MOVING OBJECT ANALYSIS IN VIDEO SEQUENCES USING SPACE-TIME INTEREST POINTS

Alain Simac-Lejeune
Litii, Alpespace, 15 rue St Exupery, 73800 Francin, France

Keywords: Video Signal Processing, Image Object Detection, Event Detection, Motion Analysis, Interest Points.

Abstract: Among all the features which can be extracted from videos, we propose to use Space-Time Interest Points (STIPs). STIPs are particularly interesting because they are simple and robust low-level features providing an efficient characterization of moving objects within videos. In this paper, after defining STIPs and after giving some of their properties, we will use STIPs to detect moving objects and to characterize specific changes in the movements of these objects. Proposed results are obtained from two very different types of videos, namely athletic videos and animation movies.

1 INTRODUCTION

The human perception system is naturally attracted by differences between parts of images and by motion or moving objects. Therefore, in the video indexing framework, interest points provide useful information which may be related to a semantic content. Different methods have been suggested to extract spatial interest points. An evaluation of these approaches is proposed in (Schmid et al., 2000). In (Laptev and Lindeberg, 2003), Laptev and Lindeberg propose a spatio-temporal extension of the interest point detection, denoted Space-Time Interrest points (STIPs) in the following. STIPs are interest points which are interesting both in the spatial and temporal domains. STIPs have been used for action recognition (Ke et al., 2005), automatic summarization (Laganiere et al., 2008) or, more generally, spatio temporal event detection (Laptev, 2005). In this paper, we propose to use STIPs to detect moving objects in videos and to characterize some specific changes in the movement of these objects. To illustrate the robustness of this approach, two very different types of videos are used : athletic videos and animation movies. The paper is organized as follows : Section 2 briefly describes the videos which are used in this study. Section 3 introduces STIPs and gives an overview of some STIPs specific properties. Section 4 and 5 show some outcomes obtained on moving objects dtection and on the localization of specific movement changes, respectively. Finally, Section 6 explores the limitations of the proposed method.

2 DATABASE

In order to characterize our work and test our assumptions, we have used three different types of data:

- synthesis videos : 60 sequences composed of synthetic images with a uniform background, one or more objects (round, square, triangle, polylines) in uniform motion or not, straight or not with a 288x288 image size;
- sport videos : 40 sequences of athletic jumps having 100 to 160 frames (about 5 seconds) with a 300x300 image size (Ramasso, 2007);
- an animation movie from the International Festival of Animated Movies of Annecy. The movie, entitled ”Le Moine et le Poisson”, lasts 6 minutes and 23 seconds (5745 frames) with a 320x240 image size.

It can also be noted that in all the following tests, performances has been evaluated on separated shots and without taking into account STIPs generated by shot transitions. Indeed, in the tested videos or movies, transitions can easily be detected.

3 SPACE-TIME INTEREST POINTS (STIPS)

3.1 Detection

On an image, spatial interest points (SIPs) can be

Simac-Lejeune A..
MOVING OBJECT ANALYSIS IN VIDEO SEQUENCES USING SPACE-TIME INTEREST POINTS.
DOI: 10.5220/0003866402010204
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2012), pages 201-204
ISBN: 978-989-8565-03-7

defined as pixels with a significant intensity variation. Examples of interest points are corners, junctions, isolated points or specific texture points. In (Harris and Stephens, 1988), Harris proposes to find such points using a second moment matrix.
In (Laptev and Lindeberg, 2003) Laptev and Lindeberg proposed a spatio-temporal extension to detect what they call "Space-Time Interest Points" (STIPs). STIPs are points which are relevant both in space and time. Theses points are especially interesting because they focus information initially contained in thousands of pixels on a few specific points which can be related to spatio-temporal events in the sequence. Typically, STIPs appear in articulated motions (walking, running or jumping person). However, it can be noted that constant motion of a corner does not produce any STIPs.

STIPs detection is performed by using the Hessian-Laplace matrix (Laptev, 2005) defined, for a pixel (x,y) at time t having intensity $I(x,y,t)$, by :

$$H(x,y,t) = \begin{pmatrix} \frac{\partial^2 I}{\partial x^2} & \frac{\partial^2 I}{\partial x \partial y} & \frac{\partial^2 I}{\partial x \partial t} \\ \frac{\partial^2 I}{\partial x \partial y} & \frac{\partial^2 I}{\partial y^2} & \frac{\partial^2 I}{\partial y \partial t} \\ \frac{\partial^2 I}{\partial x \partial t} & \frac{\partial^2 I}{\partial y \partial t} & \frac{\partial^2 I}{\partial t^2} \end{pmatrix} \quad (1)$$

In order to highlight STIPs, different criteria have been proposed. As in (Laptev, 2005), we have chosen the extension of the Harris corner function, called "*salience function*", defined by:

$$R(x,y,t) = \det(H(x,y,t)) - k * trace(H(x,y,t))^3 \quad (2)$$

where k is a parameter empirically adjusted. STIP correspond to high values of the salience function.

We are make tests for different values of the standard deviations σ_s and σ_t. These tests highlight the impact of Gaussian filters: when the values of σ_s and σ_t are low, the number of STIPs increases, but the good detection rate decreases. On the contrary, when the values ofσ_s and σ_t are high, the number of STIPs decreases and good detection rate increases up the 100%. However, the settings corresponding to a 100% rate provide a too small number of STIPs. Finally, a good compromise is $\sigma_s = 1.5$ and $\sigma_t = 1.5$. Although there are methods to make an automatic adjustment, we preferred to define them manually in order to optimize computation time.

3.2 Properties

STIP properties are well known particularly the relative stability with respect to geometric transformations. In our application, we lay interest in some specific properties, such as the robustness of STIPs against impulsive noise and contrast modification.

3.2.1 Low/High Contrast and Noise

An analysis of the effects of image quality on the STIP detection has also been done. Two situations were examined: contrast modifications and noise addition. The noise that were used is an impulsive noise because it is the most difficult type of noise relative to interest point detection. Table 1 shows the number of STIPs obtained for different contrast and noise conditions.

Table 1: Influence of contrast and impulse noise.

Contrast	50	75	100	125	150	175
STIP	1	2	29	64	68	127

a) Contrast influence

Pow	0	20	20	50	50	50
Intensity	0	20	50+	20	50	70+
STIP	29	29	33	49	78	126

b) Noise influence

80 sequences of video synthesis and athletics jump
$k = 0,04$, $\sigma_s = \sigma_t = 1.5$,salience threshold $= 150$

The evaluation is performed by observing the variations of the number of STIPs compared with the initial situation (no contrast modification and no noise : 29 STIPs by frame). It can be noticed that the STIP detection is very sensitive to contrast modification. On the contrary, the number of STIPs is relatively stable with respect to impulse noise.

3.2.2 Video Compression

The last criterion that influences STIPs generation is the compression factor of the video. Indeed, as a result of compression, straight lines show an aliasing which may, under certain circumstances, be perceived as angles (Clarke, 1995). This change causes the generation of STIPs.

Table 2: Influence of MPEG2 factor compression : average number of STIP by frame.

Compression factor (%)	10	20	30	40	50
STIP by frame (nb)	29	29	30	38	44
Compression factor (%)	60	70	80	90	100
STIP by frame (nb)	51	62	77	90	118

80 sequences of synthesis videos
and athletics jump
$k = 0,04$, $\sigma_s = \sigma_t = 1.5$
and salience threshold $= 120$

Table 2 shows the influence of MPEG2 compression factor on the number of generated STIPs. It is important to note that the sequence with square has not generated false positives. Indeed, no aliasing has occurred. These results show that the compression factor has an important influence past the threshold of

30% compression. In order not to disturb the results, it is necessary to ensure that the sequences used are not compressed beyond this threshold.

4 DETECTION OF MOVING OBJECTS

4.1 Principle

There are many methods for the detection of moving objects based on motion detection (Giai-Checa et al., 1993), the segmentation (Bugeau, 2007), the difference between successive images, etc. STIPs can be used for moving object detection. However, it only works if the object has a non regular motion, as STIPs correspond to second order variation both in space and time.

4.2 Experimental Evaluation

In athletic jumps or in animation movies, such type of motions occurs frequently, and generally corresponds to objects or persons which have an important role in the scene. Tests are performed according to the classical Precision/Recall criteria. The validation has been manually obtained in the following way:

- true positive: at least one STIP within an interesting moving object;
- false positive: at least one STIP within a non-interesting moving object;
- false negative: no STIP within an interesting moving object.

Table 3 shows that we obtained very good results using STIPs as interesting object detectors, even if there are several moving objects within the same frame.

Table 3: Object detection performances.

	Precision	Recall
animation movie	0.99	0.91
athletic movie	0.99	0.95

20 sequences of long jump (duration: 2120 frames) and 500 frames from the animated movie "Le Moine et le Poisson"
$k = 0,04$, $\sigma_s, \sigma_t = 1.5$,salience threshold $= 120$

To conclude, we can stress that the STIPs have a large enough performance to locate moving objects. Plus they will present "corner" if the number of points is important. A function determining the focus of these points can then define the approximate position of moving objects and make tracking.

5 DETECTION OF MOVEMENT CHANGES

5.1 Principle

In (Laganière et al., 2008) the activity level is defined within a video as the number of pixels altering their characteristics between two images. As a consequence, he proposes to define an activity function by the number of detected STIPs within each frame. A high (respectively low) value reflects a strong (respectively weak) activity. Moreover, the time evolution of this activity may contain some interesting information from a semantic point of view. Particularly, local maxima of this activity function are generally related to important events in the sequence. This is why we used this strategy to detect the different phases in movement. The hypothesis is that a local maxima of the activity function is related to a significative change in the non constant motion, and must correspond to a transition between two phases of a movement (for example, in a jump : running phase, ascending flight phase, descending flight phase, etc.).

5.2 Realization

Given that the activity function is generally noisy, it is first smoothed through the use of a mean filter with a filter size of 11. Let's denote $a_{filt}(t)$ the filtered activity function. Then the we look for local maxima of $a_{filt}(t)$ satisfying the following condition:

$$0.8 \times a_{filt}(t-\alpha) \leq a_{filt}(t) \leq 0.8 \times a_{filt}(t+\alpha) \quad (3)$$

with α accounting for the temporal extent determined by σ_t.

5.3 Experimental Evaluation

We used twenty sequences of different types of jumps (high jump, pole vault, long jump and triple jump) for test. In athletic jumps, such sequences generally contain a single dominant time event. The evaluation is a comparison between ground truth and detected transitions. As the transition location is not always accurate, we accepted a tolerance on the transition location. This tolerance depends on the kind of jump. Let's note that we used the same parameter set for all the sequences.

Table 4 shows the obtained results. Globally, the transitions are correctly detected with an accuracy between 3 and 10 images.
Precision and recall are relatively high. The least satisfying performances are obtained with the triple jump. This is probably due to the camera motion which is more complex for this type of jump.

Table 4: Detection of significant changes in movement.

	Precision	Recall	Tolerance
long jump	0.93	0.92	$\pm 3 frames$
high jump	0.92	0.88	$\pm 3 frames$
triple jump	0.81	0.71	$\pm 5 frames$
pole vault	0.84	0.85	$\pm 10 frames$

20 sequences of long jump (2120 frames)
$k = 0,04$, $\sigma_s = \sigma_t = 1.5$, salience threshold $= 120$

6 DISCUSSION

The proposed tool, that is STIPs, shows convincing results for the detection of moving objects and for the detection of significant changes in videos. However, it has some limitations. The first limitation comes from the setting. Indeed, the σ_s and σ_t parameters are difficult to adjust and the settings suggested in this analysis may be less effective in videos with very different characteristics. The second limitation relies on the conditions necessary shooting and the video quality (noise, contrast, compression), especially in the case of captured video in real time. These constraints can be problematic if one wishes to use this tool on videos from Web or stream videos real time. In this case, it will probably be necessary to make a pre-processing of contrast adjustment and / or noise filtering. The last limitation deals with reliability. The proposed assessments were performed on data which the events and movements were actually visible for. In the case of movement of which speeds are low or constant (for object detection) or in the case of movement which changes are not large enough (to detect change), there is no doubt that performance will be lower than proposed. Despite these limitations, the tool can be improved in many ways, this time to load very low.

7 CONCLUSIONS

In this paper, we proposed to use STIPs for video analysis. First, we examined some STIPs specific properties related to our applications. Thus, we showed that STIPs detection is sensitive to factor compression, parameter settings, specifically the variances of the gaussian filters, and intensity contrast. Conversely, STIPs detection is relatively robust against shooting condition variations and impulsive noise. Second, we used STIPs to detect moving objects in three different types of videos : synthesis videos for qualification, athletic jumps and animated movie for evaluation. The results we got were satisfying. In the specific case of athletic videos, we also resorted to STIPs to detect the transitions between the different phases of a jump, which provided good results too. The next step of this work will be to find out an adaptive setting of the most sensitive parameters.

REFERENCES

Bugeau, A. (2007). *Dtection et suivi d'objets en mouvement dans des scnes complexes, application la surveillance des conducteurs*. PhD thesis, IRISA.

Clarke, R. (1995). Digital compression of still images and video. *London : Academic press*, pages 285–299.

Giai-Checa, B., Bouthemy, P., and Vieville, T. (1993). Detection d'objets en mouvement. Technical Report INRIA-RR - 1906, INRIA.

Harris, C. and Stephens, M. (1988). A combined corner and edge detector. *In Alvey Vision Conference*.

Laganière, R., Bacco, R., Hocevar, A., Lambert, P., Païs, G., and Ionescu, B. (2008). Video summarization from spatio-temporal features. *ACM*.

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

Laptev, I. and Lindeberg, T. (2003). Space-time interest points. *ICCV'03*, pages 432–439.

Ramasso, E. (2007). *Reconnaissance de squences d'tats par le Modle des Croyances Transfrables et application l'analyse de vidos d'athltisme*. PhD thesis, University Joseph Fourier of Grenoble.

Schmid, C., Mohr, R., and Bauckhage, C. (2000). Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172.