Bray, *Visual Categorization with Bags of Keypoints*, "Workshop on Statistical Learning in Computer Vision", (2004). 1-.22.

[18] Giannakopoulos, Theodoros and Makris, Alexandros and Kosmopoulos, Dimitrios and Perantonis, Stavros and Theodoridis, Sergios, *Audio-Visual Fusion for Detecting Violent Scenes in Videos*, book Springer Berlin Heidelberg "Artificial Intelligence: Theories, Models and Applications", **6040**, (2010), 91–100.

[19] Gong, Yu and Wang, Weiqiang and Jiang, Shuqiang and Huang, Qingming and Gao, Wen, *Detecting Violent Scenes in Movies by Auditory and Visual Cues*, book "Proceedings of the 9th Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing", series PCM '08, (2008), 317–326.

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

[21] Hidaka, Shohei, *Identifying kinematic cues for action style recognition*, Cognitive Science Society, (2012).

[22] Hina Uttam Keval, *Effective, Design, Configuration, and Use of Digital CCTV*, University College London, (2008).

[23] Horn, Berthold K and Schunck, Brian G, *Determining optical flow*, book "1981 Technical symposium east", International Society for Optics and Photonics, (1981), 319–331.

[24] Hu, Jie and Li, Hailiang and Li, Ying, *Real time super resolution reconstruction for video stream based on GPU*, book "Orange Technologies (ICOT)", 2014 IEEE International Conference on, (2014), 9–12.

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

[26] Iborra, Marcelo J Armengot, *Análisis comparativo de métodos basados en subespacios aplicados al reconocimiento de caras*, Universidad de Valencia, (2006).

[27] INEI, *Informe Técnico - Estadísticas de Seguridad Ciudadana*, (2015).

[28] Jian Wu and Zhiming Cui and Victor S. Sheng and Pengpeng Zhao and Dongliang Su and Shengrong Gong, *A comparative study of SIFT and its variants*, "Measurement science review", (2013).

[29] Jianchao Wang and Bing Li and Weiming Hu and Ou Wu, *Horror video scene recognition via Multiple-Instance learning*, book "Acoustics, Speech and Signal Processing (ICASSP)", 2011 IEEE International Conference on. (2011), 1325–1328.

[30] K. Mikolajczyk and C. Schmid, *Scale and affine invariant interest point detectors*, "International Journal of Computer Vision", **60**, number 1, (2004), 63–86.

[31] Kliper-Gross, Orit and Hassner, Tal and Wolf, Lior, *The Action Similarity Labeling Challenge*, "IEEE Trans. Pattern Anal. Mach. Intell", **34**, (2012), 615–621.

[32] Laptev, I. and Marszalek, M. and Schmid, C. and Rozenfeld, B., *Learning realistic human actions from movies*, book "Computer Vision and Pattern Recognition", 2008. CVPR 2008. IEEE Conference, (2008), 1–8.

[33] Lin, Jian and Wang, Weiqiang *Weakly-Supervised Violence Detection in Movies with Audio and Video Based Co-training*, "Advances in Multimedia Information Processing - PCM", Lecture Notes in Computer Science **5879**. Springer Berlin Heidelberg, (2009), 930–935.

[34] Liu, Ce, *Beyond pixels: exploring new representations and applications for motion analysis*, Citeseer, (2009).

[35] Long Xu and Chen Gong and Jie Yang and Qiang Wu and Lixiu Yao, *Violent video detection based on MoSIFT feature and sparse coding*, book "Acoustics, Speech and Signal Processing (ICASSP)", IEEE International Conference on. (2014), 3538-3542.

[36] Lucas, Bruce D. and Kanade, Takeo, *An Iterative Image Registration Technique with an Application to Stereo Vision*, book "Proceedings of the 7th International Joint Conference on Artificial Intelligence", **2**, series IJCAI'81, Vancouver, BC, Canada, (1981), 674–679, URL: http://dl.acm.org/citation.cfm?id=1623264.1623280

[37] Matta, Federico and Dugelay, Jean-Luc, *Person recognition using facial video information: A state of the art*, "Journal of Visual Languages & Computing", **20**, number 3, (2009), 180–187.

[38] Ming-yu Chen and Alex Hauptmann, *Mosift: Recognizing human actions in surveillance videos*, (2009).

[39] O. Deniz and I. Serrano and 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", **2**, (2014). 478–485.

[40] Oro, David and Fern'ndez, Carles and Segura, Carlos and Martorell, Xavier and Hernando, Javier , *Accelerating boosting-based face detection on gpus*, "Parallel Processing (ICPP)", 2012 41st International Conference. IEEE, (2012), 309–318.

[41] Oshin, O. and Gilbert, A. and Bowden, R., *Capturing the relative distribution of features for action recognition*, book "Automatic Face Gesture Recognition and Workshops", 2011 IEEE International Conference on, (2011),111–116.

[42] P. Dollar and V. Rabaud and 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", **2**, (2005), 65–72.

[43] Penet, Cédric and Demarty, Claire-Helene and Gravier, Guillaume and Gros, Patrick, *Multimodal information fusion and temporal integration for violence detection in movies*, book "Acoustics, Speech and Signal Processing (ICASSP)", 2012 IEEE International Conference on. IEEE, (2012), 2393–2396.

[44] Perronnin, Florent and Sánchez, Jorge and Mensink, Thomas, *Improving the fisher kernel for large-scale image classification*, book "Computer Vision–ECCV 2010". Springer, (2010), 143–156.

[45] Rong Yan and Naphade, M., *Semi-supervised cross feature learning for semantic concept detection in videos*, book "Computer Vision and Pattern Recognition" CVPR, **1**, IEEE Computer Society Conference on. (2005), 657–663.

[46] Santiago Cortijo Pablo Crovetto, Daniel Palomino, *Reconocimiento de patrones faciales en tiempo real mediante transformada de wavelet y computación paralela*, Universidad Nacional de ingeniería. Centro de tecnologías de información y comunicaciones, (2010).

[47] Senst, Tobias and Eiselein, Volker and Sikora, Thomas, *A local feature based on lagrangian measures for violent video classification*, book "Imaging for Crime Prevention and Detection (ICDP-15)", 6th International Conference on. IET, (2015), 1–6.

[48] T. Giannakopoulos and D. I. Kosmopoulos and A. Aristidou and S. Theodoridis, *Violence Content Classification Using Audio Features*, "ECCV 2004 workshop Applications of Computer Vision", (2006), 502–507.

[49] Uijlings, J. R. R. and Duta, I. C. and Rostamzadeh, N. and Sebe, N., *Realtime Video Classification Using Dense HOF/HOG*, book "Proceedings of International Conference on Multimedia Retrieval", series ICMR '14, New York, NY, USA. ACM, (2014), 145:145–145:152.

[50] V. Machaca Arceda and K. Fernańdez Fabián and J.C. Gutiérrez, *Real Time Violence Detection in Video*, "IET Conference Proceedings", Institution of Engineering and Technology. (2016).

[51] Van Ouwerkerk, JD, *Image super-resolution survey*, "Image and Vision Computing" **24** number 10. Elsevier, (2006), 1039–1052.

[52] Viola, Paul and Jones, Michael, *Rapid object detection using a boosted cascade of simple features*, bookt "Computer Vision and Pattern Recognition", Proceedings of the 2001 IEEE Computer Society Conference on, **1**, (2001), I–511.

[53] Winarno, Edy and Harjoko, Agus and Arymurthy, Aniati Murni and Winarko, Edi, *Improved Real-Time Face Recognition Based On Three Level Wavelet Decomposition-Principal Component Analysis And Mahalanobis Distance*, "Journal of Computer Science)",**10**, number 5. Science Publications, (2014), 844–851.

[54] Yan Ke and Rahul Sukthankar, *PCA-SIFT: A More Distinctive Representation for Local Image Descriptors*, "Proceedings of the IEEE Computer Society Conference", (2004).

[55] Yang, Jun and Jiang, Yu-Gang and Hauptmann, Alexander G. and Ngo, Chong-Wah, *Evaluating Bag-of-visual-words Representations in Scene Classification*, book "Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval", series MIR '07, Augsburg, Bavaria, Germany. ACM, (2007), 197–206, URL: http://doi.acm.org/10.1145/1290082.1290111

[56] Yang, Zhijie and Zhang, Tao and Yang, Jie and Wu, Qiang and Bai, Li and Yao, Lixiu, *Violence detection based on histogram of optical flow orientation*, "Proc. SPIE", **9067**, (2013), 906718-906718-4.

[57] Yeffet, Lahav and Wolf, Lior, *Local trinary patterns for human action recognition*, "Computer Vision", IEEE 12th International Conference on. (2009), 492–497.