

# On-line video recognition and counting of harmful insects

Ikhlef Bechar, Sabine Moisan, Monique Thonnat and François Bremond.

*EPI Pulsar, INRIA Sophia Antipolis-Mediterranee*

*firstName.lastName@sophia.inria.fr*

## Abstract

*This article is concerned with on-line counting of harmful insects of certain species in videos in the framework of in situ video-surveillance that aims at the early detection of prominent pest attacks in greenhouse crops. However, the video-processing challenges are numerous and they mainly concern the low spatial resolution and color contrast of the objects of interest in the videos, the outdoor issues and the video-processing which needs to be done in quasi-real time. Thus, we propose a solution which makes use of an efficient pattern recognition algorithm to extract the locations of the harmful insects of interest in a video, and when coupled with some video-processings a quick on-line vision solution is achieved, and overall systems sensitiveness and accuracy are substantially increased. The system has been tested off-line against many videos of the whitefly species (one potential harmful insect), recorded under various in situ conditions (light changes, shadows, presence of outlier objects, etc.), and it has shown acceptable performance in terms of accuracy versus computational time.*

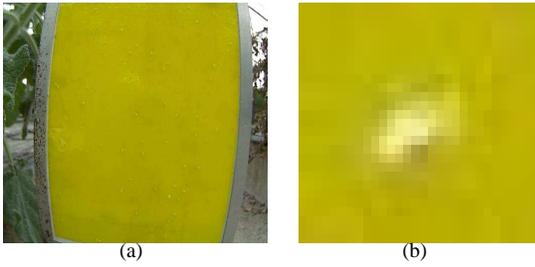
## 1. Introduction

Classically, pest monitoring is performed manually and relies heavily on the knowledge and the availability of a human expert for routinely screening every greenhouse crop in order to predict prominent pest attacks at the early stage, thereby manage to optimize the fighting operations that fall. However, with greenhouse cultures which are getting increasingly intensive motivated by the increasing need of feeding a growing population, along with the recent norms on the use of pesticides, hence an automation of the pest monitoring process is needed. On the other hand, computer vision has been successfully used in various real-life surveillance applications [4], and among its advantages from which one can benefit in order to automate the pest monitor-

ing process [1, 6], we can mention non-invasiveness, autonomy, and objectiveness. The idea is to equip a greenhouse crop with a network of video-cameras that will sense during daytime some tailored devices (sticky traps (cf. Fig.1(a))) that have the property (pertaining to their used color) to attract the insects of interest (but not only), and to fix them on their sticky surface permanently. An on-line video-processing makes it then possible to recognize the trapped insects of interest during daytime periods, and to describe statistically their spatiotemporal presence which is then used to predict a pest attack.

Despite the used big image resolution, however, the objects of interest (harmful insects) appear in the videos only like tiny and lowly contrasted spots with unclearly defined borders (see Fig.1(a)-(b)), making of their on-line recognition a big issue. Such a problem of extraction of spots from unspecific backgrounds is recurrent in various application contexts, and one distinguishes generally between two categories of approaches. In the first category of approaches, the objects of interest in an image (resp. a video) are represented directly in it by some bright spots plunged in a heavily cluttered background (e.g. fluorescent biological particles in videomicroscopy), and the challenge is to be able to separate them from the rest of the image (resp. the video) in order to answer some biological questions (see e.g. [3] and [5] for some existing approaches in biological imaging). In the second category of approaches encountered mainly in active vision, an image is firstly transformed into another image (generally a gray image) where the zones of interest in the original image will appear like bright structures which are then easier and/or quicker (mainly when real time constraint is considered) to extract by using a standard image processing technique like, for example, a local maxima extraction algorithm (see e.g. [2] for an application to facial features extraction in active vision).

Hence, our approach for the extraction of the insects of interest in videos, though sitting on the border be-

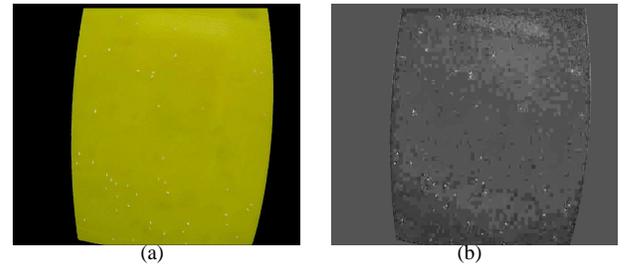


**Figure 1.** (a) An example of a typical video-frame (resolution:  $1280 \times 960$  px). The central yellow zone corresponds to the zone of the sticky trap, and the trapped harmful insects (whiteflies) correspond to the tiny white spots fixed on its surface ; (b) A zoom on the imagerie of one insect of interest in the video-frame.

tween the two above-mentioned approaches is, nevertheless, more biased towards the second category of approaches. Indeed, firstly, we propose to (linearly) transform each RGB video-frame into a gray image in such a way that the zones of the insects of interest in the original RGB frame appear like bright spots. Secondly, we propose to use a pattern recognition algorithm based on an approximate mathematical model about a bright spot of interest in the joint space-gray intensity domain in order to extract any zone of a harmful insect in the original video-frame. Some of the strengths of the proposed pattern extraction algorithm are its invariance under the rotation of an object of interest, its invariance under the adding of a constant to the image of an insect, and most of all, its dependence upon some parameters (namely the scale and the saliency parameters) that can be tuned (learned) accordingly off-line in such a way to achieve desired performance. Moreover, such an approach is significantly accelerated by coupling it with some video-processing algorithms (background subtraction and tracking).

## 2 Frame-wise detection of harmful insects

Initially, the zone of the sticky trap w.r.t. each video is extracted automatically once and for all from the first video-frame (cf. Fig.2(a)) by using some mere assumptions about similarity of color and compacity, hence all subsequent video-processings are performed in this zone of interest in a video. The algorithm for the recognition of the harmful insects of interest in a video that we propose proceeds in two main steps which are described in detail in the sequel.



**Figure 2.** (a) Projection of an RGB video-frame on the recognized zone of the sticky trap ; (b) Its RGB-into-gray transformation.

### 2.1 RGB-into-gray linear transformation

The first step consists in transforming each RGB video-frame into a gray image where the insects of interest will appear as much brighter as possible than the background so as to enhance their signal to noise ratio and to motivate the algorithm that we propose subsequently for their extraction from a video. This is motivated by the fact the harmful insects of the species of interest are characterized fundamentally by one color (e.g., the white color for the whitefly species as shown in (cf. Fig.1 and Fig.2), and the green color for the greenfly species (not shown in this article)), and so is also the background on which they lie. This is achieved by considering a linear transformation of the form  $I := t_r R + t_g G + t_b B$ , and estimating  $t_r$ ,  $t_g$  and  $t_b$  in such a way to maximize with respect to the linear coefficients  $t_r$ ,  $t_g$  and  $t_b$  the (SNR) ratio between the mean contrast over a sample of  $N_I$  insect intensities  $\mathcal{S}_I = \{(R_i, G_i, B_i); i = 1, \dots, N_I\}$  and the mean contrast over a sample of  $N_B$  background intensities  $\mathcal{S}_B = \{(r_j, g_j, b_j); j = 1, \dots, N_B\}$  as follows:

$$\frac{\sum_{i=1}^{N_I} (t_r R_i + t_g G_i + t_b B_i)^2}{\sum_{j=1}^{N_B} (t_r r_j + t_g g_j + t_b b_j)^2} \rightarrow \max_{t_r, t_g, t_b} \quad (1)$$

By rewriting the latter expression in matrix form as follows:  $\frac{t^T V_I t}{t^T V_B t}$  with  $t := (t_r, t_g, t_b)$  and  $V_I, V_B$  which stand for two 3 by 3 matrices deduced accordingly from formula (1), thus  $t$  is found as the generalized eigen vector corresponding to the greatest generalized eigen value of the following generalized eigen value problem:  $t^T V_I t \rightarrow \max_{t \in \mathbb{R}^3} ; \text{s.t. } t^T V_B t = 1$ . Note that care is taken in order to yield the vector  $t$  with the right sign. The sample of insect intensities  $\mathcal{S}_I$  being constructed off-line from many available sample videos (we use a database of 500 insect objects), and the sample of background intensities  $\mathcal{S}_B$  being available after the extraction zone of the sticky trap in the first video-frame,

hence, the linear coefficients  $t_r$ ,  $t_g$ ,  $t_b$  are estimated once and for all at the launching of the application.

## 2.2 Recognition of potential locations of insects of interest in a video-frame

After transformation of an RGB video-frame into a gray image (cf. Fig.2(b)), we would like to automatically extract from it any bright spot that may correspond to an insect of interest in the original video-frame. To do so, we model such a spot which as a contrasted rectangular pattern  $\mathcal{R} := \mathcal{R}(r, w, \theta, f(\cdot, \cdot))$ , with  $r$  and  $w$  standing for its half-width and its half-length respectively,  $\theta$  which stands for its tilt angle, and  $f(x, y)$  which stands for a  $2D$  function describing the gray intensity level at any point  $(x, y)$  of the plane. For simplicity's sake, we shall assume that  $f(x, y)$  is a piecewise constant function which is equal to a constant  $h+a$  inside the rectangle, and to a constant  $a$  outside the rectangle as follows:

$$f(x, y) = \begin{cases} h + a, & \text{if } |x \cos(\theta) + y \sin(\theta)| \leq w \text{ and} \\ & |-x \sin(\theta) + y \cos(\theta)| \leq r; \\ a, & \text{otherwise.} \end{cases}$$

where  $h$  stands for the gray contrast of  $\mathcal{R}$  and  $a$  stands for the gray level of its surrounding background. Now, in order to yield a continuously differentiable  $2D$  image which can show "singularities", namely local maxima at the rectangular zones of interest in the image and which can be extracted efficiently by using a geometric differential technique, for instance by using the Karush-Kuhn-Tucker (KKT) local maximality criterion, we propose to convolve  $f(x, y)$  with a gaussian kernel  $K_\sigma(x, y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x^2+y^2)}{2\sigma^2}}$  to obtain the following  $2D$  scale space intensity profile

$$f_\sigma(x, y) = h \times (\Phi_\sigma(x \cos(\theta) + y \sin(\theta) + w) - \Phi_\sigma(x \cos(\theta) + y \sin(\theta) - w)) (\Phi_\sigma(-x \sin(\theta) + y \cos(\theta) + r) - \Phi_\sigma(-x \sin(\theta) + y \cos(\theta) - r))$$

with  $\Phi_\sigma(t) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^t e^{-\frac{u^2}{2\sigma^2}} du$ . Now, let us perform the following change of variable:  $u = x \cos(\theta) + y \sin(\theta)$ ,  $v = -x \sin(\theta) + y \cos(\theta)$ , one can then rewrite the scale space rectangular profile, denoted by  $g_\sigma(u, v)$ , in the new referential as a tensor product of two  $1D$  functions as follows:  $g_\sigma(u, v) = h g_{\sigma,w}(u) g_{\sigma,r}(v)$  with  $g_{\sigma,w}(u) = (\Phi_\sigma(u + w) - \Phi_\sigma(u - w))$  and  $g_{\sigma,r}(v) = (\Phi_\sigma(v + r) - \Phi_\sigma(v - r))$  which is handier (but equivalent) to study than  $f_\sigma(x, y)$ . After some lengthy calculations, we found out indeed that the conditions on  $\sigma$  in order for  $g_\sigma(u, v)$  (or  $f_\sigma(x, y)$ ) to show

a clearly defined local maximum at the centre of mass of the rectangular pattern are:

$$\begin{aligned} \left(\frac{w^2}{\sigma^2} - 3\right)(2\Phi_\sigma(r) - 1) - \frac{2r}{\sqrt{2\pi}\sigma} e^{-\frac{r^2}{2\sigma^2}} &\leq 0 \\ \left(\frac{r^2}{\sigma^2} - 3\right)(2\Phi_\sigma(w) - 1) - \frac{2w}{\sqrt{2\pi}\sigma} e^{-\frac{w^2}{2\sigma^2}} &\leq 0 \end{aligned}$$

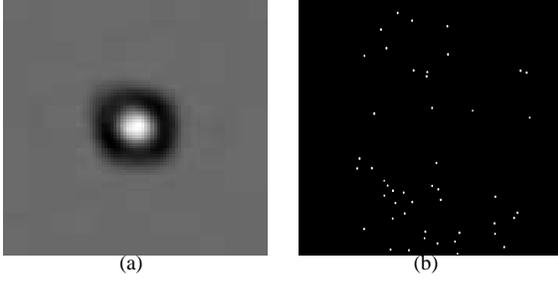
then, it is easy to check that the latter conditions are satisfied simultaneously for any  $\sigma \geq \frac{\max(w, r)}{\sqrt{3}}$ . Therefore,  $\sigma$  is chosen, in our application, as follows  $\sigma := \frac{\ell_1}{\sqrt{3}}$  with  $\ell_1$  standing for the prior about maximum half-width or maximum half-length of an insect of interest in a video, and the latter is previously available for the user from sample videos and is given as a parameter to the application. Now, the KKT sufficient conditions of local maximality of  $g_\sigma(u, v)$  at the centroidal point  $(0, 0)$  of  $\mathcal{R}$  express as follows:

$$\nabla g_\sigma(0, 0) = 0 \quad (2)$$

$$\nabla^2 g_\sigma(0, 0) < 0 \quad (3)$$

with  $\nabla$  and  $\nabla^2$  which stand respectively for the gradient and the Hessian operators with respect to  $u$  and  $v$ . For detection purposes, we shall focus more on the KKT second condition of local maximality (3) of  $g_\sigma(u, v)$  at the centroidal point of some rectangular pattern which can also be seen as a measure of its saliency. Such a criterion amounts then to saying that both eigen values of  $\nabla^2 g_\sigma(0, 0)$  are negative. Note then that the (symmetric) matrices  $\nabla^2 f_\sigma(x, y)$  and  $\nabla^2 g_\sigma(u, v)$  are related through the formula:  $\nabla^2 f_\sigma(x, y) = \mathbf{R}_\theta \nabla^2 g_\sigma(u, v) \mathbf{R}_\theta^T$  where  $\mathbf{R}_\theta$  stands for a rotation matrix which is given by  $\mathbf{R}_\theta = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$ ; and one deduces that  $\nabla^2 f_\sigma(x, y)$  and  $\nabla^2 g_\sigma(u, v)$  have the same eigen values. As a consequence, knowledge of the tilt angle  $\theta$  of a rectangular pattern is not necessary (thus, invariance of the detection criterion w.r.t.  $\theta$ ), and one may formulate the detection criterion of the centroidal region of an insect of interest as the one that both eigen values of  $\nabla^2 f_\sigma(0, 0)$  should be negative, moreover, in order to be robust against some image artifacts, one adds that its greatest eigen value should lie below some (negative) threshold " $s_*$ ", to mean that only salient enough rectangular patterns which may correspond to insects of interest in some video-frame should be considered. In practice,  $s_*$  may be estimated off-line from sampled insects in videos as we shall describe it hereafter. First of all, one notices that, as a consequence of the KKT first condition of local maximality (2), the non-diagonal values of the Hessian matrix  $\nabla^2 g_\sigma(0, 0)$  vanish. The respective eigen values of matrix  $\nabla^2 g_\sigma(0, 0)$  (or equivalently

of  $\nabla^2 f_\sigma(0,0)$  denoted by  $s_1(\sigma, w, r)$  and  $s_2(\sigma, w, r)$  are then given in closed form by:



**Figure 3.** (a) The typical response of an insect location to the measure  $-\sup \{ \text{greatest eigen value}(-\nabla^2 f_\sigma(x, y)), 0 \}$  (computed for the imagette in Fig.1(b)) ; (b) Extraction of insect locations from the video-frame in Fig.2(a).

$$s_1(\sigma, w, r) := \frac{\partial^2 g_\sigma(0,0)}{\partial u^2} = \frac{-2hw}{\sqrt{2\pi}\sigma^3} e^{-\frac{w^2}{2\sigma^2}} (2\Phi_\sigma(r) - 1)$$

$$s_2(\sigma, w, r) := \frac{\partial^2 g_\sigma(0,0)}{\partial v^2} = \frac{-2hr}{\sqrt{2\pi}\sigma^3} e^{-\frac{r^2}{2\sigma^2}} (2\Phi_\sigma(w) - 1)$$

Therefore, given *a priori* the minimum and the maximum values  $\ell_0$  and  $\ell_1$  of either  $r$  or  $w$ , the minimum and the maximum areas  $A_0$  and  $A_1$  respectively, and the minimum contrast value  $h_0$  (all these parameters being available for us off-line), the threshold value  $s_*$  may then be estimated robustly as follows (we assume uncorrelation between intensity and geometric parameters):

$$s_* := \frac{-2h_0}{\sqrt{2\pi}\sigma^3} \min_{\ell_0 \leq w, r \leq \ell_1; A_0 \leq 4wr \leq A_1} w e^{-\frac{w^2}{2\sigma^2}} (2\Phi_\sigma(r) - 1)$$

Thus, our algorithm for the extraction of the spots that may correspond to the insects of interest in some video-frame proceeds in two steps : a recognition step and a segmentation step. In the recognition step, a binary image where all potential centroidal locations of insects in some video-frame are extracted based on the criterion that at some pixel  $(x, y)$ , matrix  $\nabla^2 f_\sigma(x, y)$  is negative definite, moreover its greatest eigen values lies below  $s_*$ , hence a run of the connected components algorithm allows to group such pixels into connected components. In a second step, a box  $\mathcal{B}$  of size  $L \times L$  (for  $L := \lceil \sqrt{3}\sigma \rceil$ ) is computed around the centroid of each extracted connected component in the original video-frame, and a quick conquer-and-merge segmentation strategy allows to add to a connected component all other insect pixels lying in  $\mathcal{B}$ . Finally, too small found connected components are filtered out (cf. Fig.3(b)).

### 3 An on-line video-processing solution

So far, we have explained the principle of the recognition and counting of the harmful insects of interest in individual video-frames which, rather, is a too slow process (for information, its takes about 4 secs. to treat one video-frame of 1.3 mega pixels). Hence, we would like now to describe the on-line video-processing solution that we proposed for the video-surveillance application at hand which is made possible by the fact that the process which consists of trapping harmful insects is a sparse one, in the sense that the probability that a harmful insect is trapped during a short period of time (e.g. a few minutes) is very low. Thus, a video-processing solution which makes parsimonious use of the insect detector described in section 2 depending on the occurrence or not of some event that might be related to the trap of a new insect of interest. To do so, initially a video-frame is divided into a number of  $k \times k$  small virtual image patches (e.g.  $k = 10$ ) that can be processed much more quickly than a whole video-frame by the insect detector, and to avoid any border effect, hence patch-overlapping is used. Then, a quick background subtraction algorithm (BSA) in the spirit of the Mixture of Gaussians algorithm (MoG) [7] which runs permanently with respect to each video allows to integrate pixel intensity information over time, and to emit a signal whenever a significant change in the intensity of some pixel has been detected. Since a change that is detected at some pixel might be due either to the trap of new insect of interest or to any other useless event (noise, light change, outlier crossing the FOV of a camera, etc.), hence such a pixel undergoes a second test referred as the insect presence detection test which will classify it as "likely" or "unlikely" to be an insect pixel of interest. This is achieved by learning off-line the space of color intensities of insects of interest by means of a Principal Components Analysis (PCA). Then, under white noise assumptions with respect to the each of three PCA axes, testing if some pixel could belong to an insect of interest, amounts to testing the gaussianity of each of its three PCA residuals independently after subtraction of the mean intensity. The frame patch with maximum number of pixels that passed the insect presence detection test is then submitted to the insect detector described above in order to realize independently a precise detection of any recently trapped insect of interest, provided that this number exceeds some predefined threshold. A detection is validated if and only if it intersects with a minimal number of pixels that passed the presence detection test. Since a trapped insect manages generally to displace at least slightly from its initial position, mainly

during the few first minutes following its trap, hence a quick TBD (Track-Before-Detect) type tracking algorithm allows to keep track of it through time as follows. Firstly, a square bounding box computed in the current frame "t" around the insect and the RGB intensities lying in its inside are transformed into gray intensities as explained in section 2.1, then sorted in an ascending order and arranged in a feature vector  $Y_0^t$ . One notes the invariance of the feature vector  $Y_0^t$  with respect to a rotation or a translation of the insect. In the next frames, such a bounding box is updated by sliding it in the neighborhood of its current position (typically, a small square window of some predefined size), and for each slid box, a new feature vector  $Y^{t+1}$  is computed in the same way as explained for  $Y_0^t$ . The bounding box of the insect is then updated by maximizing a similarity criterion between  $Y_0^t$  and  $Y^{t+1}$  of the form  $\mathcal{S}(Y_0^t, Y^{t+1}) := \min \left\{ \frac{\|Y_0^t\|}{\|Y^{t+1}\|}, \frac{\|Y^{t+1}\|}{\|Y_0^t\|} \right\} \frac{Y_0^{tT} Y^{t+1}}{\|Y_0^t\| \|Y^{t+1}\|}$  where the quantity  $\frac{Y_0^{tT} Y^{t+1}}{\|Y_0^t\| \|Y^{t+1}\|}$  stands for the correlation coefficient between the two feature vectors  $Y_0^t$  and  $Y^{t+1}$ , and the factor  $\min \left\{ \frac{\|Y_0^t\|}{\|Y^{t+1}\|}, \frac{\|Y^{t+1}\|}{\|Y_0^t\|} \right\}$  allows to favor two feature vectors  $Y_0^t$  and  $Y^{t+1}$  with comparable modules so as to fight more robustly against noise. The tracking is accelerated significantly by using the information about insect presence detection yielded previously, in such a way that only boxes which intersect with a minimal number of pixels that passed the presence detection test are considered.

## 4 Method's evaluation

The currently developed version of our vision application has been tested off-line against 8 video sequences representing the whitefly species and recorded under realistic *in situ* conditions during daytime for periods ranging from 20 minutes to 1 hour (the process of trapping insects has been accelerated by placing the cameras in highly infested zones). So, by combining the video spatiotemporal information, which is gained on the one hand from the insect presence detection module, and on the other hand from the pattern extraction module, and by tuning accordingly the parameters of the application, then an assessment of the results against ground truth revealed that the false positive rate is negligible in all tested video sequences (namely, only one false positive has been found against about 250 found others), whereas the false negative rate is of order of 3% and the latter concerns mainly some insects that were not detected by the algorithm because of their too low signal to noise ratio and/or a highly illuminated neighboring background.

## 5 Conclusion

We developed a new full on-line computer vision prototype for in-situ pest monitoring and we showed its feasibility for the case of one potential harmful pest species (the whitefly), nevertheless, its extension in order to take into account other harmful pest species of interest (e.g. the greenfly species) is straightforward. As described above, the developed system relies mainly on a pattern recognition algorithm for extracting the locations of harmful insects in a video, and because of its generic aspect, we believe that the same algorithm may be used in other vision application contexts (e.g. active vision) for carrying out efficiently low-level image processing operations amounting to extracting salient points of interest from gray-scale images, moreover, by allowing the scale parameter  $\sigma$  to vary, one may achieve an efficient multiresolution feature extraction framework in active vision and video-surveillance.

## 6 Acknowledgements

The authors are grateful to CREAT (La Gaude) and INRA (Avignon) for their participation in this work.

## References

- [1] C. Bauch and T. Rath. A prototype of a vision based system for measurements of whitefly infestation. *Acta Horticulturae (ISHS)*, 0(691):773–780, Jan. 2005.
- [2] E. Z. G. Loy. A fast radial symmetry transform for detecting points of interest. *7th European Conference on Computer Vision*, Oct.. 2002.
- [3] H. B. H.-Y. Chen, N. Brandle and H. Lapp. Robust spot fitting for genetic spot array images. *Image Processing, Proceedings. International Conference on Pattern Recognition and Image Processing*, 3:412–415, Oct.. 2000.
- [4] L. V. N. Haering, A. Peter and A. Lipton. The evolution of video surveillance: an overview. *Machine Vision and Applications*, 19:279–290, Jun. 2008.
- [5] J. Olivo-Marin. Extraction of spots in biological images using multiscale products. *Pattern Recognition*, 35(9):1989–1996, Sept. 2002.
- [6] V. M. P. Boissard and S. Moisan. A cognitive vision approach to early pest detection in greenhouse crops. *Int. Journ. of Comp. Elect. in Agric.*, 2(62):81–93, Jul. 2008.
- [7] C. Stauffer and W. Grimson. Background mixture models for real-time tracking. *CVPR*, Aug. 1999.