Journal of Computer Science 10 (5): 828-839, 2014
ISSN: 1549-3636
© 2014 Science Publications
doi:10.3844/jcssp.2014.828.839 Published Online 10 (5) 2014 (http://www.thescipub.com/jcs.toc)

MOTION DETECTION USING THE SPACE-TIME INTEREST POINTS

Insaf Bellamine and Hamid Tairi

Department of Computer Science, Sidi Mohamed Ben Abdellah University, Fes, Morocco

Received 2013-07-18; Revised 2013-11-18; Accepted 2014-01-17

ABSTRACT

Space-Time Interest Points (STIP) are among all the interesting features which can be extracted from videos; they are simple, robust and they allow a good characterization of a set of regions of interest corresponding to moving objects in a three-dimensional observed scene. In this study, we show how the resulting features often reflect interesting events that can be used for a compact representation of video data as well as for tracking. For a good detection of moving objects, we propose to apply the algorithm of the detection of spatiotemporal interest points on both components of the decomposition which is based on a Partial Differential Equation (PDE): A geometric structure component and a texture component. Proposed results are obtained from very different types of videos, namely sport videos and animation movies.

Keywords: Space-Time Interest Points, Structure-Texture Image Decomposition, Motion Detection

1. INTRODUCTION

The motion analysis is a very active research area, which includes a number of issues: Motion detection, optical flow, tracking and human action recognition.

To detect the moving objects in an image sequence is a very important low-level task for many computer vision applications, such as video surveillance, traffic monitoring, video indexing, recognition of gestures, analysis of sport-events, Sign language recognition, mobile robotics and the study of the objects' behavior (people, animals, vehicles).

In the literature, there are many methods to detect moving objects, which are based on: Optical flow (Jodoin and Mignotte, 2008), difference of consecutive images (Galmar and Huet, 2007), Space-Time Interest Points (Laptev, 2005) and modeling of the background (local, semi-local and global) (Nicolas, 2007).

Our method consists to use the notion of Space-Time Interest Points; these ones are especially interesting because they focus information initially contained in thousands of pixels on a few specific points which can be related to spatiotemporal events in an image. Laptev and Lindeberg (2003) were the first who proposed STIPs for action recognition (Laptev, 2005), by introducing a space-time extension of the popular Harris detector. They detect regions having high intensity variation in both space and time as spatio-temporal corners. The STIP detector of (Laptev, 2005) usually suffers from sparse STIP detection. Later, several other methods for detecting STIPs have been reported (Dollar *et al*., 2005). Dollar *et al*. (2005) improved the sparse STIP detector by applying temporal Gabor filters and selecting regions of high responses. Dense and scale-invariant spatio-temporal interest points were proposed by Willems *et al*. (2008). An evaluation of these approaches has been proposed in (Wang, 2009).

Our approach also uses Aujol Algorithm (Aujol, 2004), this one decomposes the image f into a structure component u and a texture component v, (f = u + v). The notion of Structure-Texture Image Decomposition is essential for understanding and analyzing images depending on their content.

In this study, we propose to apply the algorithm of the detection of spatiotemporal interest points for a good detection of moving objects on both components of the decomposition: A geometric structure component and a texture component. Proposed results

Corresponding Author: Insaf Bellamine, Department of Mathematics and Computer, Sidi Mohamed Ben Abdellah University, LIIAN, Fes, Morocco

are obtained from different types of videos, namely sport videos and animation movies.

1.1. Space-Time Interest Points

The idea of interest points in the spatial domain can be extended into the spatio-temporal domain by requiring the image values in space-time to have large variations in both the spatial and the temporal dimensions. Points with such properties will be spatial interest points with a distinct location in time corresponding to the moments with non-constant motion of the image in a local spatiotemporal neighborhood (Laptev and Lindeberg, 2003). These points are especially interesting because they focus information initially contained in thousands of pixels on a few specific points which can be related to spatiotemporal events in an image.

Laptev and Lindeberg (2003) proposed a spatio-temporal extension of the Harris detector to detect what they call "Space-Time Interest Points", denoted STIP in the following.

Detection of Space-Time Interest Points is performed by using the Hessian-Laplace matrix H (Laptev, 2005), which is defined by Equation (1):

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

I (x, y, t) is the intensity of the pixel (x, y) at time t.

As with the Harris detector, a Gaussian smoothing is applied both in spatial domain (2D filter) and temporal domain (1D filter) Equation (2):

$$g(x,y,t;\sigma_s^2,\sigma_t^2) = \frac{\exp\left(-\frac{x^2+y^2}{2\sigma_s^2} - \frac{t^2}{2\sigma_t^2}\right)}{\sqrt{(2\pi)^3\sigma_s^4\sigma_t^2}} \quad (2)$$

The two parameters (σ_s and σ_t) control the spatial and temporal scale. As in (Laptev, 2005), the spatio-temporal extension of the Harris corner function, entitled "salience function", is defined by Equation (3):

$$R(x,y,t) = \det(H(x,y,t)) - k \times \operatorname{trace}(H(x,y,t))^3 \quad (3)$$

where, k is a parameter empirically adjusted at 0.04, det is the determinant of the matrix H and trace is the trace of the same matrix.

STIP correspond to high values of the salience function R and they are obtained by using a thresholding step.

1.2. Tests

In what follows, we represent some examples of clouds of space-time interest points detected in these sequences (**Fig. 1**):

- Sport video: (karate's fight) lasts for 2 min and 49 sec with 200 images and the size of each image frame is 400 by 300 pixels
- Animation movie: Lasts for 3 min and 32 sec with 230 images and the size of each image frame is 352 by 288 pixels
- KTH dataset (Schuldt *et al*., 2004): It was provided by Schuld-tetal. Schuldt *et al*. (2004) and is one of the largest public human activity video dataset. It contains six types of actions (boxing, hand clapping, hand waving, jogging, running and walking) performed by 25 subjects in four different scenarios including in door, outdoor, changes in clothing and variations in scale. Each video clip contains one subject performing a single action. Each subject is captured in a total of 23 or 24 clips, giving a total of 599 video clips. Each clip has a frame rate of 25Hz and lasts be-tween10 and 15 s. The size of each image frame is160 by 120 pixels. Two Examples of the KTH dataset are shown in **Fig. 1**

We chose the value of 1.5 for the two standards deviation σ_s and σ_t, according to a study that was done by Simac (2006).

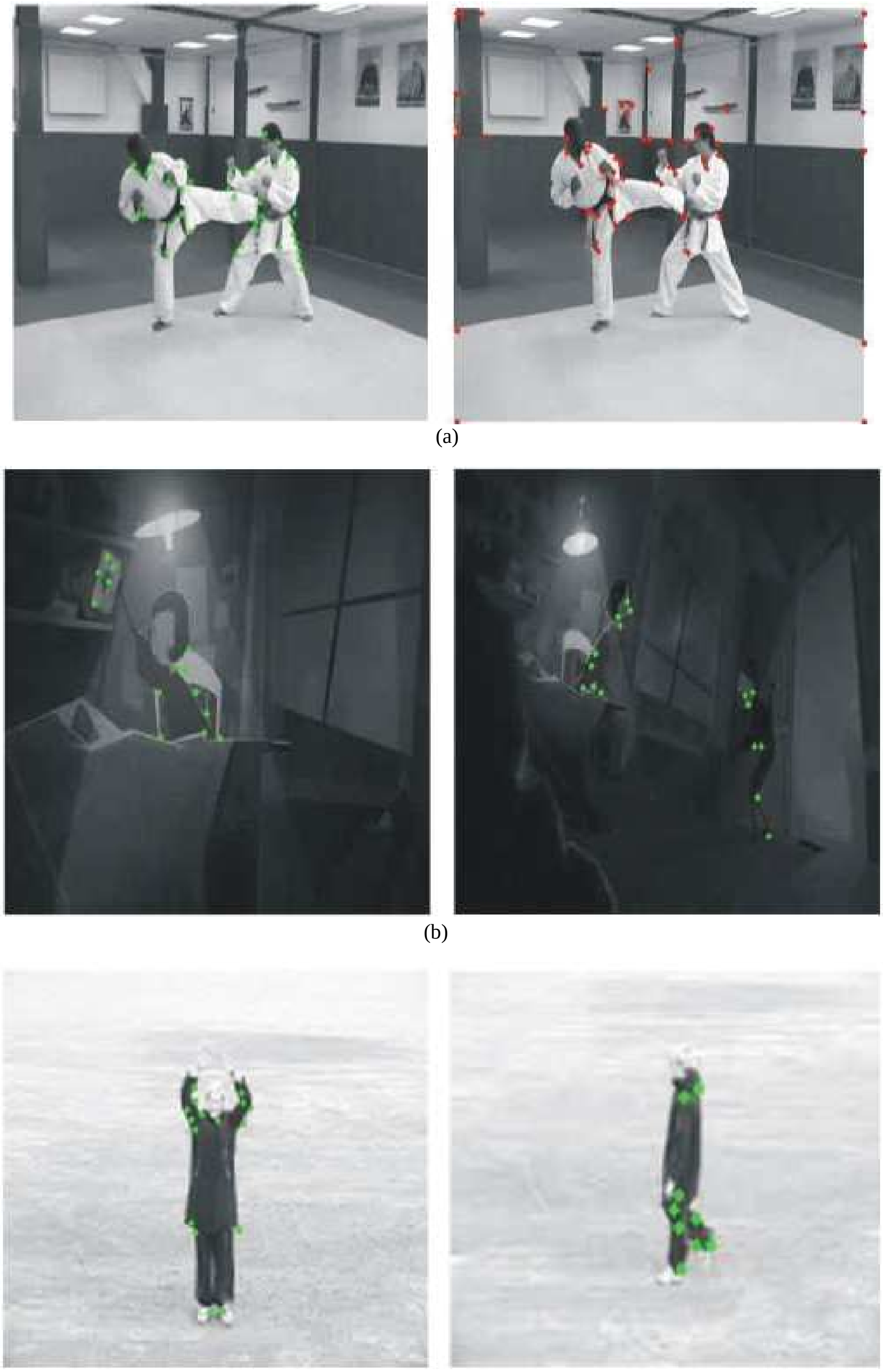
1.3. Structure-Texture Image Decomposition

Let f be an observed image which contains texture and/or noise. Texture is characterized as repeated and meaningful structure of small patterns.

Noise is characterized as uncorrelated random patterns. The rest of an image, which is called cartoon, contains object hues and sharp edges (boundaries). Thus an image f can be decomposed as:

$$f = u + v$$

where, u represents image cartoon and v is texture and/or noise.



(a)

(b)

(c)

Fig. 1. Examples of clouds of space-time interest points. We have σs = 1. 5 and σt = 1. 5 with k = 0, 04. In each frame, the red points represent the extracted Space Interest Points (Harris and Stephens (1988) detector) and the green points are the extracted Space-Time Interest Points (Laptev and Lindeberg, 2003 detector) (a) (image t = 47); karate's fight (image t = 54) (b) Animation movie (Trois petits points) (Images t = 25 et t = 30) (c) KTH dataset: Hand waving, walking (Images t = 20 and t = 60)

Decomposing an image f into a geometric structure component u and a texture component v is an inverse estimation problem, essential for understanding and analyzing images depending on their content.

Many image decomposition models have been proposed, those based on the total variation as the model of the ROF minimization was proposed by Rudin *et al.* (1992), this model has demonstrated its effectiveness and has eliminated oscillations while preserving discontinuities in the image, it has given satisfactory results in image restoration (Cha *et al*., 2001; Rudin and Osher, 1994) because the minimization of the total variation smooths images without destroying the structure edges.

In recent years, several models based on total variation, which are inspired by the ROF model, were created (Aujol, 2004; Gilles, 2006). In the literature there is also another model called Meyer (2001) that is more efficient than the ROF model. Many algorithms have been proposed to solve numerically this model. In the following, we represent the most popular algorithm, Aujol algorithm.

1.4. The Aujol Algorithm

Aujol (2004), propose a new algorithm for computing the numerical solution of the Meyer's model. An image f can be decomposed as f = u + v, where u represents a geometric structure component and v is a texture component. The algorithm of Aujol is represented as follows:

Step 1: Initialisation

$u_0 = v_0 = 0$

Step 2: Iterations

$v_{n+1} = P_{G_\mu}(f - u_n)$

$u_{n+1} = f - v_{n+1} - P_{G_\lambda}(f - v_{n+1})$

where P_{G_μ} and P_{G_λ} are the operators of projections (Chambolle, 2004)

Step 3: Stop condition

We stop the algorithm if

$\max^{f0}(|u_{n+1} - u_n|, |v_{n+1} - v_n|) \leq \varepsilon$

Or if it reaches a maximum of iterations required

The regularization of parameters (λ and μ) play an important role in the decomposition of the image: λ controls the amount of the residue $f - u - v$ and μ influences on the texture component v. The choice of λ does not pose a problem. It gives it just a small value, but the μ parameter is not easy to adjust.

Well the Aujol algorithm can extract the textures in the same way as the Osher-Vese algorithm. Moreover, this algorithm has some advantages if it is compared to Osher-Vese (Gilles, 2006):

- No problem of stability and convergence
- Easy to implement (requiring only a few lines of code)

1.5. Decomposition Results

Let f the image to decompose, then f can be written as follows: f = u + v.

The Structure-Texture Image Decomposition has been applied on the Barbara image of size 512×512. The result of the decomposition using the parameters (μ = 1000, λ = 0.1) is shown in **Fig. 2**.

The program was run in a PC with a 2.13 GHz Intel core (TM) i3 CPU with 3 GB RAM.

Decomposing an image f into a geometric structure component and a texture component requires relatively low computation time **Fig. 3**, which gives us the opportunity to use this decomposition in motion detection in real time.

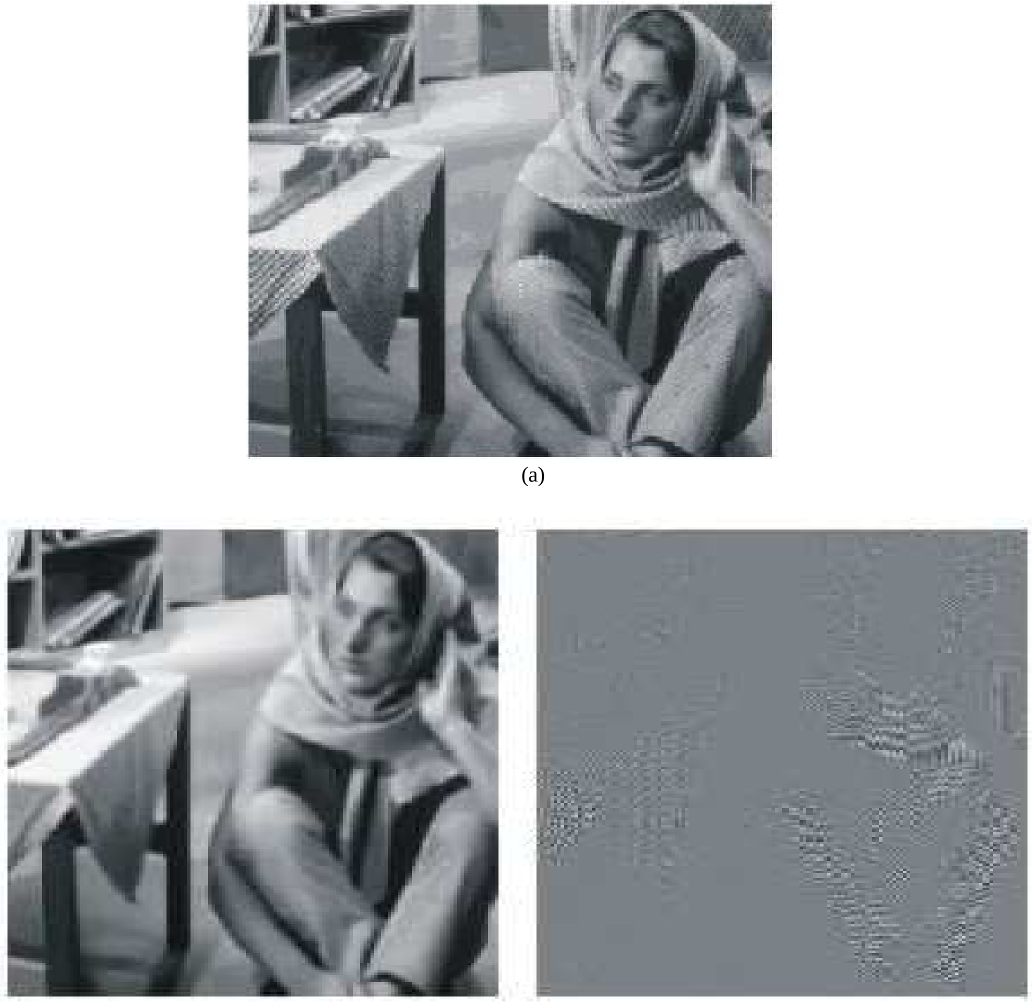
1.6. Proposed Approach

The most famous algorithm to detect Space-Time Interest Points is that of Laptev; however we can reveal three major problems when a local method is used:

- Texture, Background and Objects that may influence the results
- Noisy datasets such as the KTH dataset, which is featured with low resolution,strong shadows and camera movement that renders clean silhouette extraction impossible
- Features extracted are unable to capture smooth and fast motions and they are sparse. This also explains why they generate poor results

However, to overcome the three problems, we propose a technique based on the space-time interest points and which will help to have a good detection of moving objects and even reduce the execution time by proposing a parallel algorithm (**Fig. 4**).

A complex scene can contain various information (noise, textures, shapes, background...), these ones influence the detection of moving objects. Our goal is to apply the algorithm of the detection of spatiotemporal interest points on both components of the decomposition: A geometric structure component and a texture component.



(a)

(b)

(c)

Fig. 2. Decomposition with Aujol (2004) Algorithm: (a) Barbara image (b) the u component (c) the v component

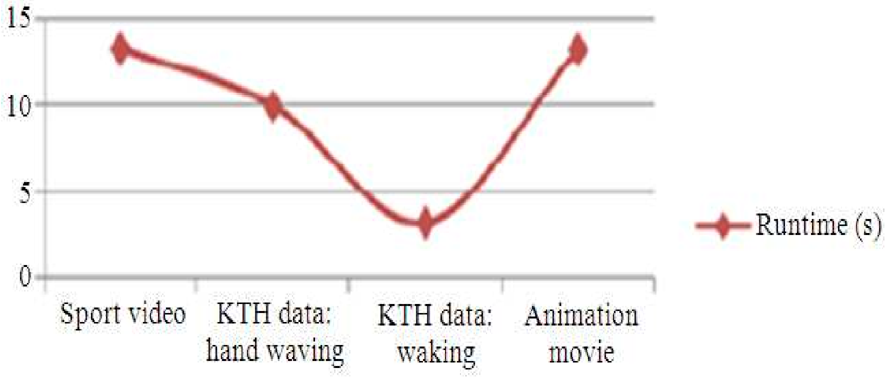


Fig. 3. Extraction time of structure-texture image decomposition

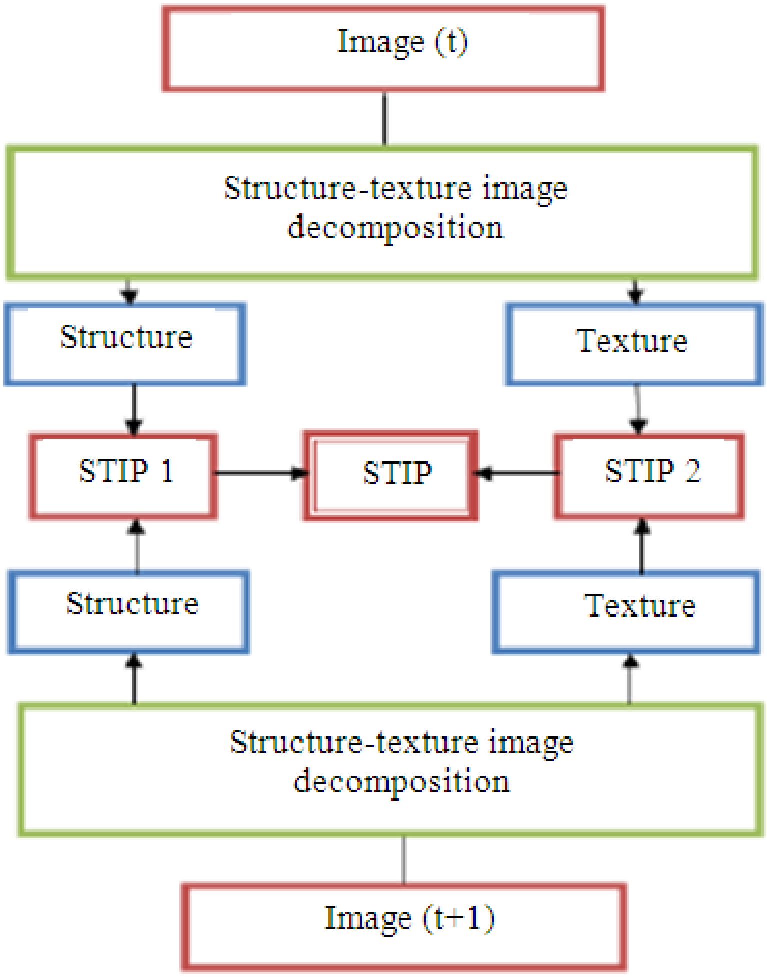


Image (t)
Structure-texture image decomposition
Structure
Texture
STIP 1
STIP
STIP 2
Structure
Texture
Structure-texture image decomposition
Image (t+1)

Fig. 4. The adapted Laptev algorithm

Let STIP denote the final Space-Time Interest Points, STIP1 denotes the extracted Space-Time Interest Points between the components of the decomposition of structure1 and structure 2, also STIP2 denotes the extracted Space-Time Interest Points using Texture1 and Texture 2 components.

Our new space-time interest points will be calculated as the following:

$$STIP = STIP1 \cup STIP2 \quad (4)$$

1.7. Enhanced Laptev Algorithm

In the first step, the Structure-Texture Image Decomposition method is applied to the two consecutive frames of the video sequence. In the second step, two processes based on structures (Structure1 and Structure2) and textures (Texture1 and Texture2) had to be made equivalent to the two matching modes. Each process provides, as output result, the STIP1 extracted from the first mode and the STIP 2 extracted from the second mode. For the last step, the final STIP are computed by the Equation (4). **Figure 5** shows the steps of the enhanced Laptev algorithm. The results illustrated in **Fig. 6**, show that we come to locate moving objects in both sequences, we note that: The objects moving (the two players) are detected with our approach, for against just one player who has been detected with Laptev detector (**Fig. 1a**).

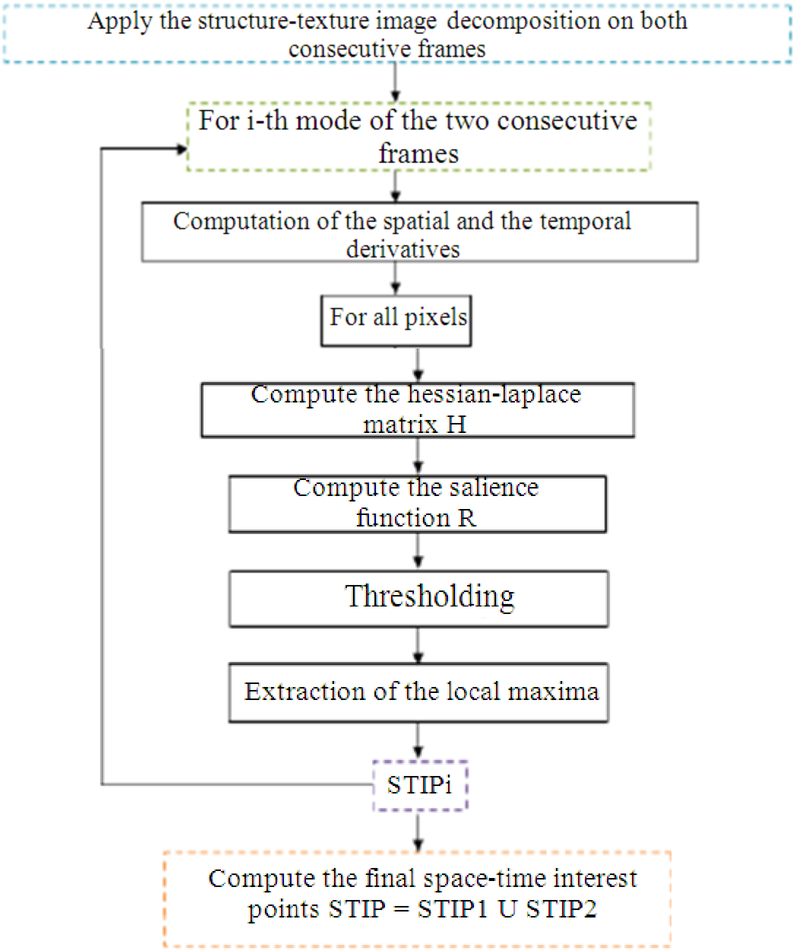
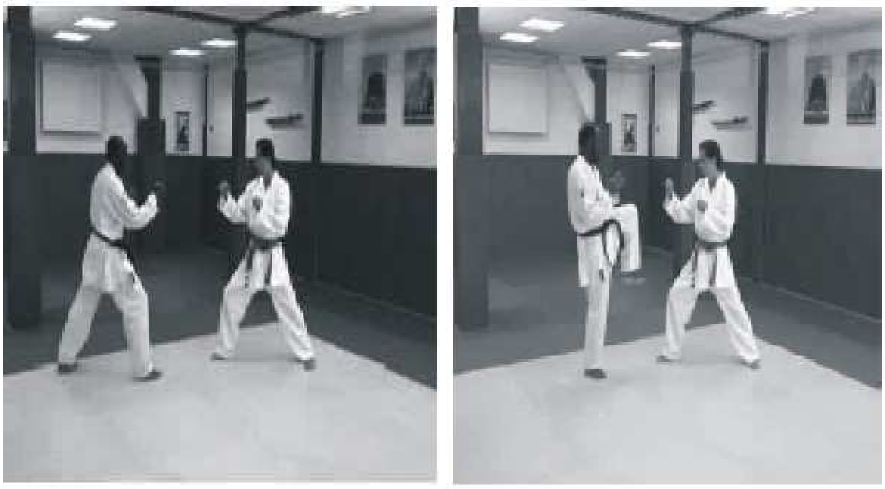
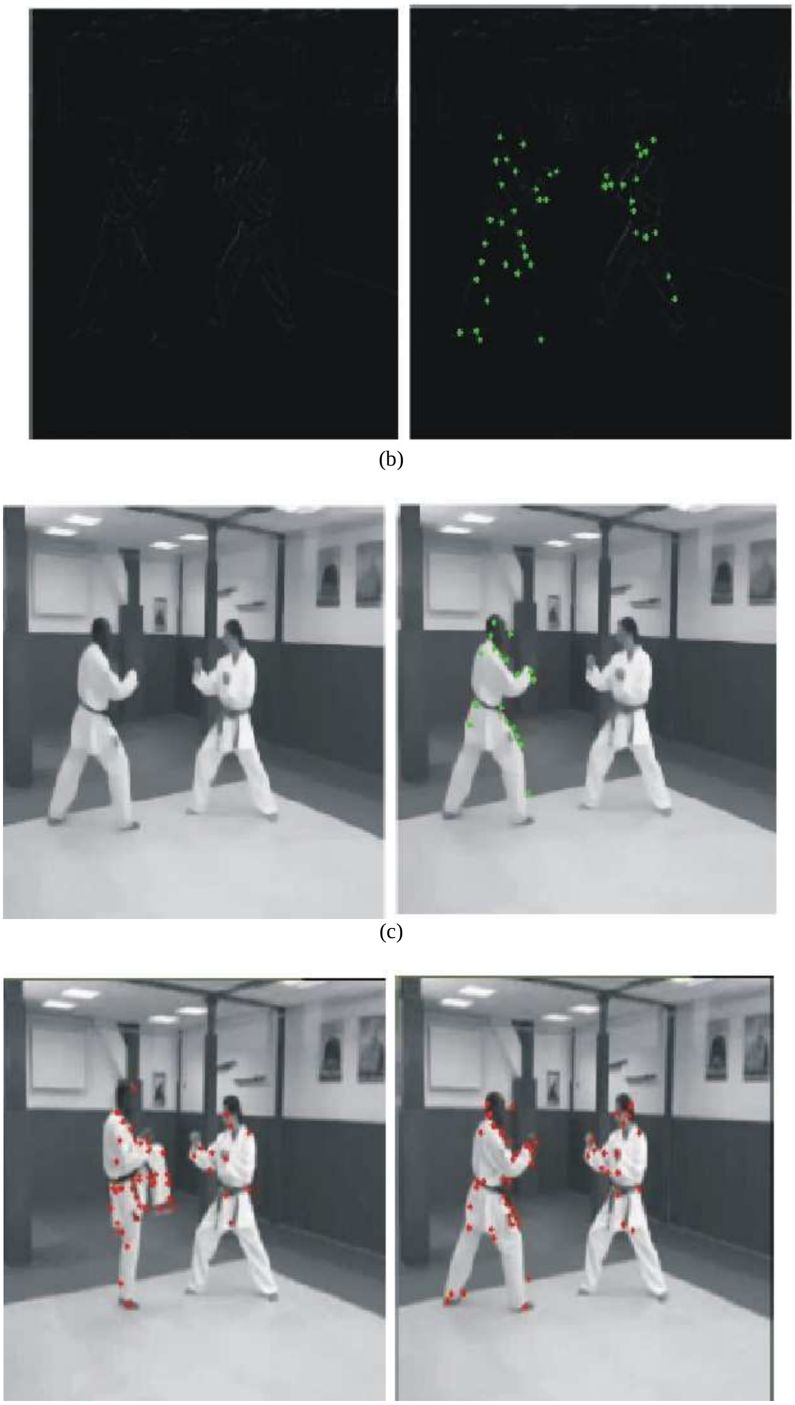

Apply the structure-texture image decomposition on both consecutive frames
For i-th mode of the two consecutive frames
Computation of the spatial and the temporal derivatives
For all pixels
Compute the hessian-laplace matrix H
Compute the salience function R
Thresholding
Extraction of the local maxima
STIPi
Compute the final space-time interest points STIP = STIP1 U STIP2

Fig. 5. Enhanced Laptev algorithm



(a)



(b)

(c)

(d)

Fig. 6. Examples of clouds of space-time interest points, in each frame, we use the parameters (σt = 1. 5, k = 0, 04 and σs = 1. 5) (a) Image (t = 44)/Image (t = 45) (b) The green points are the detected Space-Time Interest Points on the texture components (c) The green points are the detected Space-Time Interest Points on the structure components (d) The red points represents the extracted Space-Time Interest Points with Our Proposed Approach

2. EXPERIMENTAL RESULTS

Let f be an observed image which contains texture and/or noise. The rest of an image, which is called geo-metric structure, contains object hues and sharp edges (boundaries). Thus an image f can be decomposed as f = u + v, where u represents structure image and v is texture and/or noise.

We propose to apply the algorithm of the detection of spatiotemporal interest points on both components of the decomposition: A geometric structure component and a texture component. In what follows, we represent some examples of clouds of space-time interest points detected in the first sequence (**Fig. 1**).

2.1. Comparison and Discussion

In order to correctly gauge performance of our proposed approach, we will proceed with a Comparative study to use the mask of the moving object and the precision.

The mask of the moving object is obtained by the Markovian approach (Luthon, 2001) (**Fig. 7**). We distinguish two cases:

- True positive: The space-time interest point is in the mask, so it is on a moving object
- False positive: The space-time interest point isn't in the mask, so it isn't on a moving object

For each moving object, we have a number of the space-time interest points detected in the moving object (NTP) and a number of the space-time interest points extracted off the moving object (NFP).

The precision is defined by Equation (5):

$$\text{Precision} = \frac{\text{NTP}}{\text{NTP} + \text{NFP}} \quad (5)$$

NTP is the number of the true positives (good detections), NFP the number of the false positives (false detections).

The test is performed on four examples of sequences and gives the following results.

The results, illustrated in **Fig. 8**, show that the real moving objects (the two players, the cars, the truck and branches of the tree) are better detected with the proposed approach than with (Laptev and Lindeberg, 2003) detector.

Still, the proposed approach is much less sensitive to noise and the reconstruction of the image, it also extracts the densest features (**Fig. 9**).

The results, illustrated in **Table 1**, show that our approach allows a good detection of moving objects.

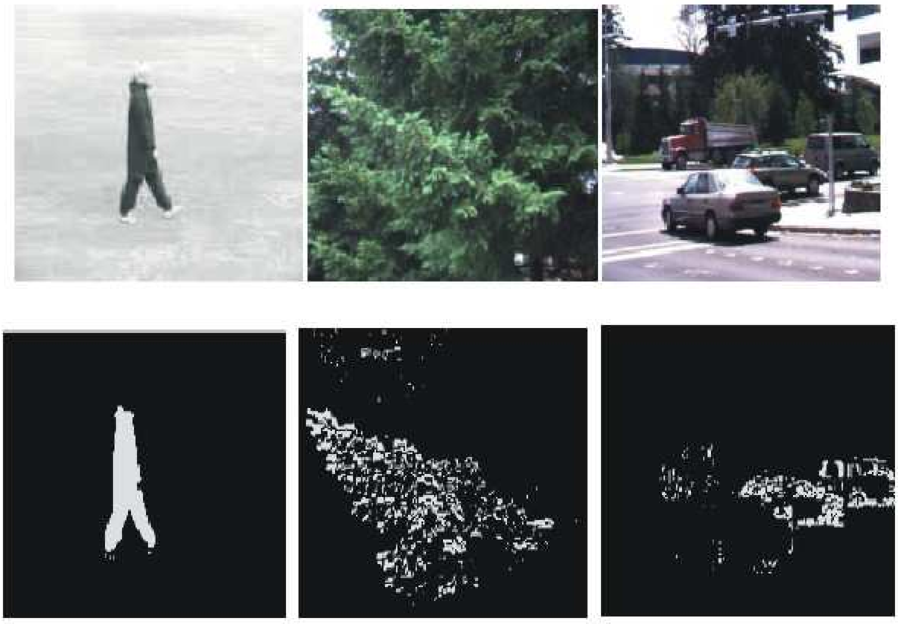
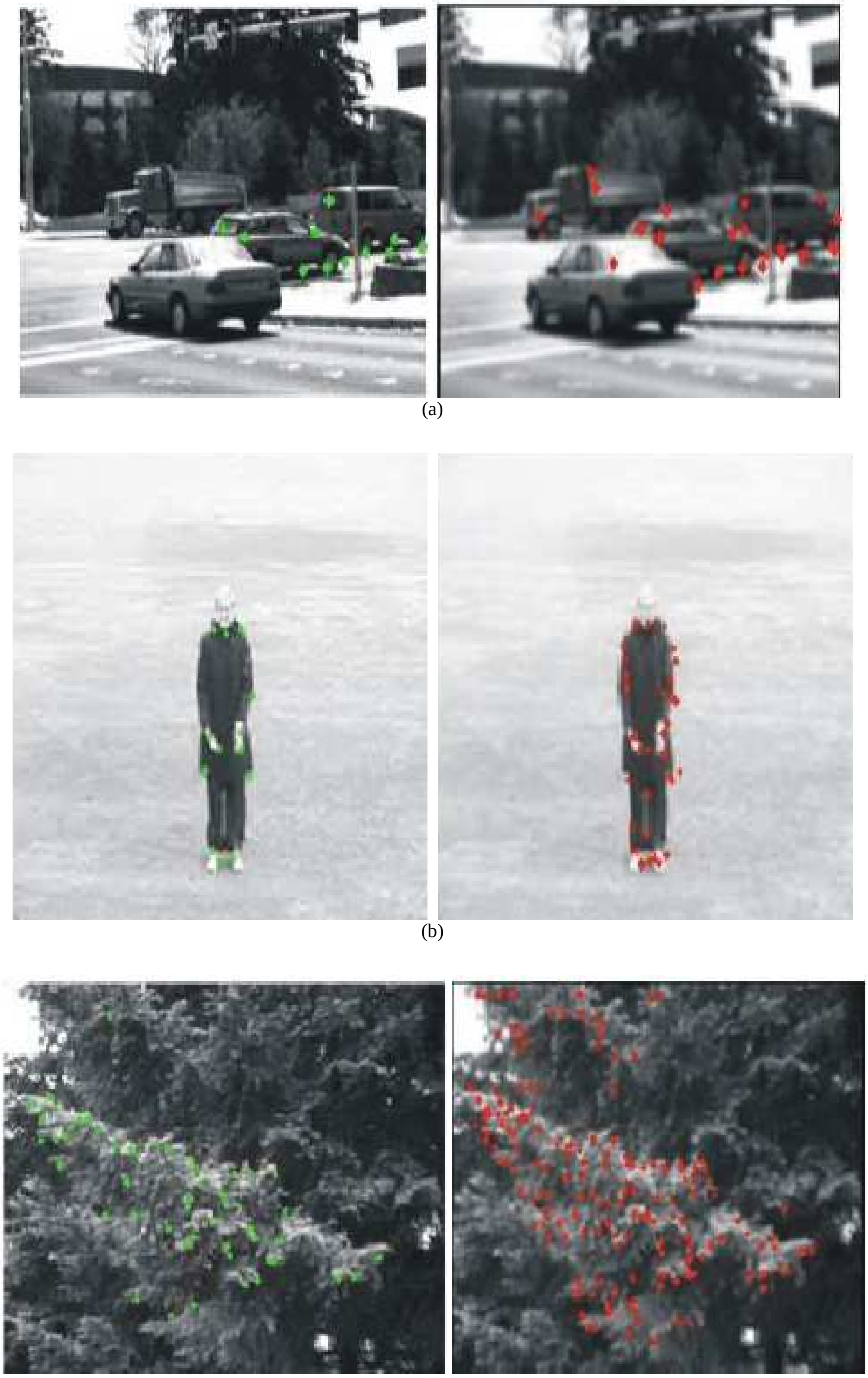


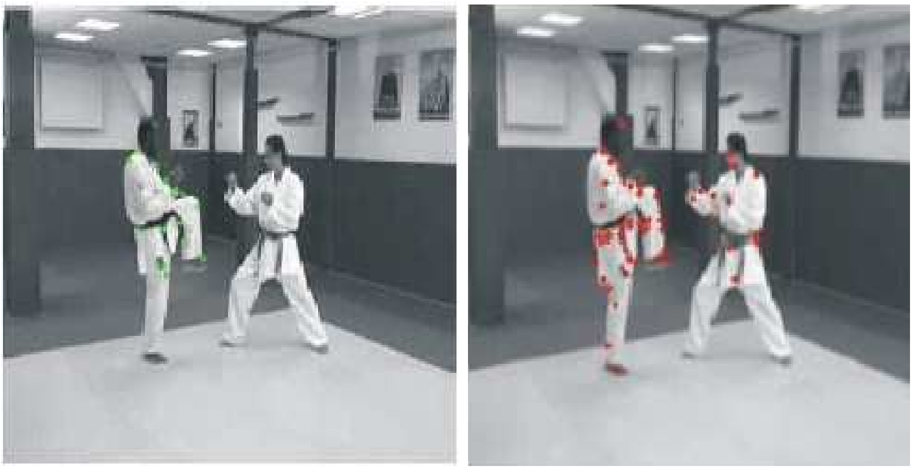
Fig. 7. Example of moving objects' masks



(a)

(b)

(c)



(d)

Fig. 8. Examples of clouds of space-time interest points. We have σs = 1. 5 and σt = 1.5 with k = 0, 04. In each frame, the green points represent the extracted Space-Time Interest Points by Laptev and Lindeberg (2003) detector and the red points are the extracted Space-Time Interest Points by Laptev and Lindeberg (2003) detector and the red points are the extracted Space-Time Interest Points by Our Approach (a) Urban transport (b) Hand waving (c) Tree (d) Karate

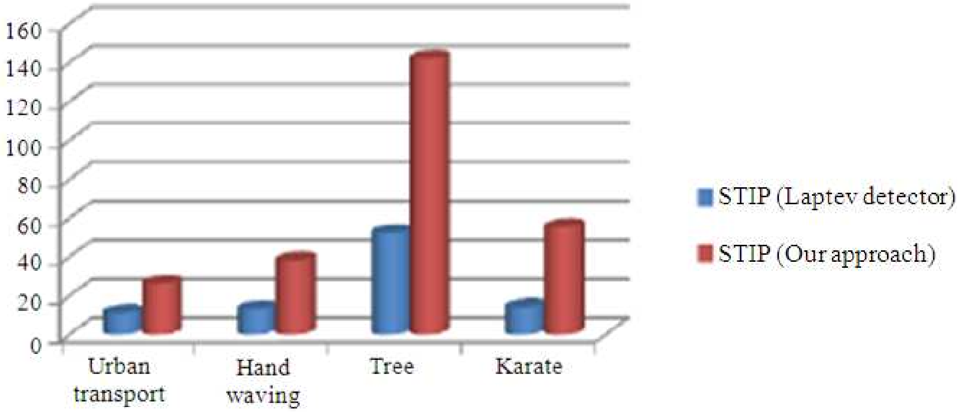


Fig. 9. Number of Space-Time Interest Points extracted in each frame

Table 1. Compared results

Precision videos	Precision (Laptev detector)	(Our approach)
Urban transport	81%	89%
Hand waving	92%	93%
Tree	96%	98%
Karate	92%	96%

3. CONCLUSION

In the experimental part, the results are obtained from very different types of videos, namely sport videos and animation movies.

Our Approach improved the sparse STIP detector by applying the algorithm of the detection of spatiotemporal on both components of the decomposition: A geometric structure component and a texture component. This Approach is less sensitive to the noise effects and the parallel implementation requires low computation time.

4. REFERENCES

Aujol, J.F., 2004. Contribution à l’analyse de textures en traitement d’images par méthodes variationnelles et équations aux dérivées partielles. Thèse de Doctorat, Université de Nice Sophia Antipolis

Cha, T.F., S. Osher and J. Shen, 2001. The digital TV filter and nonlinear denoising. IEEE Trans. Image Process., 10: 231-241. DOI: 10.1109/83.902288

Dollar, P., V. Rabaud, G. Cottrell and S. Belongie, 2005. Behavior recognition via sparse spatio-temporal features. Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Oct. 15-16, IEEE Xplore Press, pp: 65-72. DOI: 10.1109/VSPETS.2005.1570899

Galmar, E. and B. Huet, 2007. Analysis of vector space model and spatiotemporal segmentation for video indexing and retrieval. Proceedings of the 6th ACM International Conference on Image and Video Retrieval, Jul. 9-11, ACM Press, Amsterdam, Netherlands, pp: 433-440. DOI: 10.1145/1282280.1282344

Gilles, J., 2006. Décomposition et détection de structures géométriques en imagerie. Thèse de Doctorat, Ecole Normale Supérieure de Cachan.

Jodoin, P.M. and M. Mignotte, 2008. Optical-flow based on an edge-avoidance procedure. Comput. Vis. Image Understand., 113: 511-531. DOI: 10.1016/j.cviu.2008.12.005

Laptev, I. and T. Lindeberg, 2003. Space-time interest points. Proceedings of the 9th IEEE International Conference on Computer Vision, Oct. 13-16, IEEE Xploer Press, Nice, France, pp: 432-439. DOI: 10.1109/ICCV.2003.1238378

Laptev, I., 2005. On space-time interest points. Int. J. Comput. Vis., 64: 107-123. DOI: 10.1007/s11263-005-1838-7

Luthon, F., 2001. Module de Traitement d'Image, Laboratoire d'Informatique de l'Universite de Pau et des Pays de l'Adour.

Meyer, Y., 2001. Oscillating Patterns in Image Processing and Nonlinear Evolution Equations. 1st Edn., American Mathematical Society, Providence, RI., ISBN-10: 0821829203, pp: 122.

Nicolas, V., 2007. Suivi d'objets en mouvement dans une séquence vidéo. Thèse de doctorat, Université Paris Descartes.

Rudin, L., S. Osher and E.E. Fatimi, 1992. Nonlinear total variation based noise removal algorithms. Physica D, 60: 259-268. DOI: 10.1016/0167-2789(92)90242-F

Rudin, L.I. and S. Osher, 1994. Total variation based image restoration with free local constraints. Proceedings of the IEEE International Conference Image Processing, Nov. 13-16, IEEE Xplore Press, Austin, TX., pp: 31-35. DOI: 10.1109/ICIP.1994.413269

Schuldt, C., I. Laptev and B. Caputo, 2004. Recognizing human actions: A local SVM approach. Proceedings of the 17th International Conference on Pattern Recognition, Aug. 23-26, IEEE Xplore Press, pp: 32-36. DOI: 10.1109/ICPR.2004.1334462

Simac, A., 2006. Modélisation et gestion de concepts, en particulier temporels, pour l'assistance à. Thèse de Doctorat, Université de Grenoble,

Wang, H., 2009. Evaluation of local spatio-temporal features for action recognition.

Willems, G., T. Tuytelaars and L.V. Gool, 2008. An efficient dense and scale-invariant spatio-temporal interest point detector. Proceedings 10th European Conference on Computer Vision, Oct. 12-18, Springer Berlin Heidelberg, Marseille, France, pp: 650-663. DOI: 10.1007/978-3-540-88688-4_48