Simultaneous Localization and Object Detection using an a-contrario approach

S. Le Hégarat-Mascle*
University Paris-South 11
IEF (bat.220)
91405 Orsay Cedex, France
sylvie.le-hegarat@u-psud.fr

A. Robin
University of the Witwatersrand, SCAM
Wits 2050, Johannesburg, South Africa
amandine.robin@wits.ac.za

R. Reynaud
University Paris-South 11
IEF (bat.220)
91405 Orsay Cedex, France
roger.reynaud@u-psud.fr

ABSTRACT

In the context of automotive driver assistance, we focus on the problem of simultaneous localization and object detection considering a video sequence acquired by an on-board camera. This paper presents an original approach permitting localization and object detection by using coarse resolution images. It is based on an a-contrario model previously introduced for land cover monitoring using remote sensing data. Applied to the problem of detecting scene changes from the acquisition of a video sequence from an on-board camera, we show that such an approach permits to detect appearing objects even when the illumination and the geometry of the scene vary, and this in a much more robust way than keeping full resolution data. Results obtained in the context of real data acquired using a frontal camera on-board a car illustrate these statements.

1. INTRODUCTION

One of the major applications of on-board automotive driver assistance systems is to alert the driver about driving environment events and possible collision with obstacles (other vehicles, pedestrian, etc.). In this context, Advanced Driver Assistance Systems (ADASs) are able to provide more and more precise information to the driver [15, 12, 3], such as the road position, the distance to other vehicles [26], the presence of pedestrians on the road [11, 10, 6]. Part of this information can be derived by simple sensors (such as radar or lidar) providing estimations of the obstacle/vehicle distance). Their main advantage is that they allow measurements (*e.g.* distance) without requiring powerful computing resources. However, their spatial resolution is generally low, leading to detection failure when the number of targets to detect is large typically.

Besides, vision-based systems allow to derive various types of information using a single sensor (or double in case of stereovision). For instance, obstacles can be detected as well as horizontal or vertical road signs (e.g. road surface markings or traffic signs) by using a frontal on-board camera. For such systems, stereovision provides information on the distance to the objects present in the scene. For instance, the GOLD system [18] addressed both lane and obstacle detection at the same time by using stereovision. Stereovision methods are based either on disparity maps or on inverse perspective mapping, or on a combination of both [2]. In all cases, their computational cost is important. For instance, the computation of a disparity map requires to solve the matching problem in every pixel and its complexity can even be increased in case more than two cameras are being used [21].

Some methods are motion-based, using optical flow computation (e.g. [1]). They are capable to distinguish moving objects all the more easily so as this movement is different from the global scene movement (due to the vehicle movement). Thus, a strong drawback of these methods is that they would not detect stationary obstacles. Finally, some other methods are based on specific features present in the researched objects: symmetry [14], color or shadow, vertical or horizontal edges [25, 19]. In [6], the authors distinguish six main steps for pedestrian detection: (i) preprocessing; (ii) foreground segmentation; (iii) object classification; (iv) verification/refinement; (v) tracking; and (vi) application. The preprocessing step includes camera calibration and extrinsic parameter updating. Then, foreground segmentation aims at extracting 'Regions Of Interest' (ROI). It can either be 2D-based or, preferentially, stereo-based combined possibly with other information, e.g. Thermal Infra-Red imaging [13], or motion analysis [9]. The two following steps are even more specific to the pedestrian detection: object classification for instance uses silhouette matching or appearance measured through various descriptors, e.g. Haar wavelets [22] or histograms of oriented gradients [17], that are the SIFT-inspired features. The most used classification approaches are supervised ones (database learning) and based on SVM (Support Vector Machines) and AdaBoost. Then, as a complement to the previous classification step, the verification step aims at removing false positives whereas the refinement one aims at obtaining a fine segmentation of the pedestrian and an accurate distance estimation. Tracking is generally done via a Kalman filter or a particle filter. Its interests are related to the other steps, e.g. removing remaining false positives, predicting next ROI, decision at the application level.

*Corresponding author

ICVGIP '10, December 12-15, 2010, Chennai, India

This study is motivated by the creation of a vehicle empowered with an autonomous system allowing self-positioning in a learned environment and the detection of changes arising in this environment (in particular, the arrival of new obstacles). Therefore, relatively to the motion-based approach described earlier, in this paper we focus on the 'extraction of ROI' step. More precisely, we consider a vehicle with an on-board camera acquiring a video sequence as it is moving and a reference image sequence of the same route without "obstacles" (refered to as 'learning image sequence' in the following). Given a newly acquired image at time t, our goal is to identify the closest image in the learning image sequence and then to detect the actual changes between this new image and the learning sequence. We refer to this problem as SLOD, as the acronym of Simultaneous Localization and Object Detection (as opposed to the SLAM acronym of Simultaneous Localization And Mapping). A similar problem has also been considered by Wang et al. [5] for robotics, associating SLAM with object tracking. However, they focus on the motion modeling of a robot and generalized objects, whereas we focus on the detection of objects encountered along the route of the vehicle.

More precisely, as the newly acquired sequence might differ from the learning sequence of reference, our main problem here is to detect changes that are due to the emergence of a new object when various other changes attributed to a "normal variability" are likely to occur also, due to changes in acquisition conditions (illumination and geometry). To that aim, first, we propose to lower the spatial resolution of the learning sequence in order to filter minor changes (e.g. leaves in the trees) and to quantify the newly acquired images for robustness to changes in illumination conditions. Then, the most significative changes will be detected by using an a-contrario criterion to evaluate the consistency between a new quantified -but high spatial resolution- image (QHR) and each low spatial resolution image (LR) of the learning sequence.

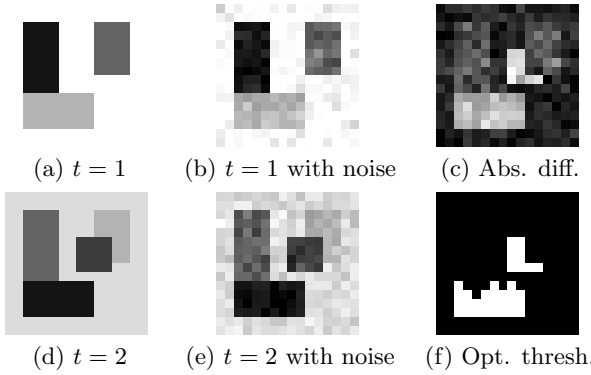
In Section 2, the choice of using a LR learning sequence as well as a QHR image is further explained and motivated through synthetic examples. Then, Section 3 details the model used for the SLOD problem which ensure a control of the average number of false detections. Section 4 states the algorithm before some results are shown in Section 5 through examples using actual data acquired by an on-board camera. Finally, Section 6 gathers the conclusions and perspectives of this study.

2. PROPOSED APPROACH

The cornerstone of our approach stands in lowering the resolution of both the learning sequence (spatially) and of the new image (in grey-levels, by quantification) for robustness and computation time purposes. In this section, we explain the reasons of this choice and illustrate them through several simple examples which characterize typical improvements obtained by using an image quantification and lowering the spatial resolution.

The acquisition conditions of the learning sequence may differ from those of the image at instant t as both the illumination conditions and the geometry of acquisition change. Most methods in the literature cannot handle that type of situation and come up with false detections (detections that do not correspond to actual changes). To avoid such a drawback, we suggest to introduce a tolerance to changes in grey-levels by adopting a class-based comparison instead of a greyscale-based one.

Let us illustrate typical drawbacks of a greyscale-based approach through the simple example shown in Figures 1, 2, 3. Different types of changes are present in these three figures: some grey-level changes that appear within an object are not to be detected whereas the changes that correspond to the arrival of a new object (*e.g.* square about in the center of the image, in this synthetic example) must be detected. This latter type will be referred to as an *actual* change in the following. Figure 1 illustrates the different type of errors (false negatives and false positives) obtained typically when processing the images at a grey-scale level (without quantification). Other authors have proposed to consider the image edge information (e.g. [25, 16, 7]). Such an approach is illustrated in Figure 2 together with its sensitivity to image noise.



(a) $t = 1$ (b) $t = 1$ with noise (c) Abs. diff.

(d) $t = 2$ (e) $t = 2$ with noise (f) Opt. thresh.

Figure 1: Detection of an appearing square in synthetic images 16×16: figures (a) and (d) correspond to the state of a scene resp. at time $t = 1$ (with grey-level values in $\{20, 100, 180, 255\}$) and $t = 2$ (with grey-level values in $\{20, 60, 100, 180, 220\}$). Notice the appearing square and the change of grey-level values; images (b) and (e) are simulated resp. from (a) and (b) by using a Gaussian noise of standard deviation equal to 15; (c) shows the absolute difference image ($|(e) - (b)|$) and (f) results from the "optimal" thresholding of (c). Note that the actual change (square) is only partly detected due to grey-level conflicts and that the bottom rectangle is improperly detected.

Now, to introduce the idea of a class-based approach, the third raw of Figure 2 shows the result obtained considering a QHR image with four levels of quantification instead of the 255 grey levels. Let us remark that the result has been improved a lot simply by using image filtering and quantification. However, the edge binary difference is not trivial to interpret in term of scene changes. Hence, we rather adopted a region-based approach, following Robin's idea [23] where a classification image is used to infer the structure of the image. Following this approach, the grey-level values are not important *per se* but relatively to the average grey-level values of other classes. Thus, the detection process is robust to changes of illumination (e.g. due to an automatic re-calibration of the camera) that would affect all grey-level values of the image.

As announced in Section 1, in this paper any 'new' image

acquired at time t will be quantified using k levels, assimilated to classes in the following. Relatively to an approach that only considers edges, this adds the constraint for the different objects of a same quantification class to be represented by a same value of quantification at each of the considered dates (but this value may varies between the two dates).

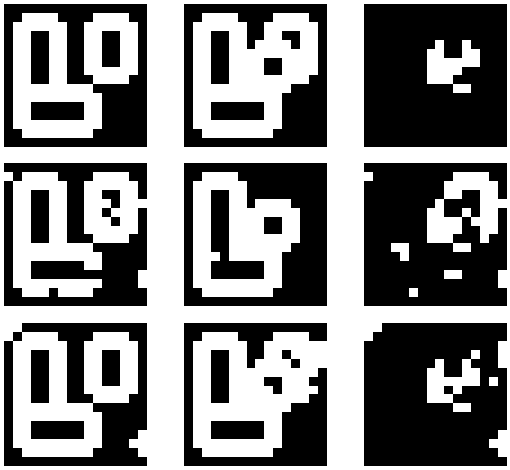


Figure 2: Binary images of edges of size 16×16 (obtained using a morphological gradient and tuning the threshold for edge detection) resp. acquired at t_1 (1^{st} column) and t_2 (2^{nd} column) and binary difference of the edge images (3^{rd} column); 1^{st} raw: synthetic images; 2^{nd} raw: simulated images assuming a Gaussian noise of standard deviation equal to 15; 3^{rd} raw: quantified images on 4 levels. We note that edge-based approach fails as soon as the noise level is too important. Due to the image filtering and simplification, a better result is obtained using only few levels of quantification.

Having explained the interest of image quantification and of a classification-based approach, let us now illustrate the interest of a LR learning sequence, through the synthetic example presented Figure 3. Here, in order to separate the sources of errors, we ignore changes in illumination conditions but misregistration errors have been introduced. Hence two images acquired at time t_1 and t_2 have the same levels of quantification but the image at t_2 is $1/4$ pixel shifted relatively to the image at t_1. It appears clearly that the false positives that are present using full resolution images disappear by applying a spatial resolution reduction of factor 2×2 to the initial images. In the following, we perform both image transformations: intensity quantification and reduction of the spatial resolution (by window averaging), but each separately on a different image:

- The 'new' image acquired at time t is quantified into k levels (or classes). Then, the algorithm estimates automatically the grey-levels characterizing each class by using the values of the learning sequence (using the a-contrario criterion described in Section 3). The image comparison will then be based on the consistency of the classification induced by the quantification relatively to the images of the learning database.

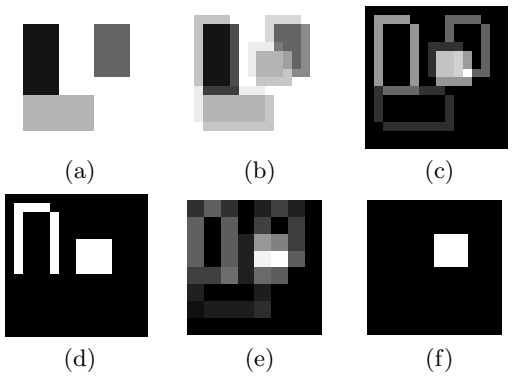


Figure 3: Case of misregistration: images 16×16 resp. acquired at time t_1 (a) and t_2 (b), where a square appeared in (b) together with a $1/4$ pixel shift relatively to $t = 1$; (c) shows the absolute difference $|(b) - (a)|$ and (d) the result of the '0−false negative' thresholding; (e) shows the absolute difference between grey-level pixel values of the $2 \times 2-$reduced images and (f) the '0−false negative' thresholding result. Notice that the shift that is wrongly detected in (d) is no longer detected in (f), showing the potentiel of using a reduced spatial resolution to be robust to slight misregistration.

- The reduction of the spatial resolution is performed on the images of the learning sequence. Then, the spatial transformation between a 'new' image and its 'closest' image in the LR learning sequence is approximated by a translation. Such an approximation is,at the same time, reasonable since we can search for a position of an acquisition viewpoint in the learning sequence that would be close to those considered at t and coarse if the route followed by the vehicle is not parallel to the one of the learning sequence. Therefore, comparing with LR images leads to some robustness relatively to noise registration (in addition to an interest in term of management of the memory resources).

In summary, our approach is based on the two following ideas: firstly, the transformation of the acquisition conditions can be estimated only roughly and, secondly, the imprecision on the previously mentioned transformation can be partly 'filtered' by considering some images with lowered precisions both spatially and in intensity. In other words, the fact to consider lowered resolution allows to introduce some 'fuzzyness' in the image models.

3. MODEL

The proposed model comes from a previous study performed in the context of unsupervised sub-pixel change detection using time series of satellite images [23, 24]. In this latter work, a high resolution classification image is used as a description of the reference state of the scene and a LR sequence of images is used to monitore the changes along with time. A consistency measure has been introduced between the classification and the sequence, permitting an estimation of the image sub-domain where the LR sequence corresponds significantly to the classification of reference while controlling the average number of false alarms. The esti-

mated change domain is then directly obtained by taking the complementary part of the latter image sub-domain. Notice that in [23, 24], all images correspond to the same scene and are assumed to be perfectly registered. One major property of this consistency measure is that its value can be compared and interpreted for different images or data sets. Indeed, it corresponds to an expectation which, conversely to a probability measure, can be interpreted in itself. Here, the scene is no longer static as we consider a moving camera. Thus, our goal is now to detect appearing objects in a 'moving' background. In this context, this method presents the noticeable advantage of allowing to measure the consistency of different images even though they are not superimposable, and to compare the levels of consistency of each pair of images thanks to the previously mentioned property of the expectation. In this paper, as introduced in Section 2, we aim at comparing a LR image sequence with a QHR image. As in [23, 24], we propose to measure the degree of consistency between the image at an instant t and the images of the learning sequence that have been acquired earlier, thus with different acquisition conditions -lightning or geometrical- (e.g. from different viewpoints if the vehicle is positioned differently on the road, at different hours of the day or with a different weather). More precisely, we suggest to search the LR image in the learning sequence that is the most 'consistent' with the QHR image, according to the defined consistency criterion. This consistency of a LR image relatively to the QHR image is measured in term of contradiction of an unstructured model The arrival of new objects with regard to the learning sequence is then detected and localized as the complementary part of areas where the LR image is the most consistent with the QHR image. Let Ω and Ω' denote respectively the high resolution and low resolution image domains. The QHR image is defined on Ω with values in $\mathcal{L}$ (set of all levels of quantification, of cardinal $|\mathcal{L}|$). The LR sequence is denoted by $(v_t)_t$, where for each instant t the image v_t is a real-valued function defined on Ω'. As each LR image is obtained by block-averaging a high resolution image, the expectation of the measurement performed over a LR pixel is the average of the measure corresponding to each level of quantification (in the QHR image) weighted by its occupation rate in the pixel. The value observed within a LR pixel x at an instant t can hence be estimated by

$$\hat{v}_t(x) = \sum_l \alpha_l(x) M_t(l), \tag{1}$$

where $\alpha_l(x)$ denotes the relative area of the LR pixel x corresponding to the level of quantification l (by construction, $\sum_{l\in\mathcal{L}} \alpha_l(x) = 1$) and $M_t(l)$ represents the average intensity corresponding to the each level of quantification l at the instant t. Note that as the QHR image is given, the proportions $\alpha_l(x)$ are known. The minimal residual error between the observations and the reconstruction obtained from a given distribution of the quantification levels can be measured over a sub-domain $\omega \subset \Omega$ by using the squared Euclidean norm by

$$E_\omega = \min_M ||(v_t(x) - \hat{v}_t(x))\mathbb{1}_\omega(x)||_2^2. \tag{2}$$

Then, the average intensity $M_t(l)$ can be estimated by minimizing the residual error over the sub-domain $\omega \in \Omega'$. From there, the sub-domain for which the residual error is *particularly* small is assumed *unchanged*, following the idea that the image v_t is then well approximated from the QHR image (through $\hat{v}_t$). Then, a core issue is the choice of a threshold from which deciding that the residual error is acceptable for an unchanged domain. Following the general framework of a-contrario modeling, let us consider the probability to observe the residual error E_ω *by chance*, denoted by $\mathbb{P}_{H_0}(E_\omega)$. This can be done by assuming the a-contrario random model (H_0): a LR image v is a random field V of $|\Omega|$ independent Gaussian centered variables with a given variance σ^2. The purpose of this model is not to reasonably model the data but only to define a noise model against which detecting significant structures in the data. From there, we define the consistency measure over a spatial sub-domain ω by

$$NFA(\omega, E_\omega) = \eta(|\omega|) \cdot \mathbb{P}_{H_0}(E_\omega), \tag{3}$$

where η is a normalization term chosen so that the expected number of false alarms can be controled. After computation,

$$\mathbb{P}_{H_0}(E_\omega) = \frac{1}{\Gamma(\frac{|\omega|-|\mathcal{L}|}{2})} \int_0^{E_\omega/2\sigma^2} e^{-t} t^{\frac{|\omega|-|\mathcal{L}|}{2}-1} dt, \tag{4}$$

where Γ is the usual Euler function. Here, we choose $\eta = |\Omega|\binom{|\omega|}{|\Omega|}$ in order to distribute the risk uniformly with respect to the domain size. This measure depends on the size of the considered sub-domain $|\omega|$, on the number of quantification levels $|\mathcal{L}|$ and on the variance of the a-contrario model. All these parameters are obtained directly from the data, except the variance σ^2 which is chosen arbitrary, fixed equal to the empirical variance of the current LR image. The model is then free of parameters meanwhile ensuring a robust control of the average number of false alarms. As mentioned earlier, the value of this NFA function has a meaning in itself (in absolute terms). Thus, it can be evaluated and compared for different images. This is the principle we use in order to find the image of the LR sequence that is the most consistent with the QHR image (maximizing the NFA over all images). In addition, when all images of the sequence are not superimposable, this property can also be used in order to estimate the transformation to apply to an image to register it with the reference one (among a finite set of simple transformations). In practice, Section 4 details how this model is used for SLOD.

4. ALGORITHM

As the consistency measure (4) corresponds to the expectation of the number of false alarms, with respect to the defined a-contrario model (Section 3), its value can be compared and interpreted for different images. Thus, looking for the image sub-domain that minimizes the NFA permits to detect changes whereas comparing the NFA minimum value obtained for different dates gives the date for which a LR image is the most consistent with a QHR image, *i.e.* the image in the learning database corresponding to the QHR image of interest (in which objects need to be detected). Moreover, in the same spirit, the NFA measure can be used to register two images. Indeed, as the camera is moving with time, the images acquired are slightly shifted from one date to another. Here, as a first approximation, we assume that the transformation linking two images is a simple translation (tx, ty), and for a given set of translations to explore, we decide for the translation which minimizes the NFA. The algorithm is based on a random sampling strategy (*cf.* [8]) for pixel selection. This strategy, combined with the a-contrario

model, leads to a robust detection method. Notice that all parameters of the NFA can be obtained directly from the data except the cumulated quadratic residue E_ω which depends on the means $M_t(l)$ corresponding to each level of quantification l, *a priori* unknown. The mean estimation and the detection itself are two closely linked problems as the quality of the estimation has a strong impact on the performance of the detection. The algorithm described below achieves both tasks simultaneously and is thus fully unsupervised. It takes a learning sequence and a *current* sequence as inputs and returns, for each image of the current sequence, the closest image in the learning sequence (t), the translation to apply to it for registration ((tx, ty)) and the detected objects (domain ω).

- Initialize table $NFA[tx][ty][t]$ to $+\infty$.
- For translations $(tx, ty, t) = (0, 0, 0)$ to (tx_{max}, ty_{max}, nt),
  - Shift the QHR image of (tx, ty),
  - Repeat N times
    1. draw randomly $|\mathcal{L}|$ LR pixels x, denoted by $I = (x_1, \cdots, x_{|\mathcal{L}|})$;
    2. estimate the mean values $(M_t(l))_{t,l}$ by linear regression
    3. compute the residuals $r_t(x) = (v_t(x) - \hat{v_t})^2$, for $x \in \Omega'$;
    4. sort Ω' into a vector $(x_i)_{1 \leq i \leq |\Omega'|}$ by increasing error $r(x_i)$;
    5. initialize $E = \sum_{i=0}^{|\mathcal{L}|} r_t(x_i)$;
    6. for each index $i \in \{|\mathcal{L}| + 1, \ldots, |\Omega'|\}$,
       * set $E = E + r(x_i)$;
       * if $E < E_{min}[i]$ then
         · set $E_{min}[i] = E$;
         · compute the corresponding $NFA[tx][ty][t]$ value;
         · update NFA_{min} and ω
       * end if
    7. end for
  - end repeat
- end for

Notice that, in practice and thanks to the high robustness due to spatial resolution lowering, time computation can be considerably reduced by first optimizing the NFA criterion over the time t in order to find the closest image in the learning sequence (even if it contains a shift from the QHR image) and then optimize the NFA criterion over a finite set of translations (tx, ty) in order to precise the object detection.

5. RESULTS

In this section, let us consider a sequence acquired along several round trips on a same road. The vehicle containing the camera onboard is a dedicated vehicle for research about autonomous or quasi-autonomous systems [4, 20]. The camera onboard is a standard color camera (three channels in the visible domain) acquiring 25 images per second, each one of size 240 lines by 320 columns. The learning image database is composed of the images acquired along the first round trip (images acquired at $t \in [1300, 1600]$). The SLOD algorithm is then applied for images acquired at $t \in \{5620, 5750, 5800, 5870\}$ in order to find the *closest* image in the learning database. For first experiments, we consider quantification images with 6 levels and a spatial resolution ratio equal to 8×8, each LR image being obtained from a high resolution one by block averaging. To ensure convergence, the algorithm was run using a number of iterations equal to 100000. For the localization step, we look for the image in the learning sequence which, for a given quantified image, minimizes the NFA (thus minimizing the consistency, see Figure 5). In figure 4, the NFA values obtained using the quantified image corresponding to the time index $t = 5750$ and the learning sequence are plotted as a function of time. Notice that the minimum is sharply obtained for the closest image (indexed 10). Figure 5 illustrates the

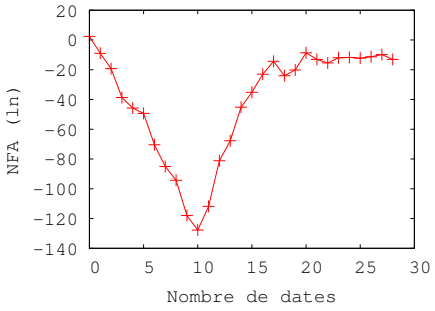


Figure 4: Evolution of the NFA values obtained (in ln scale) as a function of the time index in the sequence. Note the sharpen figure of the minimum.

Object Detection part of the SLOD proposed algorithm. It shows, from top to bottom: the image at time t, the image at t after quantification (6 levels), the LR image (8×8 block averaging) found as the corresponding image in the learning sequence, and the corresponding LR image found after registration, by optimizing the NFA over a set of spatial translations of the image at t. On these two last images, the black pixels are the pixels of change or detected objects. The objects to detect are the moving cars and the pedestrians since these latter were not present during the learning sequence acquisition. Note that in general the objects to detect are well detected. On the left example ($t = 5620$), a pedestrian and a car are both well detected. On the right example ($t = 5750$), the three cars are well detected once the images are registered (4th row) but only the pedestrian in the foreground is well detected (the second pedestrian on the edge of the road is a missed detection, probably too assimilated to the edge). Notice the very small number of false alarms as the black pixels on the right side of the image (column) correspond to the geometric transformation between the two acquisitions. Besides let us remark, in these two cases, that the spatial translation optimization allows to reduce the number of false positives (removal of the detection of the sky pixels). Similarly, Figure 6 shows other examples with images 5800 and 5870 in the same sequence. The approach has been run optimizing directly over the set of translations. On these examples, all objects are well-detected: the pedestrian, the motorbike and the cars, close as those about at the horizon, showing that the approach is

Figure 5: SLOD, case of moving cars and pedestrians: row-1) images at time $t = 5620$ (left) and $t = 5750$ (right); row-2) same images after quantification with 6 levels; row-3) LR images (resolution ratio: 8×8) found in the learning sequence as the most consistent with row-2 images; row-4) Detected *objects* (black pixels) in the same LR images after registration with the image at t (by estimating the spatial translation optimizing the NFA). The moving cars and pedestrians are well detected in both cases. Detected pixels on the borders are due to the fact that the left and right images are not perfectly superimposable. Notice the improvement by using the registration step: *e.g.* the sky detected in row-3 is no longer detected in row-4.

Figure 6: Some results of the SLOD: from top to bottom, the image at t (1^{st} column: image 5800, 2^{nd} column image 5870), the image at t after quantification (6 levels), and the 'correspondent' image LR after estimation of spatial translation of the image at t. The black pixels are the pixels of change or detected objects (except at the image border). Notice that even cars near the horizon and motorbike are well-detected.

robust to the size of the object to detect. Figure 7 shows the index of the image found in the learning sequence versus the index of the image at t. It illustrates the robustness of our approach for the *Localization* part of the SLOD proposed algorithm. Indeed, the obtained curve is monotonous (non-decreasing) almost everywhere, that is consistent with the fact that in both sequences (learning and current) the car did not go backwards. Besides, the flat part of the curve with dates $t \in [5650, 5700]$ corresponds to a stop of the car at the crossroads.

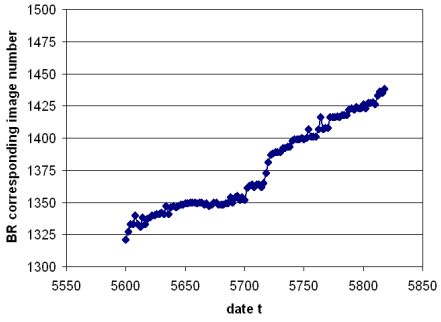


Figure 7: Index of the image found in the learning sequence versus the index of the image at t. Note the non-decreasing feature of the curve is consistent with the absence of going backward, and the flat part corresponding to a car stop.

We now illustrate qualitatively the necessity of the lowering (both for the spatial resolution and for the number

of quantification levels) in the case of the *Object Detection* part of the SLOD proposed algorithm. Figures 8 and 9 show the sensitivity to the spatial resolution ratio: from 2×2 to 16×16 (the number of quantification levels being equal to 6), and to the number of quantification levels: from 4 to 24 (the spatial resolution ratio being equal to 8×8). Concerning the spatial resolution ratio, we see that for too low ratio (2×2 or even 4×4) there are numerous false positives. Even worse sometimes these false positives form block pixels (this is particularly clear on the image at $t = 5620$) that are not present increasing the spatial resolution ratio. However, we also note that the localization of the detected objects in the image is as fuzzy as the resolution ratio is important (as illustrated in the case of the 16×16 ratio in the shown example). Concerning the number of quantification levels, we see that for too numerous levels (18 or greater) there are very numerous false positives. In the shown example, almost all the image is detected (every pixel labeled 'change') mainly due to the wrong estimation of the radiometric values of the classes. Decreasing the number of quantification levels the class value estimation is more robust and allows good result achievement. However, for too low number of quantification levels (this is obvious in the case of 1 level, and it is illustrated for 4 levels in the shown example), the information is so lowered that no change can be detected.

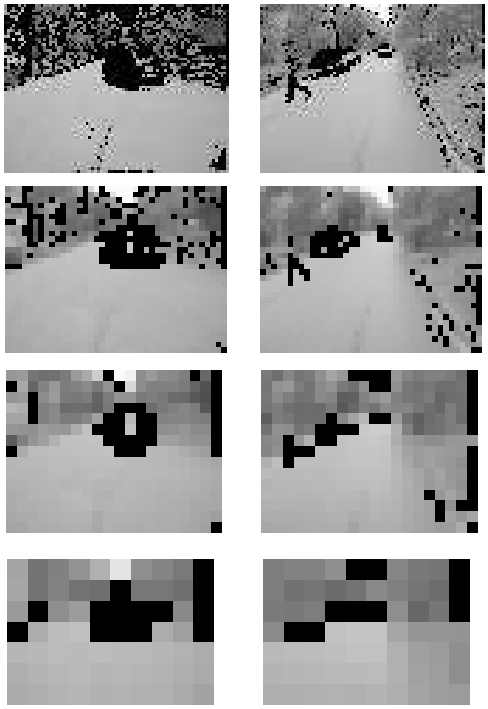


Figure 8: Sensitivity of SLOD to the spatial resolution ratio, defined as the number of HR pixels included in a BR pixel; from top to bottom: 1- ratio 2×2 (1^{st} column: $t = 5620$, 2^{nd} column $t = 5800$); 2- ratio 4×4; 3- ratio 8×8; 4- ratio 16×16. Notice the numerous false positives using a spatial resolution ratio lower than 8×8 and the object location imprecision increase with the spatial resolution ratio.

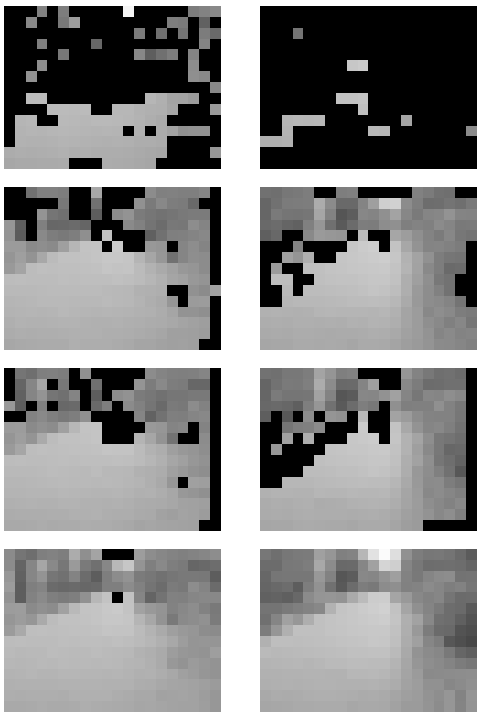


Figure 9: Sensitivity of SLOD to the number of quantification levels of the HR image. From top to bottom: row-1) 18 levels (1^{st} column: $t = 5710$, 2^{nd} column $t = 5870$); row-2) 12 levels; row-3) 6 levels; row-4) 4 levels. Notice the numerous false positives due to the class grey value wrong estimation using a too high quantification level number (here 18) and the numerous false negatives due to the information imprecision using a too low quantification level number (here 4).

6. CONCLUSIONS

In this study, we defend the idea that in order to be robust to *non-significative* changes, such as a change in illumination conditions of the scene or a change in the geometrical viewpoint of the image acquisition, the actual change or object detection could be performed between slightly imprecise images, as far as this imprecision somehow boils down to an image "filtering". We suggest two different types of such filterings: a reduction of the number of grey levels that allows a class-based comparison, and a reduction of the spatial resolution that increases the tolerance to slight misregistrations between images. Besides, using on the one hand a high spatial resolution classification and on the other hand a coarse resolution image, the algorithm firstly developed for change detection in remote sensing data in [23] can be applied.

Such an idea was applied in the context of vision-based Advanced Driver Assistance Systems to detect the objects appeared in the environment since the date of acquisition of a video learning sequence: typically pedestrians or cars on the road. From this video learning sequence (stored at coarse resolution) and a new image (quantified) we solve simultaneously the localization of the current vehicle and the detection of the changes. Our solution is based on the use of the NFA criterion that allows the comparison between different solutions in terms of image geometrical translation spatio-temporal and unchanged pixel sub-domains. Using actual data, we show the good performance of the proposed approach and establish that some compromises should be found for the spatial resolution ratio (between high and artificial coarse resolution) and for the number of quantification levels or classes considered at high resolution. Perspectives deals with the reduction of the processing time, first by subsampling of the learning sequence, and then by implementing the proposed algorithm on GPU (Graphic Processor Unit).

7. REFERENCES

[1] M. C. A. Giachetti and V. Torre. The use of optical flow for road navigation. *IEEE Trans. Robotics and Automation*, 14(1):34–48, 1998.

[2] M. Bertozzi and A. Broggi. Gold: A parallel real-time stereo vision system for generic obstacle and lane detection. *IEEE Trans. Image Processing*, 7:62–81, 1998.

[3] R. Bishop. *Intelligent Vehicle Technologies and Trends*. Artech House, Inc., 2005.

[4] L. B. E. A. Bouaziz S., Reynaud R. Experimental platform car for automatic control applications. In *Proc. of the IEEE Information and Communication Technologies Int. Symp. ICTISŠ07*, volume CDROM, pages 1–4, Fes, Maroc, 2007.

[5] C. chih Wang, C. Thorpe, M. Hebert, S. Thrun, and H. Durrant-whyte. Simultaneous localization, mapping and moving object tracking.

[6] A. S. David Gerónimo, Antonio López and T. Graf. Survey of pedestrian detection for advanced driver assistance systems. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(7):1239–1258, 2010.

[7] F. Dibos, G. Koepfler, and S. Pelletier. Fast detecting and tracking of moving objects in video scenes. 2006.

[8] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24:381–385, 1981.

[9] U. Franke and S. Heinrich. Fast obstacle detection for urban traffic situations. *IEEE Trans. Intelligent Transportation Systems*, 3(3):173–181, 2002.

[10] T. Gandhi and M. Trived. Pedestrian protection systems: Issues, survey, and challenges. *IEEE Trans. Intelligent Transportation Systems*, 8:413–430, 2007.

[11] D. Gavrila. Sensor-based pedestrian protection. *IEEE Intelligent Systems*, 16(6):77–81, 2001.

[12] W. Jones. Building safer cars. *IEEE Spectrum*, 39(1):82–85, 2002.

[13] S. Krotosky and M. Trivedi. On color-, infrared-, and multimodal-stereo approaches to pedestrian detection. *IEEE Trans. Intelligent Transportation Systems*, 8(4):619–629, 2007.

[14] A. Kuehnle. Symmetry-based recognition for vehicle rears. *Pattern Recognition Letters*, 12:249–258, 1991.

[15] M. P. L. Vlacic and F. Harashima. *Intelligent Vehicle Technologies*. Butterworth-Heinemann, 2001.

[16] J. Lisani and J. Morel. Detection of major changes in satellite images. In *IEEE ICIP*, pages 941–944, 2003.

[17] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int'l J. Computer Vision*, 60(2):91–110, 2004.

[18] A. B. M. Bertozzi and A. Fascioli. Obstacle and lane detection on argo autonomous vehicle. *IEEE Intelligent Transportation Systems*, 1997.

[19] E. H. M. Betke and L. Davis. Real-time multiple vehicle detection and tracking from a moving vehicle. *Machine Vision and Applications*, 12(2), 2000.

[20] R. R. Mounier H., Bouaziz S. A first step towards anytime invariant quasi static feedback for real time tracking. In *Proc. of the IEEE Information and Communication Technologies Int. Symp ICTISŠ04*, volume CDROM, pages 1–6, Damas, 2004.

[21] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15:353–363, 1993.

[22] C. Papageorgiou and T. Poggio. A trainable system for object detection. *Int'l J. Computer Vision*, 38(1):15–33, 2000.

[23] A. Robin. *Sub-pixel change detection qnd classification. Application to remote-sensing land-cover monitoring.* PhD thesis, Université Paris-Descartes, May 2007.

[24] A. Robin, L. Moisan, and S. Le Hégarat-Mascle. An a-contrario approach for sub-pixel change detection in satellite imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, January 2010.

[25] C. T. M. W. U. Handmann, T. Kalinke and W. Seelen. An image processing system for driver assistance. *Image and Vision Computing*, 18(5), 2000.

[26] G. B. Z. Sun and R. Miller. On-road vehicle detection: A review. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(5):694–711, 2006.