We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

5

Intelligent Surveillance System Based on Stereo Vision for Level Crossings Safety Applications

Nizar Fakhfakh, Louahdi Khoudour,
Jean-Luc Bruyelle and El-Miloudi El-Koursi
French Institute of Science and Technology for Transport, Development and Networks (IFSTTAR)
France

1. Introduction

Considered as a weak point in road and railway infrastructure, level crossings (LC) improvement safety became an important field of academic research and took increasingly railways undertakings concerns. Improving safety of persons and road-rail facilities is an essential key element to ensure a good operating of the road and railway transport. Statistically, nearly 44% of level crossings users have a bad perception of the environment which consequently increases the accidents risks Nelson (2002). However, the behavior of pedestrians, road vehicle drivers and railway operators cannot be previously estimated beforehand. According to Griffioen (2004), the human errors are the causes of 99% of accidents at LC whose 93% are caused by road users. It is important also to note the high cost related to each accident, approximately one hundred million euro per year in the EU for all level crossing accidents. For this purpose, road and railway safety professionals from several countries have been focused on providing a level crossings as safer as possible. Actions are planned in order to exchange information and provide experiments for improving the management of level crossing safety and performance. This has enabled us to discuss sharing knowledge gained from research into improving safety at level crossings.

High safety requirements for level crossing systems mean a high cost which hinders the technological setup of advanced systems. High technology systems are exploited and introduced in order to timely prevent collisions between trains and automobiles and to help reduce levels of risk from railroad crossings. Several conventional object detection systems have been tested on railroad crossings. These techniques provide more or less significant information accuracy. Any proposed system based on a technological solution is not intended to replace the present equipment installed on each level crossing. The purpose of such an intelligent system is to provide additional information to the human operator; it can be considered as support system operations. This concerns the detection and localization of any kind of objects, such as pedestrians, people on two-wheeled vehicle, wheelchairs and car drivers on the dangerous zone Yoda et al. (2006). Today, there are a number of trigger technologies installed at level crossings, but they all serve the same purpose: they detect moving object when passing at particular points in the LC. Indeed, those conventional obstacle detection systems have been used to prevent collisions between trains

and automobiles. In Fakhfakh et al. (2010), the conventional technologies applied at LC are discussed and both the advantages and drawbacks of each are highlighted.

One of the main operational purposes for the introduction of CCTV (Closed Circuit Television) at LC is the automatic detection of specific events. Some object detection vision-based systems have been tested at level crossings, and provide more or less significant information. In video surveillance, one camera, or a set of cameras, supervise zones considered as unsafe in which security must be increased Fakhfakh et al. (2011). Referring to the literature, little research has focused on passive vision to solve the problems at LC. Among the existing systems, two of them based on CCTV cameras are to be distinguished: a system using a single camera Foresti (1998). It uses a single grayscale CCD camera placed on a high pole in a corner of the LC, classifying objects as cars, bikes, trucks, pedestrians and others, and localizing them according to the camera calibration process, assuming a planar model of the road and railroad. This system is prone to false and missed alarms caused by fast illumination changes or shadows. In Ohta (2005), a second system using two cameras with a basic stereo matching algorithm and 3D background removal. This system allows detecting more or less vehicles and pedestrians, but it is extremely sensitive to adverse weather conditions. The 3D localization module is not very accurate because of the simplicity of the proposed stereo matching algorithm.

We propose in this chapter an Automatic Video-Surveillance system (AVS) for an automatic detection of specific events at level crossing. The system allows automatically and accurately detecting and 3D localizing obstacles which are stopped or in motion at the level crossing. This information can be timely transmitted to the train's driver, in a form of red lighting in the cabin, and, on his monitor, the images of such hazardous situation. So, we would be able to evaluate the risk and to warn the appropriate persons. This chapter is organized as follows: after an introduction covering the problem and the area of reserach, we describe in section 2 an overview of our proposed system for object localization at LC. Section 3 will focus on detailing the background subtraction algorithm for stationary and moving object detection from real scenes. Section 4 is dedicated to outlining a robust approach for 3D localization the objects highlighted in section 3. Results concern the object extraction and 3D localization steps are detailed in Section 6. The conclusion is devoted to a discussion on the obtained results, and perspectives are provided.

2. Overview of the AVS system

Our research aims at developing an Automatic Video-Surveillance (AVS) system using the passive stereo vision principle. The proposed imaging system uses two color cameras to detect and localize any kind of object lying on a railway level crossing. The system supervises and estimates automatically the critical situations by localizing objects in the hazardous zone defined as the crossing zone of a railway line by a road or path. The AVS system is used to monitor dynamic scenes where interactions take place among objects of interest (people or vehicles). After a classical image grabbing and digitizing step, this architecture is composed of the two following modules:

– *Background Subtraction for Moving and Stationary object detection:* the first step consists in separating the motion and stationary regions from the background. It is performed using Spatio-temporal Independent Component Analysis (stICA) technique for high-quality motion detection. The color information is introduced in the ICA algorithm that models

the background and the foreground as statistically independent signals in space and time. Although many relatively effective motion estimation methods exist, ICA is retained for two reasons: first, it is less sensitive to noise caused by the continuously environment changes over time, such as swaying branches, sensor noise, and illumination changes. Second, this method provides clear-cut separation of the objects from the background, and can detect objects that remain motionless for a long period. Foreground extraction is performed separately on both cameras. The motion detection step allows focusing on the areas of interest, in which 3-D localization module is applied.

– *3-D localization of Moving and Stationary object detection:* this process applies a specific stereo matching algorithm for localizing the detected objects. In order to deal with poor quality images, a selective stereo matching algorithm is developed and applied to the moving regions. First, a disparity map is computed for all moving pixels according to a dissimilarity function entitled Weighted Average Color Difference (WACD) detailed in Fakhfakh et al. (2010). An unsupervised classification technique is then applied to the initial set of matching pixels. This allows to automatically choose only well-matched pixels. A pixel is considered as well-matched if the pair of matched pixels have a confidence measure higher than a threshold. The classification is performed applying a Confidence Measure technique detailed in Fakhfakh et al. (2009). It consists in evaluating the result of the likelihood function, based on the *winner-take-all* strategy. However, the disparities of pixels considered as badly-matched are then estimated applying a hierarchical belief propagation technique detailed further. This allows obtaining, for each obstacle, a high accurate dense disparity map.

3. Background subtraction by spatio-temporal independent component analysis

3.1 State of the art

Complex scenes acquired in outdoor environments require advanced tools to be dealt with, for instance, sharp brightness variation, swaying branches, shadows and sensor noise. The use of stationary cameras restricts the choice of techniques to those based on temporal differencing and background subtraction. The latter aims at segmenting foreground regions corresponding to moving objects from the background, somehow by evaluating the difference of pixel features between a reference background and a current scene image. This kind of technique requires updating the background model over time by modeling the possible states that a pixel can take. A trade-off is to be found between performing a real time implementation and handling background changes which are caused by gradual or sudden illumination fluctuations and moving background objects.

The pixel-based techniques assumes statistical independence between the intensity at each pixel throughout the training sequence of images. The main drawback is that it is not effective to model a complex scene. A mixture of Gaussian distribution (GMM) Stauffer & Grimson (2000) have been proposed to model complex and non-static scenes. It consists of modeling the background as a constant or adaptive number of Gaussians. A relatively robust non-parametric method has been proposed in Elgammal et al. (2000). The authors estimate the density function of a distribution given only very recent history information. This method allows obtaining a sensitive detection. In Zhen & Zhenjiang (2008) the authors use an improved GMM and Graph Cut to minimize an energy function to extract foreground objects. The main disadvantage is that the fast variations cannot be accurately modeled.

Another kind of technique, called codebook model, has recently been proposed. It consists in registering, over a long period of time, the possible states of each pixel in what is called a *codebook* Kim et al. (2005) consisting of a set of *codewords*. A pixel is classified into either background or foreground classes by evaluating the difference between a given pixel and the corresponding codebook. A color metric and brightness boundaries are used as criteria for classification. Hence, the existing techniques can handle gradual illumination changes, but remain vulnerable to sudden changes. Several works are mainly focused on how to make foreground object extraction unaffected by background changes. These methods are very sensitive on the background model so that a pixel is correctly classified when a given image is coherent to its corresponding background model. Another issue is in the huge computational time of the background updating process.

In recent years, another kind of technique has emerged to deal with this issue. The Independent Component Analysis (ICA) technique, known for its robustness in the signal processing field, is getting much attention in the image processing field. The purpose of ICA is to restore statistically independent source signals, given only a mixture of these signals. For a short time, ICA is applied to fMRI data by McKeown et al. (1998) and have been then introduced for solving problems related to image processing. Hence, ICA finds applications in many emerging new application areas, such as feature extraction Delfosse & Loubaton (1995), speech and image recognition Cardoso (1997), data communication Oja et al. (1991), sensor signal processing Cvejic et al. (2007) and biomedical signal processing Dun et al. (2007)Waldert (2007).

More recently, ICA has been introduced in video processing to cope with the issue of foreground estimation. Zhang and Chen Zhang & Chen (2006) have introduced a spatio-temporal independent component analysis method (stICA) coupled with multiscale analysis as a postprocessing for automated content-based video processing in indoor environments. Their system is computationally demanding so that the data matrix, from which the independent components must be estimated using ICA, is of a very large dimension. Recently, Tsai and Lai Tsai & Lai (2009) have proposed an ICA model for foreground extraction without background updating in indoor environments. The authors have proposed an algorithm for estimating the de-mixing matrix, which gives the independent components, directly measuring the statistical independence by estimating the joint and marginal probability density functions from relative frequency distributions. But, neither detail of an automated system is proposed. These two related works do not handle the background changes over time and are limited to monochrome images. Furthermore, their algorithms are only tested in indoor environments characterized by small environmental changes.

3.2 Overview of the proposed background subtraction algorithm

The proposed scheme is a complete modelization of the background subtraction task from an image sequence in real-world environments. While considering the acquisition process achieved, the block diagram of the proposed framework is given by Figure 1. The algorithm can be devided into two complementary steps: training step and detection step.

− The first step consists of the estimation of the de-mixing matrix parameter by performing the ICA algorithm on background images only. While any foreground object may appear in

the background images, the ICA algorithm allows estimating a source which represents the temporal difference between pixels. Typically, only the five most recent background images seem to be sufficient in our experiments. The matrix which allows separating the foreground from its background, termed de-mixing matrix, is estimated in the following way: the ICA algorithm is performed only once on a data matrix from which the independent components, i.e. the background and the foreground, will be estimated. The data matrix is constructed from two images which are the most recent background, and another on which a foreground object is arbitrarily added. The de-mixing matrix will be used in the detection step.

– The detection step consists of the approximation and the extraction of foreground objects. However, the data matrix is constructed from two images; one is an incoming image from the sequence and the other is the most recent available background. The approximated foreground is then obtained simply by multiplying the data matrix with the de-mixing matrix. The approximated foreground is filtered in order to effectively segment the true foreground objects. This is performed by the use of a spatio-temporal belief propagation method. The principal guidelines of our framework can be explained and summarized in Algorithm 1.

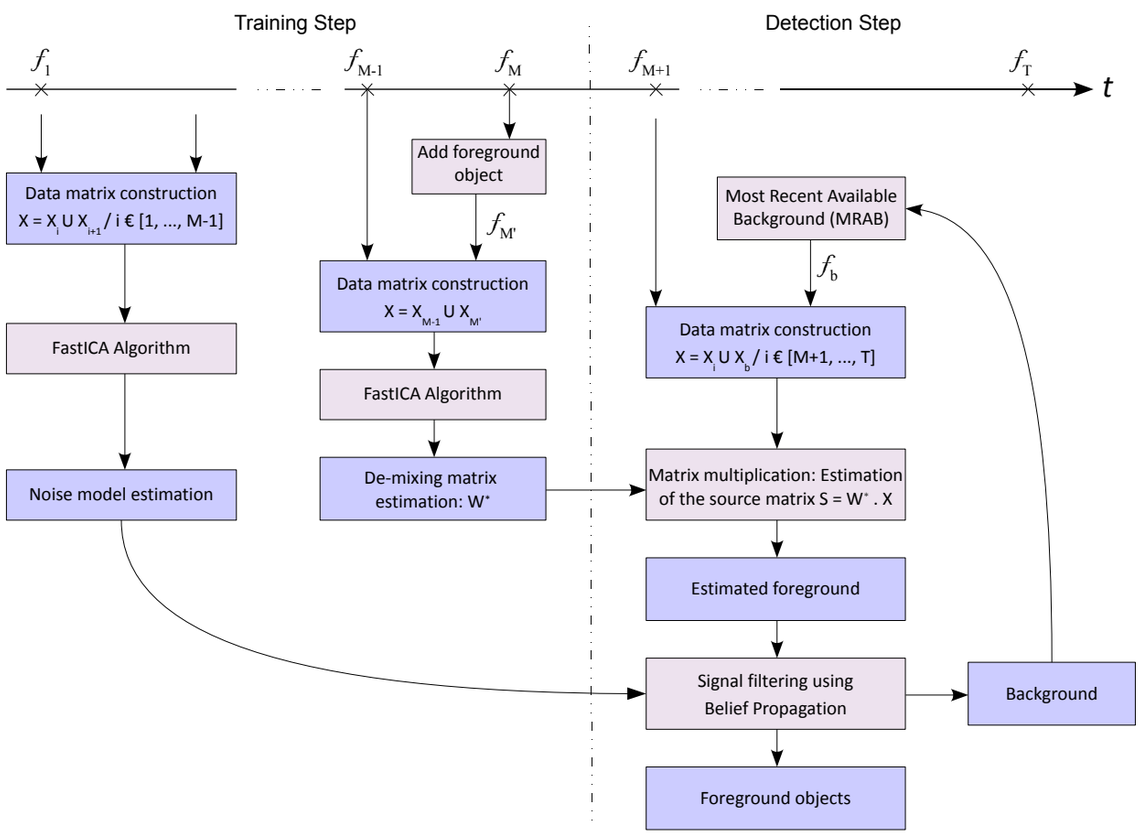


Fig. 1. The block diagram of the proposed background subtraction agorithm.

ICA can be defined as a statistical and computational technique for revealing hidden factors that underlie sets of random variables, measurements or signals. ICA defines a generative

Algorithm 1 Background subtraction for Foreground objects segmentation.

1. Perform the *FastICA* algorithm on a data matrix obtained from two consecutive background images for noise model estimation. The noise model, obtained from the k previous frame, corresponds to the mean and the standard deviation of each color component.

2. Excecute the *FastICA* algorithm only once in order to estimate the de-mixing matrix. The data matrix, from which the *FastICA* algorithm is performed, is constructed of two images: the one corresponds to the background and the other corresponds to the background on which a foreground object is added.

3. Construct the data matrix for foreground approximation. The data matrix is composed of the Most Recent Available Background and an incoming image from the sequence.

4. The approximated foreground is obtained by multiplying the data matrix with the de-mixing matrix obtained from step 2.

5. Filtering of the estimated foreground by the use of a spatio-temporal belief propagation.

model for separating the observed multivariate data that are mixtures of unknown sources without any previous knowledge. It aims to find the source signals from observation data. In that model, the observed data are assumed to be linear mixtures of some unknown latent variables, and the mixing system is also unknown. The latent variables are assumed to be non-Gaussian and mutually independent; they are called the independent components of the observed data. The problem boils down to finding a linear representation in which the components are as statistically independent as possible. Formally, the data are usually represented by a matrix X, the unknown separated source signals by $\tilde{S}$ and the unknown matrix that allow obtaining the independent components by $\tilde{A}$. Thus, every component, say $\tilde{s}_{i,t}$ from $\tilde{S}$, is expressed as linear combination of the observed variables, the problem can be reformulated as Equation 1:

$$X = \tilde{A} \times \tilde{S} \tag{1}$$

After estimating the matrix A, its inverse, called W, can be computed for obtaining the independent components. The model becomes:

$$\tilde{S} = \tilde{A}^{-1} \times X = \tilde{W} \times X \tag{2}$$

3.3 Proposed background subtraction using ICA model

We explain in this section the different parts of our proposed model and how the *FastICA* algorithm is performed for solving ICA model. The number of independent components to be estimated are the two following components: the background and the foreground. Moreover, these two components are assumed to be independent. Indeed, the presence of motion is

characterized by a high variation of the intensity of a given pixel. It is to be noted that the presence of an arbitrary foreground object over a background is an independent phenomenon. That is, the intensity of a pixel in a foreground object does not depend on the intensity of its corresponding background.

To begin with, we formulate a video sequence as a set $\mathcal{I}$ of sequential images. An image captured at time t is termed I_t, where $t = 1, \ldots, T$, T is the number of frame in the sequence and $\mathcal{I} = \bigcup_{t \in T} I_t$. We extend the ICA model to exploit the color information for video objects extraction. Each image I_t is a matrix of size $K = h \times w$, where h and w are respectively the height and the width of the image. An observation, noted $I_{p,t}$, corresponds to a pixel sample at location $p = (i, j)$ at time t. Knowing that the color information is considered in the design of our framework, we introduce the parameter c which represents a component of a color space or a normalized color space. For instance, c means either the Red, Green or Blue component in the RGB color space, i.e. $\mathbb{R}^3$. For reasons of simplicity, c^i means one component in the RGB color space where c^1 means the Red, c^2 means the Green and c^3 means the Blue. The data matrix, termed X, is a matrix of two lines and $w \times h \times 3$ columns. To fit the data matrix, each color component of the image I^c at time t is resized to be a column vector of size $h \times w$. Each line of the matrix X is a column vector, V, consisting of the three adjaxent color components $V_k = \langle I_k^{c^1}, I_k^{c^2}, I_k^{c^3} \rangle_t$, where $k \in \{bg, bg + ob\}$ so that the first line represents the background image while the second line represents an image having an arbitrary foregrounds. bg and $bg + ob$ correspond to the background, and the the background on which an object is added. The estimated de-mixing matrix W is a 2-by-2 square matrix and the estimated sources signals $\tilde{S}$ has the same size as the data matrix X. The matrix X obtained at time t is given as follows:

$$X = \begin{pmatrix} V_{bg} \\ V_{bg+ob} \end{pmatrix}_t = \begin{pmatrix} I_{bg}^{c^1}, & I_{bg}^{c^2}, & I_{bg}^{c^3} \\ I_{bg+ob}^{c^1}, & I_{bg+ob}^{c^2}, & I_{bg+ob}^{c^3} \end{pmatrix}_t \tag{3}$$

3.3.1 Foreground approximation

The *FastICA* algorithm is performed only once for initializing the detection process which allows estimating the de-mixing matrix. The data matrix in the detection step is constructed in a different way from that of the training step. The data matrix is formed by both an image containing only the recent background image in the sequence, and another image containing an arbitrary foreground object, if any. The estimated source images correspond to the separated background and only foreground: the one represents only the background source and the other highlights the foreground object, without the detailed contents of the reference background. Figure 2 illustrates the inputs and outputs of the algorithm. The estimated de-mixing matrix will be used to extract the foreground objects from their background during the detection step. In the detection step, the source components are extracted simply by multiplying the data matrix with the estimated de-mixing matrix. Therefore, the data matrix is updated for each incoming image in the sequence in the following way: the second row of the data matrix corresponds to the recent incoming image from the sequence, while the first row corresponds to the Most Recent Available Background (MRAB).

Using this configuration, the two images which constitute the data matrix are not very different because of their temporal proximity. The existing noise among two consecutive images does not degrade the ICA performances and still allows an estimation of the hidden

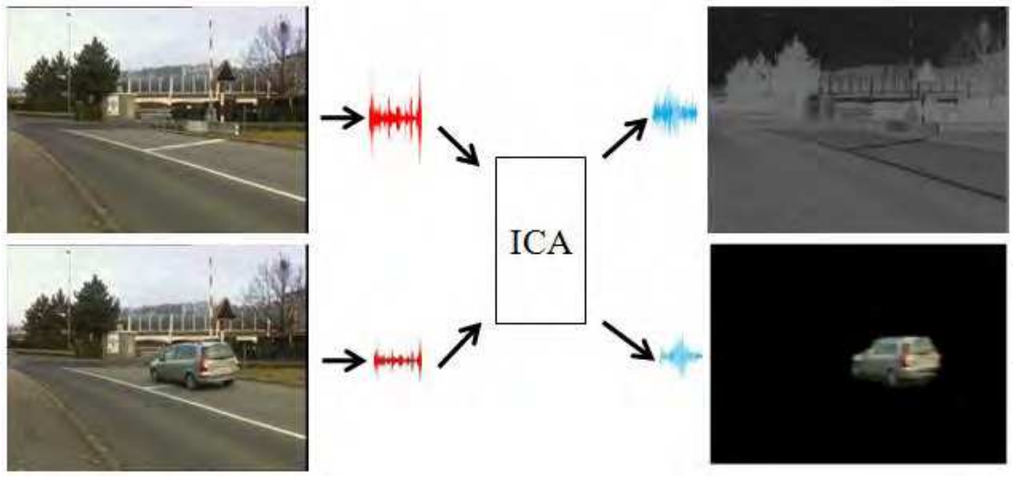


Fig. 2. Principle of the background subtraction using Independent Component Analysis.

sources. The estimated signals are obtained by multiplying the data matrix and the de-mixing matrix.

$$\tilde{S} = \begin{pmatrix} \tilde{V}_{bg} \\ \tilde{V}_{bg+ob} \end{pmatrix} = \begin{pmatrix} \mathrm{w}_1 \\ \mathrm{w}_2 \end{pmatrix} \times \begin{pmatrix} V_{bg} \\ V_{bg} \end{pmatrix} \tag{4}$$

The first row of the matrix $\tilde{S}$ corresponds to the background model, while the second row represents the estimated foreground signal in which only moving or stationary objects are highlighted. From then on, only the estimated foreground will be taken into account and will be called an "Approximate Foreground". The second row of the matrix $\tilde{S}$ is reshaped from a vector of size $h \times w \times 3$ to a 2D color image of size (h, w) by the linear transformation given by Equation 5:

$$I^c_{i,j,t} = \tilde{S}(1, (l * K) + (i * h) + j) \tag{5}$$

where $c \in \{R, G, B\}$, $K = h \times w$, and l is an integer which takes values 1 if $c = R$, 2 if $c = G$, and 3 if $c = B$. Figure 3 despicts an example of foreground objects extracted by multiplying a de-mixing matrix, estimated using *FastICA* algorithm, and a data matrix formed by the images (a) and (b). The two vectors that form the de-mixing matrix are respectively $\mathrm{w}_1 = [0.0285312, 0.0214363]$ and $\mathrm{w}_2 = [-0.519938, 0.548299]$. Vector w_1 allows obtaining the estimated background image while vector w_2 allows obtaining the estimated foreground image. Typically, the estimated foreground highlight moving and stationary objects which are smothered by a noisy and uniform background.

The estimated signal is characterized by the presence of zones corresponding to a high intensity variation of the background together with a lot of noise. We have no *a priori* about the noise distribution, making the foreground extraction a task difficult to solve.

3.3.2 Foreground extraction

A. MRF Formulation and Energy Minimization

In this part, we propose a robust framework to accurately extract foreground objects from the estimated foreground signal. This module aims at filtering the estimated foreground by reducing the noise obtained from the ICA model. The problem can be expressed in terms of

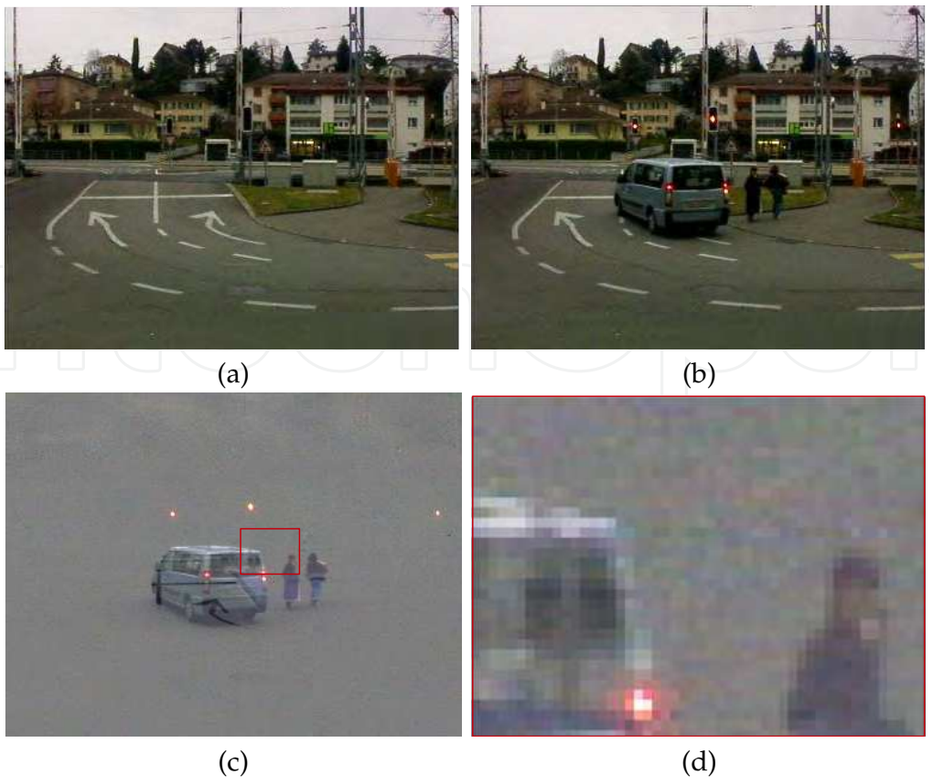


Fig. 3. Background subtraction and foreground approximation (a) background image from the "Pontet" dataset, (b) scene image from the same dataset containing a car and two pedestrians, (c) estimated foreground image obtained by mutliplying the data matrix, formed by the images (a) and (b), and a de-mixing matrix, where $w_2 = [-0.519938, 0.548299]$, and (d) zooming on a part of a background, a pedestrian and a car that we attempt to separate.

a Markov Random Field (MRF) in which an inference algorithm is applied to find the most likely setting of the model. Several robust inference algorithms such as Graph Cuts and Belief Propagation have emerged and proved their efficiency especially in the realm of stereo Yang et al. (2009) and image restoration Felzenszwalb & Huttenlocher (2006). The formulation we propose aims at clairly separating the foreground from its background by introducing spatial and temporal dependencies between the pixels. The rest of this section will be focused on the spatio-temporal formulation and the algorithm used for minimizing such energy. In what follows, the problem will be formulated as graphical model which consists of an undirected graph on which the inference is performed by approximating the MAP estimate for this MRF using loopy belief propagation. The bipartite graph is denoted by $\mathcal{G} = (\mathcal{P}, E)$ where $\mathcal{P}$ is a set of nodes, i.e. pixels, and E is a set of undirected edges between nodes. Each pixel x is modeled by a state noted $\dot{s}_x$. In computer vision, the edges allow establishing spatial dependencies between nodes. During the message passing procedure, a label is assigned to each node which is the vector of three color components. A state $\dot{s}_x = \langle l_t, \ldots, l_{t-k} \rangle$ of a pixel x is modeled by a vector of labels such as a label l_t corresponds to the color components of pixel p at time t.

Referring to the standard four-connected rectangular lattice configuration, the joint probability of the MRF is given by the product of one- and two-nodes having spatial and temporal dependencies as follows:

$$P(\mathcal{G}) = \prod_{x \in \mathcal{P}} \Phi(\dot{s}_{x(t)}) + \prod_{\substack{x \in \mathcal{P} \\ y \in \mathcal{N}_{s,x}}} \Psi(\dot{s}_{x(t)}, \dot{s}_{y(t)}) + \prod_{x(t-i) \in \mathcal{N}_{t,x}} \Theta(\dot{s}_{x(t)}, \dot{s}_{x(t-i)}) \tag{6}$$

where Φ, Ψ and Θ are functions which describe the dependency between nodes, which will be detailed in section *B*. $\dot{s}_x$ and $\dot{s}_y$ are the state of node x and y respectively, given that y is one of the four spatial neighbors of node x, $\dot{s}_{x(t)}$ is the state of the most recent node x at time t. For a given pixel x, the spatial four-connected nodes form a set of spatial neighboring noted $\mathcal{N}_{s,x}$, and the consecutive temporal neighboring denoted by $\mathcal{N}_{t,x}$. Typically, the optimization is performed by computing the *a posteriori* belief of a variable, which is NP-hard. This has generally been viewed as being too slow to be practical for early vision. The idea is to approximate the optimal solution by inference using belief propagation which is one of the most efficient methods of finding the optimum solution. This allows minimizing the energy function using either the Maximum A Posteriori (MAP) or the Minimum Mean Squared Error (MMSE) estimator.

B. Energy Minimization using Spatio-Temporal Belief Propagation

Intuitively, the objectives can be reformulated, in terms of energy minimization, as the research of the optimal labeling f^* that assigns each pixel $x \in \mathcal{P}$ a label $l \in \mathcal{L}$ by minimizing an energy function. In our case, the energy to minimize is represented as a linear combination of three terms: data term, spatial smoothness term, and temporal filtering term. The data term measures how well state $\dot{s}$ fits pixel x, given its observed data, the spatial smoothness term measures the extent to which the state $\dot{s}$ is not spatially piecewise smooth, and the temporal filtering term evaluates the temporal dependencies of the consecutive states of a pixel x over time by using its known previous optimal labels. Checking both the piecewise spatial smoothness and temporal filtering allows obtaining a robust motion-based classification.

Intuitively, the optimal labeling can be found by maximizing a probability. The MAP estimate is equivalent to minimizing the Equation 6 by taking the negative log, so writing $\phi = -\log \Phi$, $\psi = -\log \Psi$, and $\theta = -\log \Theta$, the objectibe can be reformulated as minimizing the posterior log as a function of the form:

$$E(\mathcal{G}) = \sum_{x \in P} \phi(\dot{s}_{x(t)}) + \sum_{\substack{x \in \mathcal{P} \\ y \in \mathcal{N}_{s,x}}} \psi(\dot{s}_{x(t)}, \dot{s}_{y(t)}) + \sum_{x(t-i) \in \mathcal{N}_{t},x} \theta(\dot{s}_{x(t)}, \dot{s}_{x(t-i)}) \tag{7}$$

The formulation we consider in our framework consists in optimizing the data term denoted by $E_{data}(\hat{f})$, the spatial smoothness term denoted by $E_{s\_smooth}(\hat{f})$ and the temporal filtering term denoted by $E_{t\_filtering}(\hat{f})$. By treating the problem in the context of filtering, the outputs from previous frames can be incorporated by adding an extra term to the energy function. The Global Energy Minimization function can be formulated as follows:

$$E(\mathcal{G}) = E_{data}(\hat{f}) + E_{s\_smooth}(\hat{f}) + E_{t\_filtering}(\hat{f}) \tag{8}$$

However, Equation 7 can be expressed as:

$$E(\mathcal{G}) = \sum_{x \in P} D_x(\dot{s}_{x(t)}) + \sum_{\substack{x \in \mathcal{P} \\ y \in \mathcal{N}_{s,x}}} V_s(\dot{s}_{x(t)}, \dot{s}_{y(t)}) + \sum_{x(t-i) \in \mathcal{N}_t, x} V_t(\dot{s}_{x(t)}, \dot{s}_{x(t-i)}) \tag{9}$$

– Data term : In stereo problems, data term usually corresponds to the cost of matching a pixel in the left-image to another in the right-image. Typically, this term, i.e. cost, is based on the intensity differences between the two pixels. Just like in this present case, the data term $D_x(\dot{s}_x)$ is defined as the cost of assigning a label l_x to pixel x. It can be expressed as the Euclidean distance between the color components of the pixel x and a given label l_x. In order to control the evolution of the energy function, the data term can be written as:

$$D_x(\dot{s}_x) = \begin{cases} 0 & \text{if } |\dot{s}_{x(t)} - l_x| \leq \varepsilon \\ |\dot{s}_{x(t)} - l_x| & \text{otherwise} \end{cases} \tag{10}$$

where ε is constant. The data term depends only on the parameter α of the state of the node. The rest of parameters are not considered for the computation of this term.

– Spatial smoothness term : The choice of the smoothness term is a critical issue. Several cost terms have been proposed which heavily depend on the problem to be solved. Assuming the pair-wise interaction between two adjacent pixels x and y, the spatial smoothness term can be formulated using the Potts model Wu (1982). This is motivated by the fact that this piecewise smooth model encourages a labeling consisting of several regions where pixels in the same moving region have similar labels. The cost of the spatial smoothness term is given by:

$$V_s(\dot{s}_{x(t)}, \dot{s}_{y(t)}) = \begin{cases} 0 & \text{if } \Delta_{x,y} \leq \xi \\ \Delta_{x,y}.T & \text{otherwise} \end{cases} \tag{11}$$

where $\Delta_{x,y}$ is the Euclidean distance between the two neighboring pixels x and y at the same time, ξ is a constant, and T is the temperature variable of the Potts model which can be estimated appropriately by simulating the Potts model. In our case, we choose to take T as a constant.

– Temporal filtering term Making use of an additional term which represents the temporal filtering in the energy function is very useful for improving the performances of the passing message procedure. The optimal labels obtained for a node during the k previous images are used for quickly reaching the optimal label in the current node at time t. Each current node uses its previous best set of labels obtained for the same node during the k previous images. The temporal filtering term is given by the following:

$$\theta\left(\dot{s}_{x(t)}, \dot{s}_{x(t-i)}\right) = \sum \kappa . \left\| \dot{s}_{x(t)} - \dot{s}_{x(t-i)} \right\| \tag{12}$$

where κ is a binary parameter which takes values 0 or 1. In the case where the most temporally neighboring pixel is classified as a background, the parameter κ is set to 1 for all temporally neighboring pixels which are classified as background, 0 otherwise.

C. Foreground/Background Classification

This final step consists in automatically extracting all foreground objects from their background in the aforementioned filtered foreground image, noted I_f. The classification step is preceded by postprocessing the output color image on which the background is uniform, while all moving and stationary objects are well highlighted. The postprocessing aims at binarizing the image and classifying all pixels into two classes: foreground and background. To this end, the color components $I_f^{c_i}$, where c_i is the i^{th} color component, are extracted from the color image I_f. Then, a Sobel operator *Sob* is applied to each color component which aims at performing a 2-D spatial gradient measurement on an image and so emphasizes regions of high spatial frequency that correspond to edges. For each color component, two edge images are obtained which represent the edges obtained from the horizontal and the vertical directions. The union of these two images forms the V-H (Vertical-Horizontal) edges of the image. Thus, an edge image is obtained for each color component. The final edge image is obtained by intersecting the points of the edges of the three images. The final edges E of the image is obtained as follows.

$$E(I_f^{c_i}) = Sob(I_f^{c_i})_{dx} \cup Sob(I_f^{c_i})_{dy} \tag{13}$$

$$E(T_f) = E(I_f^{c_1}) \cap E(I_f^{c_2}) \cap E(I_f^{c_3}) \tag{14}$$

The final map which contains only moved and stationary onjects is obtained by merging $E(T_f)$ and another map $S(T_f)$. The $S(T_f)$ map is obtained by a color segmentation process applied on the obtained filtering image.

4. 3D localization of obstacles by stereo matching

The use of more than one camera provides additional information, such as the depth of objects in a given scene. Dense or sparse stereovision techniques can be used to match points. When a point is imaged from two different viewpoints, its image projection undergoes a displacement from its position in the first image to that in the second image. Each disparity, determined for each pixel in the reference image, represents the coordinate difference of the projection of a real point in the 3-D space, i.e. scene point, in the left- and right-hand images of the cameras. A depth map is obtained from the two images and, for each disparity, the corresponding real 3-D coordinates are estimated according to the intrinsic and extrinsic parameters, such as the focal length and the baseline. The amount of displacement, alternatively called disparity, is inversely proportional to distance and may therefore be used to compute 3D geometry. Given a correspondence between imaged points from two known viewpoints, it is possible to compute depth by triangulation.

Several well-known stereo algorithms compute an initial disparity map from a pair of images under a known camera configuration. These algorithms are based loosely on local methods, such as window correlation, which take into account only neighborhood points of the pixel to be matched. The obtained disparity map has a lot of noise and erroneous values. This noise concerns mostly the pixels belonging to occluded or textureless image regions. An iterative process is then applied to the initial disparity map in order to improve it. These methods use global primitives. Some research has used a graph-based method Foggia et al. (2007) and color segmentation based stereo methods Taguchi et al. (2008) which belong to what is called

global approaches. Other approaches have been proposed: they are based on a probabilistic framework optimization, such as belief propagation Lee & Ho (2008). These methods aim to obtain high-quality and accurate results, but are very expensive in terms of processing time. It is a real challenge to evaluate stereo methods in the case of noise, depth discontinuity, occlusions and non-textured image regions.

In our context, the proposed stereo matching algorithm is applied only on moving and stationary zones automatically detected referring to the technique discussed in the previous section. However, costs and processing time are decreasing at a steady pace, and it is becoming realistic to believe that such a thing will be commonplace soon. The stereo algorithm presented further, springs from relaxation and energy minimization fields. Our approach aims to represent a novel framework to improve color dense stereo matching. As a first step, disparity map volume is initialized applying a new local correlation function. After that, a confidence measure is attributed to all the pairs of matched pixels, which are classified into two classes: well-matched pixel and badly-amtched pixel. Then, the disparity value of all badly-matched pixels is updated based only on stable pixels classified as well-matched in an homogeneous color region. The well-matched pixels are used as input into disparity re-estimation modules to update the remaining pixels. The confidence measure is based on a set of original local parameters related to the correlation function used in the first step of our algorithm. This paper will focus on this point. The main goal of our study is to take into account both quality and speed. The global scheme of the stereo matching algorithme we propose is given by figure 4.

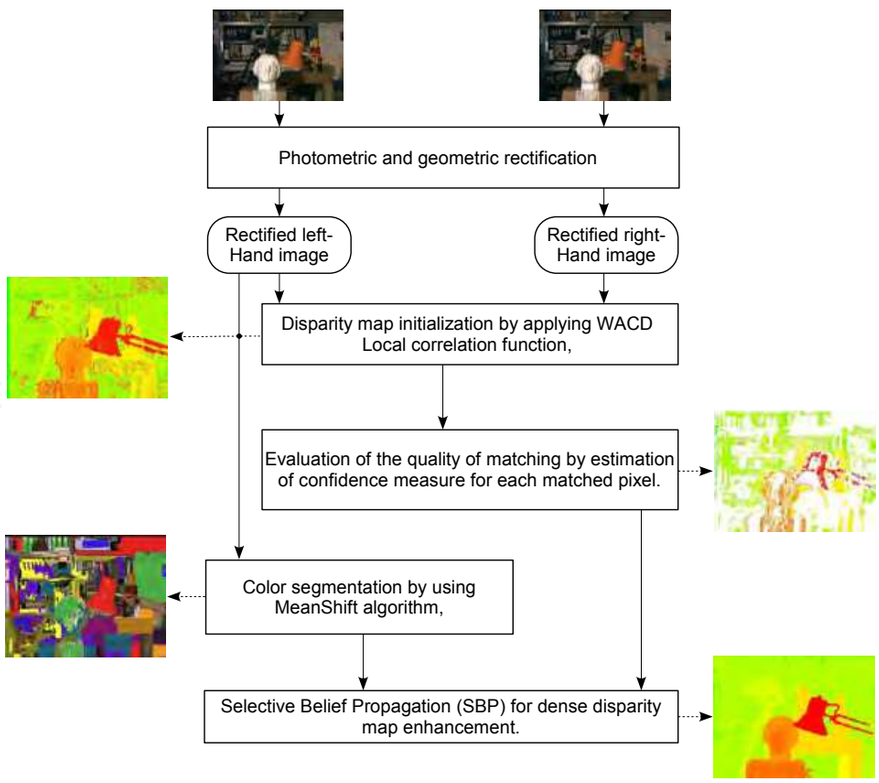


Fig. 4. Global scheme of the stereo matching algorithm for 3D localization.

4.1 Global energy formulation

The global energy to minimize is composed of two terms: data cost and smoothness constraint, noted f and $\hat{f}$ respectively. The first term, f, allows evaluating the local matching of pixels by attributing a label l to each node in a graph $\mathcal{G}$. The second term, $\hat{f}$, allows evaluating the smoothness constraint by measuring how well label l fits pixel p given the observed data. Some works Felzenszwalb & Huttenlocher (2006) Boykov et al. (2001) consider the smoothness term as the amount of difference between the disparity of neighboring pixels. This can be seen as the cost of assigning a label $l^{'}$ to a node during the inference step, where $\mathcal{L}$ is the set of all the possible disparity values for a given pixel. The Global Energy Minimization function can be formulated as follows (Equation. 15):

$$E(\mathcal{G}) = E_{l\in\mathcal{L}}(f) + E_{l'\in\mathcal{L}}(\hat{f}) \tag{15}$$

The minimization of this energy is performed iteratively by passing messages between all the neighboring nodes. These messages are updated at each iteration, until convergence. However, a node can be represented as a pixel having a vector of parameters such as, typically, its possible labels (i.e. disparities). However, reducing the complexity of the inference algorithm leads in most cases to reduced matching quality. Other algorithm variants can be derived from this basic model by introducing additional parameters in the message to be passed. One of the important parameters is the spatio-colorimetric proximity between nodes Trinh (2008).

– The data term we propose can be defined as a local evaluation of attributing a label l to a node. It is given by Equation 16:

$$E_{l\in\mathcal{L}}(f) = \sum_{p} \alpha\ \phi^{x\to x',y}(z_1) \tag{16}$$

Where $\phi^{x\to x',y}(z_m)$ is the m^{th} dissimilarity cost obtained for each matched pixel $(p^{x',y}, p^{x,y})$. Parameter α is a fuzzy value within the [0,1] interval. It allows to compute a confidence measure for attributing a disparity value d to the pixel p. α is given by Equation 17:

$$\alpha = \begin{cases} \psi(p^{x\to x',y}) & \text{if } \psi(p^{x\to x',y}) \geqslant \varrho \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

Where $\psi(p^{x\to x',y})$ is a confidence measure computed for each pair $(p^{x,y}, p^{x',y})$ of matched pixels and ϱ is a confidence threshold. The way of computing the confidence measures.

– The smoothness term is used to ensure that neighboring pixels have similar disparities.

The two-frame stereo matching approaches allow computing disparities and detecting occlusions, assuming that each pixel in the input image corresponds to a unique depth value. The stereo algorithm described in this section stems from the inference principle based on hierarchical belief propagation and energy minimization.

It takes into account the advantages of local methods for reducing the complexity of the Belief Propagation method which leads to an improvement in the quality of results. A Hierarchical Belief Propagation (HBP) based on a Confidence Measure technique is proposed: first, the data term (detailed in Section 4.1) is computed using the WACD dissimilarity function. The

obtained disparity volume allows initializing the Belief Propagation graph by attributing a set of possible labels (i.e. disparities) for each node (i.e. pixels). The originality is to consider a subset of nodes among all the nodes to begin the inference algorithm. This subset is obtained thanks to a confidence measure computed at each node of a graph of connected pixels. Second, the propagation of messages between nodes is performed hierarchically from the nodes having the highest confidence measure to those having the lowest one. A message is a vector of parameters (e.g. possible disparities, (x, y) coordinates, etc.) that describes the state of a node. To begin with, the propagation is performed inside an homogeneous color region and then passed from a region to another. The set of regions are obtained by a color-based segmentation using the MeanShift method Comaniciu & Meer (2002). A summary of our algorithm is given in Algorithm 2:

Algorithm 2 Hierarchical Belief Propagation.

1) Initialize the data cost for nodes in the graph using the method in Fakhfakh et al. (2010).
2) Compute a Confidence Measure $\psi(p^{x \to x',y})$ for each node.
3) Repeat steps a, b, c and d for each node
  a) Select node (i.e. pixel) $Node_i$, i being the
    node number, having a data term lower than a confidence threshold ϱ.
  b) Select the k-nearest neighbor nodes within a cubic 3D support
    window that have a $\psi(p^{x \to x',y})$ greater than ϱ.
  c) Update the label of the current node.
  d) Update the weight of the current node.
4) Repeat step 3) until reaching minimal energy.

4.2 Selective matching approach

Using the WACD dissimilarity function allows initializing the set of labels. It represents a first estimate of the disparity map which contains matching errors. Then, each pair of pixels is evaluated using the Confidence Measure method described in Fakhfakh et al. (2009). The likelihood function used to initialize the disparity set is applied to each pixel of the image. Furthermore, for each matched pair of pixels a confidence measure is computed. It is termed $\psi(p_l^{x,y}, p_r^{x',y})$ which represents the level of certainty of considering a label l as the best label for pixel p. This confidence measure function depends on several local parameters and is given by Equation 18:

$$\psi(p_l^{x,y}, p_r^{x',y}) = P(p_r^{x',y} / p_l^{x,y}, \rho, min, \sigma, \omega) \tag{18}$$

The confidence measure with its parameters is given by Equation 19:

$$\psi(p_l^{x,y}, p_r^{x,y'}) = \left(1 - \frac{min}{\omega}\right)^{\tau^2 log(\sigma)} \tag{19}$$

Where

– *Best Correlation Score (min):* The output of the dissimilarity function is a measure of the degree of similarity between two pixels. Then, the candidate pixels are ranked in increasing order according to their corresponding scores. The couple of pixels that has the minimum score is considered as the best-matched pixels. The lower the score, the better the matching.

The nearer the minimum score to zero, the greater the chance of the candidate pixel to be the actual correspondent.

– *Number of Potential Candidate Pixels (τ):* This parameter represents the number of potential candidate pixels having similar scores. τ has a big influence because it reflects the behavior of the dissimilarity function. A high value of τ means that the first candidate pixel is located in a uniform color region of the frame. The lower the value of τ, the fewer the candidate pixels. If there are few candidates, the chosen candidate pixel has a greater chance of being the actual correspondent. Indeed, the pixel to be matched belongs to a region with high variation of color components. A very small value of τ and a *min* score close to zero, means that the pixel to be matched probably belongs to a region of high color variation.

– *Disparity variation of the τ pixels (σ):* A disparity value is obtained for each candidate pixel. For the τ potential candidate pixels, we compute the standard deviation σ of the τ disparity values. A small σ means that the τ candidate pixels are spatially neighbors. In this case, the true candidate pixel should belong to a particular region of the frame, such as an edge or a transition point. Therefore, it increases the confidence measure. A large σ means that the τ candidate pixels taken into account are situated in a uniform color region.

– *Gap value (ω):* This parameter represents the difference between the τ^{th} and $(\tau+1)^{th}$ scores given with the dissimilarity function used. It is introduced to adjust the impact of the minimum score.

To ensure that function ψ has a value between 0 and 1, a few constraints are introduced. The *min* parameter must not be higher than ω. If so, parameter ω is forced to $min+1$. Moreover, the $log(\sigma)$ term is used instead of σ, so as to reduce the impact of high value of σ and obtain coherent confidence measures. The number τ of potential candidate pixels is deduced from the $\mathcal{L}$ scores obtained with the WACD likelihood function. The main idea is to detect major differences between successive scores. These major differences are called main gaps. Let ϕ denote a discrete function which represents all the scores given by the dissimilarity function in increasing order. We introduced a second function denoted η, which represents the average growth rate of the ϕ function. η can be seen as the ratio of the difference between a given score and the first score, and the difference between their ranks. This function is defined in Equation 20:

$$\eta(\phi^{x',y}) = \frac{\phi^{x',y}(z_m) - \phi^{x',y}(z_1)}{z_m - z_1} \qquad m \in \mathcal{L} \tag{20}$$

where $\phi^{x',y}(z_m)$ is the m^{th} dissimilarity cost among the $\mathcal{L}$ scores obtained for the pair of matched pixels $(p^{x,y}, p^{x',y})$. z_m is the rank of the m^{th} score. $\eta(\phi^{x',y})$ is a discrete function that allows to highlight the large gaps between scores. It is materialized using Equation 21:

$$\xi(\phi^{x',y}) = \begin{cases} \frac{\nabla\eta^{x',y}}{m^2} & \text{if } \nabla\eta^{x',y} \geqslant 0 \\ -1 & \text{otherwise} \end{cases} \tag{21}$$

The previous function (Equation 7) is used to characterize the major scores and is applied only in the case where the gradient $\nabla\eta^{x',y}$ has a positive sign. We have introduced parameter m^2 in order to penalize the candidate pixels according to their rank. The number of candidate

pixels is given by Equation 22:

$$\tau = \arg\max_{m} \xi(\phi^{x',y}) \tag{22}$$

4.3 Hierarchical belief propagation

The inference algorithm based on a belief propagation method Felzenszwalb & Huttenlocher (2006) can be applied to reach the optimal solution that corresponds to the best disparity set. A set of messages are iteratively transmitted from a node to its neighbors until convergence. This global optimization is NP-hard and far from real time. Referring to this basic framework, all the nodes have the same weight. The main drawback is that several erroneous messages might be passed across the graph, leading moreover to an increase in the number of iterations without guarantee of reaching the best solution. Several works have tried to decrease the number of iterations of the belief propagation method. The proposed HBP technique allows an improvement in both quality of results and processing time compared with the state of the art. The main ideas of the HBP are as follows:

– The confidence measure is used to assign a weight to each node in the graph. At each iteration, messages are passed hierarchically from nodes having a high confidence measure (i.e. high weight) to nodes having a low confidence measure (i.e. small weight). A high weight means a high certainty of the message to be passed. The weights of the nodes are updated after each iteration, so that a subset of nodes is activated to be able to send messages in the next iteration.

– The propagation is first performed inside a consistent color region, and then passed to the neighboring regions. The set of regions is obtained by a color-based segmentation using the MeanShift method Comaniciu & Meer (2002).

– In our framework, the messages are passed differently from the standard BP algorithm. Instead of considering the 4-connected nodes, the k-nearest neighboring nodes are considered. These k-nearest neighboring nodes belong to a 3D support window. We assume that the labels of nodes vary smoothly within a 3D support window centered on the node to be updated.

5. Experimental results

5.1 Evaluation of the obstacle extraction module

The first proposed module is evaluated on real-world data in outdoor environments in various weather conditions. The datasets concern videosurveillance system in which a stationary camera monitors real-worlds, such as a level crossing. The four datasets include: "Pontet" and "Chamberonne", which are two level crossings in Switzerland in cloudy weather, given that test images were 384×288 pixels; a dataset entitled "Pan", which represents a level crossing in France in sunny weather, given that test images were 720×576; and a dataset taken in snowy weather in EPFL–Switzerland is also considered for the evaluation. The test images are the same size as the "Pontet" and "Chamberonne" datasets. For a qualitative evaluation purpose, 1000 of foreground ground truths images have been obtained by manual segmentation from the "Pontet" and "Chamberonne" datasets. This allows computing the recall and the precision of the detection. In the experiments, the proposed framework is compared with the Mixture Of Gaussians (MOG) and Codebook algorithms. Furthermore, ICA model is evaluated on

different color spaces and different color constancy. The obtained results are also compared with those obtained from gray scale images. The algorithms are implemented in Visual Studio C++ 2008, using the OpenCV and IT++ libraries. The four datasets are given in Figure 5.

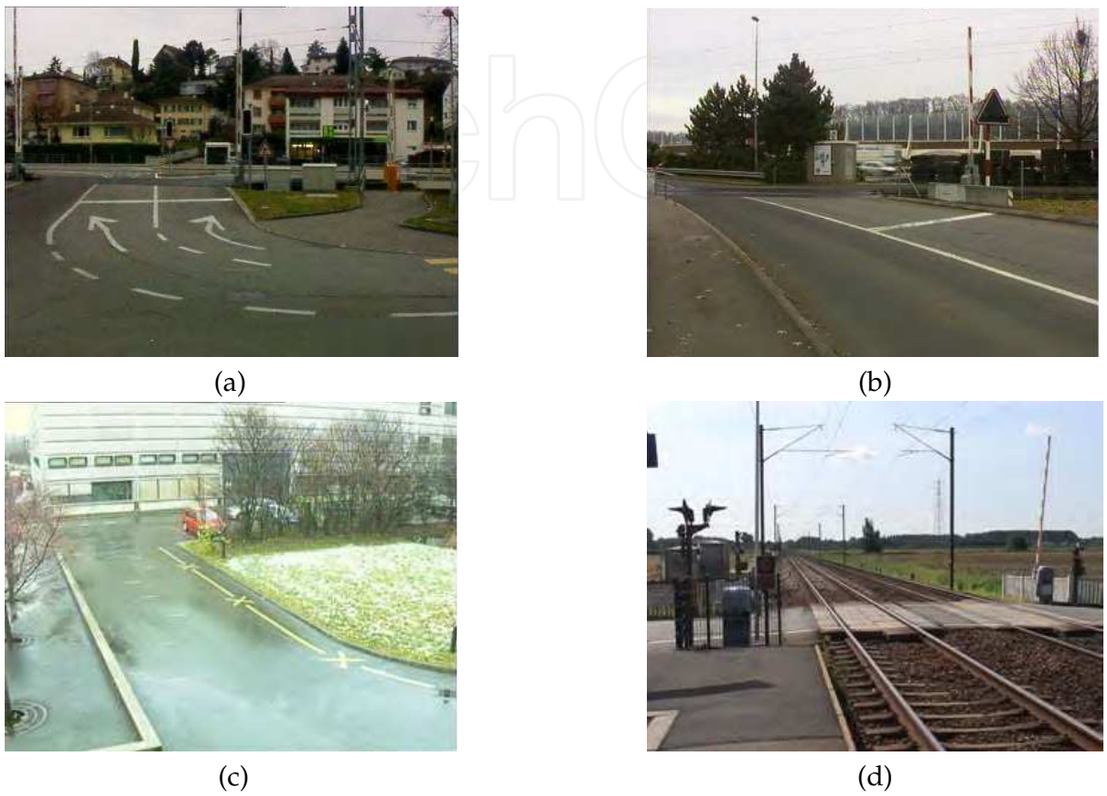


Fig. 5. The four datasets used for the evaluation (a) pontet: a level crossing in Lausanne–Switzerland (b) chamberonne: a level crossing in Lausanne–Switzerland (c) EPFL–Parking (d) pan: a level crossing in France.

5.1.1 Quantitative evaluation

The performance of the proposed scheme is compared to the two best-known techniques, which are the pixel-wise Gaussian Mixture Model (GMM) and Codebook. These algorithms are chosen because of their algorithmic complexity, which is close to ours. This qualitative comparison is performed by comparing the segmentation results of each algorithm with a ground truth. The ground truth is established by a manual segmentation of moving and stationary foreground objects in the image sequence. The evaluation criterion is expressed by the *Recall* and *Precision* parameters which describe the way the segmented image matches the corresponding ground truth. The *Recall* measures the ability of an algorithm to detect the true foreground pixels, while the *Precision* is an intrinsic criterion which gives a clue to the accuracy of the detection. These parameters can be expressed in terms of true positives T_p, true negatives T_n, false positives F_p and false negatives F_n terms. The *Recall* and *Precision* are

obtained by Equation 23 and 24:

$$Recall = \frac{T_p}{T_p + F_n} \tag{23}$$

$$Precision = \frac{T_p}{T_p + F_p} \tag{24}$$

where T_p represents the number of well pixels classified as foreground, compared to the ground truth, F_n is the number of pixels classified as background, whereas they are really foreground pixels while referring to the ground truth, and F_p is the number of pixels classified as foreground, whereas they are really background pixels. $T_p + F_n$ can be seen as the number of the true foreground pixels obtained by the ground truth, while $T_p + F_p$ is the foreground pixels classified by a given algorithm. The image samples used for computing these two previous parameters are taken from the two datasets *Pontet* and *Chamberonne*, given that five hundred images from the each dataset are used for a manual extraction of foreground objects. This allows obtaining a ground truth dataset from which the different algorithms are evaluated. Table 1 shows the qualitative evaluation of the foreground extraction process, given by *Recall* and *Precision* measures:

	MOG	**Codebook**	**ACI+Filtering**
Recall	94.76%	93.49%	96.14%
Precision	95.87%	91.72%	97.34%

Table 1. Qualitative evaluation given by *Recall* and *Precision* measures.

A visual comparison of our method compared with two other methods from the literature is given by fugure 6.

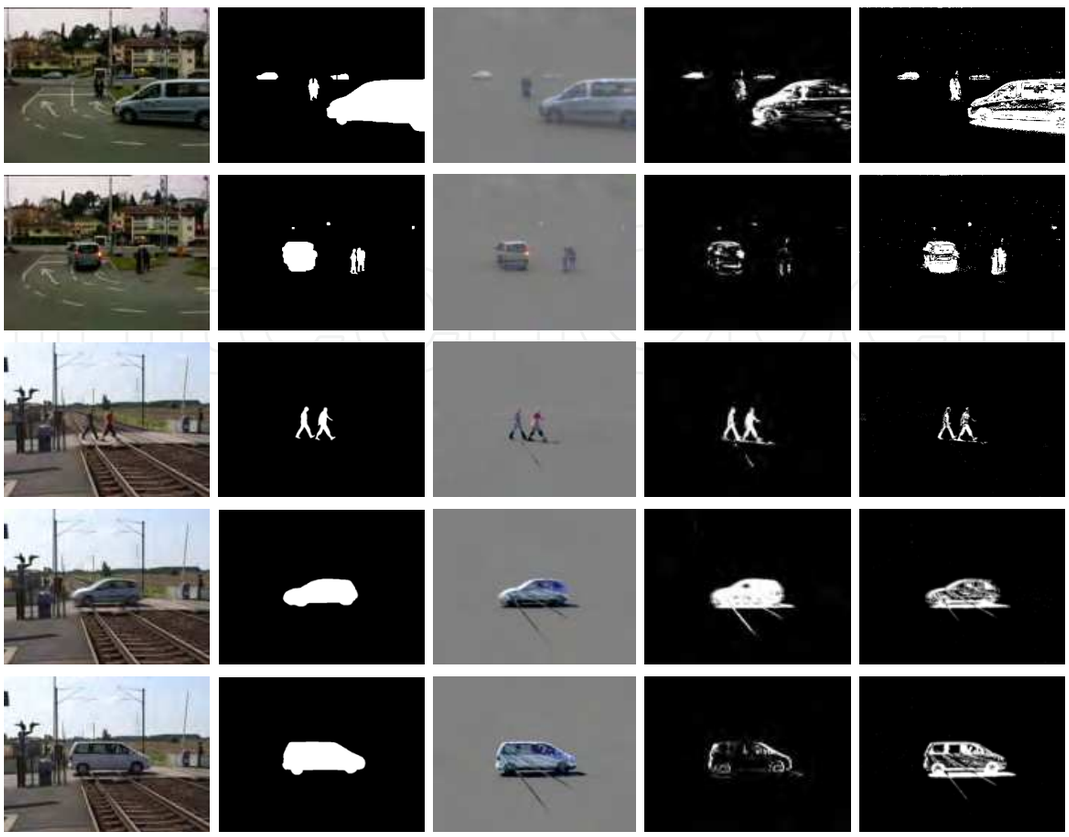
The implementation of the proposed framework runs on a personal computer with an Intel 32-bit 3.1-GHz processor. For the Pontet dataset, the proposed algorithm runs at a speed of 13 fps (frame per second). The processing time of our algorithm is compared with MOG and Codebook algorithms. Table 2 shows that our algorithm is faster than the other algorithms.

Algorithms	Proposed algorithm	MOG	Codebook
Processing time	88.687 *ms*	286.588 *ms*	118.402 *ms*

Table 2. Processing time of Pontet dataset.

5.2 Evaluation of the 3D localization module

The proposed depth estimation for 3D localization algorithm is first evaluated on the Middlebury stereo benchmark (`http://www.middlebury.edu/stereo`), using the Tsukuba, Venus, Teddy and Cones standard datasets. The evaluation concerns non occluded regions (nonocc), all regions (all) and depth-discontinuity regions (disc). In the first step of our algorithm, the WACD likelihood function is first performed on all the pixels. Applying the *winner-take-all* strategy, a label corresponding to the best estimated disparity is attributed to each pixel. The second step consists in selecting a subset of pixels according to their confidence



(a) (b) (c) (d) (e)

Fig. 6. Visual comparison of our algorithm with other methods. (a) original images (b) ground truth (c) changes detection obtained with our method (d) with MOG method and (e) with Codebook method.

measure. Indeed, the pixels having a low confidence measure belongs to either occluded or textureless regions. However, the subset corresponding to the well-matched pixels is taken as the starting point of the hierarchical belief propagation module.

Quantitatively, our method was compared to several other methods from the literature. These methods are H-Cut Miyazaki et al. (2009), max-product Felzenszwalb & Huttenlocher (2006) and PhaseBased El-Etriby et al. (2007). Table 3 provides quantitative comparison results between all four methods. This table shows the percentage of pixels incorrectly matched for the non-occluded pixels (nonocc), the discontinuity pixels (disc), and for all the matched pixels (all). More specifically, the proposed method is better for Tsukuba in "all" and "disc" pixels, in Venus for "disc" pixels and in Cones for "all" pixels.

Figure 7 illustrates an example of two objects extracted from the Cones and Teddy images, respectively. The face extracted from Cones corresponds to an non-occluded region while the teddy bear corresponds to a depth discontinuity. This proves that the propagation of disparities preserves the discontinuity between regions and gives a good accuracy in terms of matching pixels in the non-occluded regions.

Algorithm	*Tsukuba*			*Venus*			*Teddy*			*Cones*		
	nonocc	*all*	*disc*	*nonocc*	*all*	*disc*	*nonocc*	*all*	*disc*	*nonocc*	*all*	*disc*
H-Cut	2.85	4.86	14.4	1.73	3.14	20.2	10.7	19.5	25.8	5.46	15.6	15.7
Proposed	**4.87**	**5.04**	**8.47**	**3.42**	**3.99**	**10.5**	**17.5**	**20.8**	**28.0**	**7.46**	**12.5**	**13.3**
Max-Product	1.88	3.78	10.1	1.31	2.34	15.7	24.6	32.4	34.7	21.2	28.5	30.1
PhaseBased	4.26	6.53	15.4	6.71	8.16	26.4	14.5	23.1	25.5	10.8	20.5	21.2

Table 3. Algorithm evaluation on the Middlebury dataset

(a) (b) (c) (d) (e) (f)

Fig. 7. Different steps of our algorithm in different types of regions. (a) Left image (b) Segmented face and teddy bear extracted from the Cones and Teddy images, respectively, using Mean Shift (c) Dense disparity map obtained using WACD (d) Sparse disparity map corresponding to the well-matched pixels, with 60% confidence threshold (e) Dense disparity map after performing the HBP (f) Corresponding ground truth.

The disparity allows estimating the 3-D position and spatial occupancy rate of each segmented object. The transformation of an image plane point $p = \{x, y\}$ into a 3-D reference system point $P = \{X, Y, Z\}$ must be performed. The distance of an object point is calculated by triangulation, assuming parallel optical axes:

$$Z = \frac{b.f}{d} \tag{25}$$

Where
– Z is the depth, i.e. the distance between the sensor camera and the object point along the Z axis,
– f is the focal length, i.e. the distance between the lens and the sensor, supposed identical for both cameras,

– b is the baseline, i.e. the distance separating the cameras.
– d is the estimated disparity.

The proposed 3D localization algorithm is evaluated on image sequences of long of one hour acquired on two real level crossings. We have used a system composed of two cameras of model Sony DXC-390/390P 3-CCD with an optical lens of model Cinegon 3 CCD Lens 5.3mm FL. The cameras are fixed on a metal support of 1.5 meter of height, and the distance between their optical axis is fixed at 0.4 meter. The whole is placed at around 20 meters far from the dangerous zone. We illustrate in figure 8 an example of two pedestrians extracted by the stICA algorithm from the left-hand image. The image (b) is estimated by applying the WACD local stereo matching algorithm allows us to obtain a first disparity map which contains a lot of errors of matching. Much of them are identified by applying the confidence measure function. Only the pairs of matched pixels having a confidence measure higher than 60% as threshold, are kept (image c). These retained pixels are considered as a starting point for the belief propagation algorithm leading to estimate the disparity of the remaining pixels (image d). This example show the accuracy of the 3D localization in a case of occlusion, knowing that the two pedestrians are at two different distances from the cameras.

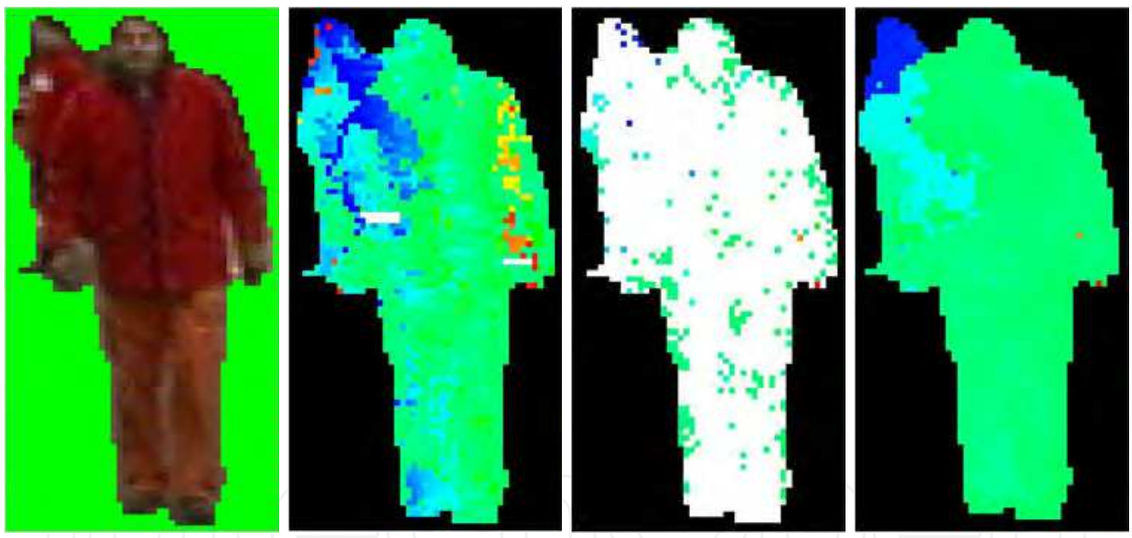


Fig. 8. 3D localization of two pedestrians partially occluded. (a) pedestrians extracted by stICA, (b) first disparity map obtained with the WACD algorithm, (c) sparse disparity map obtained after applying the confidence measure function, (d) final disparity map obtained with the selective belief propagation algorithm.

We illustrate in figure 9 some examples of obstacles which are extracted and localized with the proposed algorithms (from (a) to (d)). The first column corresponds to the left-hand images acuiqres from the left camera. The middle column represents the first disparity map obtained from the WACD stereo matching algorithm. Hence, the last column correspond to the final disparity map which will allows localizing all of obstacles in the 3D space.

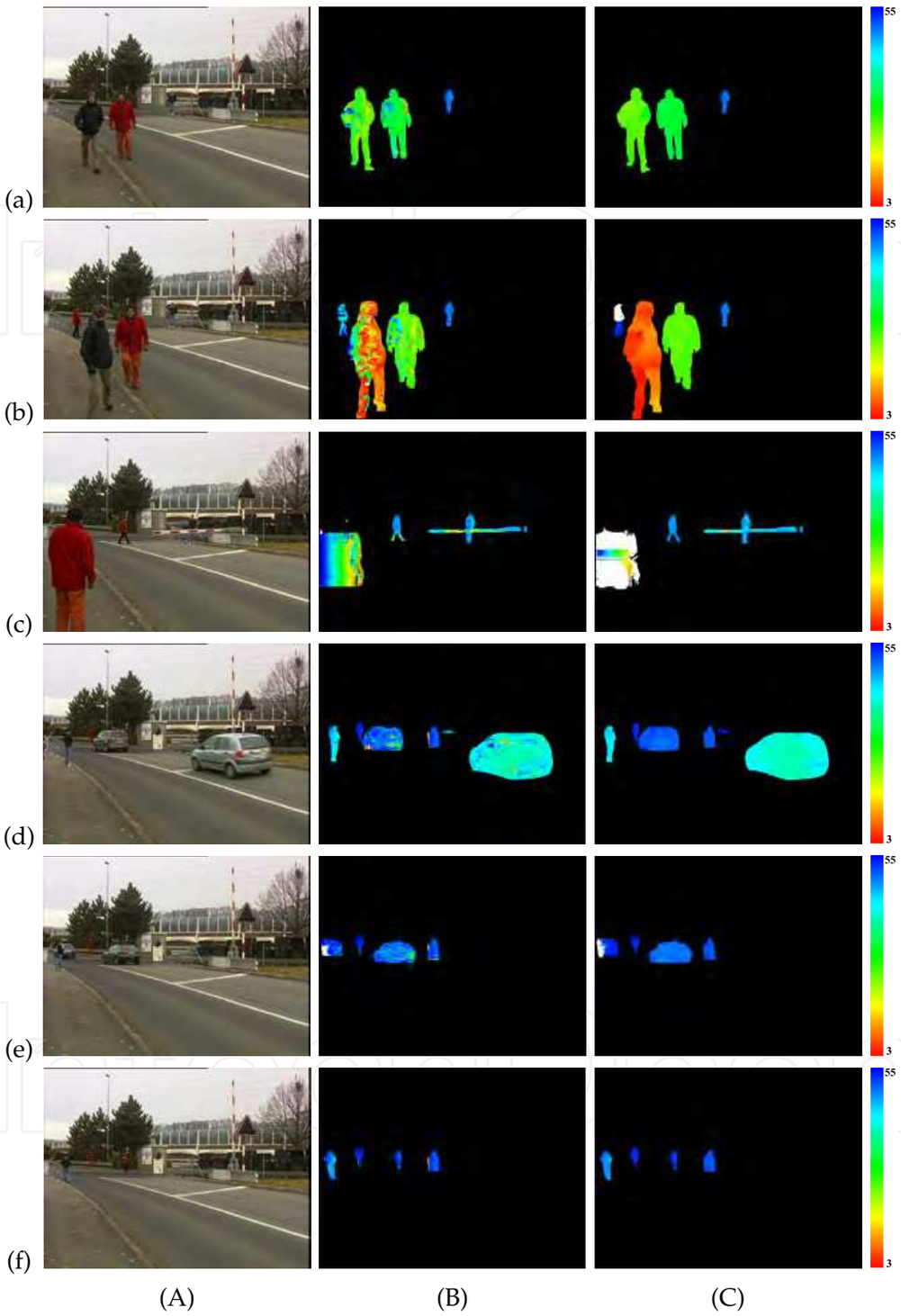


Fig. 9. 3D localization steps of a given scenario. (A) left-hand image, (B) dense disparity map obtained by applying WACD correlation function, (C) final disparity map obtained after applying the Selective Belief Propagation.

6. Conclusion

In this chapter we have proposed a processing chain addressing safety at level crossings composed of two steps : a background subtraction based on Spatio-temporal Idependentent Component Analysis and a robust 3D localization by global stereo matching algorithm. It is to be noted that the 3D localization is only applied on stationary and moving obstacles. The foreground extraction method based on stICA has already been evaluated in terms of Recall (95%) and Precision (98%), on a set of 1000 images with manually elaborated ground truth. Real-world datasets have been shot at three different level crossings and a parking at the EPFL institute including a hundred scenarios per level crossing under different illumination and weather conditions. This first step is compared with two well-known robust algorithms, entitled GMM and Codebook, from which it proves it effectiveness in term of precision of foreground extraction and processing time. The stereo matching algorithm is first applied on a standard dataset known as the Middlebury Setreo Vision which represents an unique framework for comparison with the state-of-the-art. The latter proves it effectiveness compared to stereo matching algorithms found in the literature.

The experimentations showed that the method is applicable to real-world scenes in level crossing applications. The main output of the proposed system is an accurate 3D localization of any object in, and around a level crossing. According to the experimentations, the localization of some objects may fail. However, the localization of one among sixty objects fails, this is due to the smaller number of pixels having confidence measure larger than a fixed threshold or the occlusion problem. The starting point of the belief propagation process highly depends on the number and repartition of pixels, having hight confidence measure, inside an object. This drawback can be handled by introducing the temporal dependency in the belief propagation process.

For safety purposes, the proposed system will be coupled with already existing devices at level crossings. For instance, the status of the traffic light and the barriers will be taken as input in our vision-based system. The level of such an alarm depends on the configuration of the different parameters. For instance, the presence of an obstacle in the crossing zone when the barriers are lowering is a dangerous situation and the triggered alarm must be of high importance. A Preliminary Risk Analysis (PRA) seems to be an interesting way to categorize the level of alarms. In the frame of the French project entitled PANSafer, these different parameters will be studied. In particular, telecommunication systems will be used to inform road users on the status of the level crossing. Such informations could also be shared with train driver and control room. The communication tool and the nature of information to be transmitted are in study.

7. References

Boykov, Y., Veksler, O. & Zabih, R. (2001). Fast approximate energy minimization via graph cuts, *IEEE Transactions on PAMI* 23(11): 1222–1239.

Cardoso, J.-F. (1997). Adaptive blind separation of independent sources: a deflation approach, *IEEE Letters on Signal Processing*, Vol. 4, pp. 112–114.

Comaniciu, D. & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 24(5): 603–619.

Cvejic, N., Bull, D. & Canagarajah, N. (2007). Improving fusion of surveilliance images in sensor networks using independent component analysis, *IEEE Trans. On Consumer Electronics* 53(3): 1029–1035.

Delfosse, N. & Loubaton, P. (1995). Infomax and maximum likelihood for sources separation, *IEEE Letters on Signal Processing*, Vol. 45, pp. 59–83.

Dun, B. V., Wouters, J. & Moonen, M. (2007). Improving auditory steady-state response detection using independent component analysis on multichannel eeg data, *IEEE Trans. On Biomedical Engineering* 54(7): 1220–1230.

El-Etriby, S., Al-Hamadi, A. & b. Michaelis (2007). Desnse stereo correspondance with slanted surface using phase-based algorithm, *In : IEEE International Symposium on Indistrual Electronic*.

Elgammal, A., Harwood, D. & Davis, L. (2000). Non-parametric model for background subtraction, *ECCV*.

Fakhfakh, N., Khoudour, L., El-Koursi, E., Bruyelle, J.-L., Dufaux, A., & Jacot, J. (2011). 3d objects localization using fuzzy approach and hierarchical belief propagation : application at level crossings, *In EURASIP Journal on Image and Video Processing* 2011(4): 1–15.

Fakhfakh, N., Khoudour, L., El-Koursi, E., Jacot, J. & Dufaux, A. (2010). A video-based object detection system for improving safety at level crossings, *In Open Transportation Journal* 5: 1–15.

Fakhfakh, N., Khoudour, L., El-Koursi, M., Jacot, J. & Dufaux, A. (2009). A new selective confidence measure-based approach for stereo matching, *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Springer-Verlag Berlin Heidelberg*, Vol. 5711, Santiago, Chile, pp. 184–191.

Felzenszwalb, P. & Huttenlocher, D. (2006). Efficient belief propagation for early vision, *International Journal of Computer Vision (IJCV)* 70(1): 41–54.

Foggia, P., Jolion, J., Limongiello, A. & Vento, M. (2007). Stereo vision for obstacle detection: A graph-based approach, *Lecture Notes in Computer Science, Springer Berlin / Heidelberg*, pp. 37–48.

Foresti, G. (1998). A real-time system for video surveillance of unattended outdoor environments, *IEEE Transactions on Circuits and Systems for Video Technology* 8(6): 697–704.

Griffioen, E. (2004). Improving level crossings using findings from human behaviour studies, *Proc. of 8th Inter. Level Crossing Symposium*.

Kim, K., Chalidabhongse, T., Harwood, D. & Davis, L. (2005). Real-time foregroundŰbackground segmentation using codebook model, *Journal of Real-Time Imaging, Special Issue on Video Object Processing* 11(3): 172–185.

Lee, C. & Ho, Y. (2008). Disparity estimation using belief propagation for view interpolation, *In: ITC-CSC*, Japan, pp. 21–24.

McKeown, M., Makeig, S., Brown, G., Jung, T., ndermann, S., Bell, A. & Sejnowski, T. (1998). Analysis of fmri data by blind separation into independent spatial components, *Human Brain Mappin* 6(3): 160–188.

Miyazaki, D., Matsushita, Y. & Ikeuchi, K. (2009). Interactive shadow removal from a single image using hierarchical graph cut, *In : Asian Conference on Computer Vision (ACCV)*.

Nelson, A. (2002). The uk approach to managing risk at passive level crossings, *Inter. Symposium on RailRoad-Highway Grade Crossing Research and Safety, 7th*.

Ohta, M. (2005). Level crossings obstacle detection system using stereo cameras, *Quarterly Report of RTRI* 46(2): 110–117.

Oja, E., Ogawa, H. & Wangviwattana, J. (1991). Learning in nonlinear constrained hebbian networks, *T.Kohonen, et al .,editor, Artificial Neural Networks,Proc.ICANN*, Espoo, Finland, Amsterdam, Holland, pp. 385–390.

Stauffer, C. & Grimson, W. (2000). Learning patterns of activity using real-time tracking, *IEEE Trans. Pattern Anal. Mach. Intell.* 22(8): 747–757.

Taguchi, Y., Wilburn, B. & Zitnick, C. L. (2008). Stereo reconstruction with mixed pixels using adaptive over-segmentation, *CVPR*, Anchorage, Alaska, pp. 1–8.

Trinh, H. (2008). Efficient stereo algorithm using multiscale belief propagation on segmented images, *Proceedings of the Brith Machine Vision Conference (BMVC)*.

Tsai, D. & Lai, S. (2009). Independent component analysis-based background subtraction for indoor surveillance, *IEEE Trans. On Image Processing* 18(1): 158–167.

Waldert, S. (2007). Real-time fetal heart monitoring in biomagnetic measurements using aadaptive real-time ica, *IEEE Trans. On Biomedical Engineering* 54(107): 1964–1874.

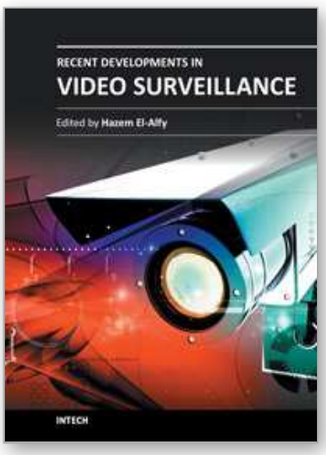
Wu, F. (1982). The potts model, *Reviews of Modern Physics* 54(1): 235–268.

Yang, Q., Wang, L., Yang, R., Stewenius, H. & Niste, D. (2009). Stereo matching with color-weighted correlation, hierachical belief propagation, *IEEE Trans. on PAMI* 31(3): 492–504.

Yoda, I., Sakaue, K. & Hosotani, D. (2006). Multi-point stereo camera system for controlling safety at railroad crossings, *IEEE ICVS*.

Zhang, X. & Chen, Z. (2006). An automated video object extraction system based on spatiotemporal independent component analysis and multiscale segmentation, *EURASIP Journal on Applied Signal Processing* 2006(2): 1–22.

Zhen, T. & Zhenjiang, M. (2008). Fast background subtraction using improved gmm and graph cut, *Congress on Image and Signal Processing, CISP*, Vol. 4, Sanya, China, pp. 181–185.



Recent Developments in Video Surveillance

Edited by Dr. Hazem El-Alfy

ISBN 978-953-51-0468-1
Hard cover, 122 pages
Publisher InTech
Published online 04, April, 2012
Published in print edition April, 2012

With surveillance cameras installed everywhere and continuously streaming thousands of hours of video, how can that huge amount of data be analyzed or even be useful? Is it possible to search those countless hours of videos for subjects or events of interest? Shouldn't the presence of a car stopped at a railroad crossing trigger an alarm system to prevent a potential accident? In the chapters selected for this book, experts in video surveillance provide answers to these questions and other interesting problems, skillfully blending research experience with practical real life applications. Academic researchers will find a reliable compilation of relevant literature in addition to pointers to current advances in the field. Industry practitioners will find useful hints about state-of-the-art applications. The book also provides directions for open problems where further advances can be pursued.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen