HIDDEN MARKOV RANDOM FIELD APPLICATION TO RAILWAY INFRASTRUCTURE DIAGNOSIS

E. Côme [*] **L. Bouillaut** [*] **L. Oukhellou** [*,**]
P. Aknin [*]

[*] *French National Institute for Transport and Safety Research, 2 av. Malleret-Joinville, 94114 Arcueil Cedex, France*
[**] *University of Paris XII - CERTES, 61 av du Général de Gaulle, 94100 Créteil, France*

Abstract: Hidden Markov Random Fields (HMRF) are widely used in solving various problems. Image segmentation is an example of such HMRF success. This paper presents a post-processing tool based on such a model and designed to increase the relevancy of a diagnosis system for rail defects detection. In this application, the hidden Markov field is not only used to define a spatial smoothness prior as it is often done in image segmentation, but , it is used to learn the spatial interaction between track singular points, and so the track label patterns. For this, an approach based on a semi-parametric model is presented.

Keywords: Railways, Diagnosis, Patterns, Markov Field,

1. INTRODUCTION AND APPLICATIVE CONTEXT

Infrastructure maintenance is an important field for railway operators. In particular, rail integrity is critical for train control as well as operating quality and availability. To assist in both these areas, an eddy current sensor has been specifically developed to detect any variations in the electromagnetic characteristics of the rail (Oukhellou *et al.* 1999).

Thus, real defects (such as broken rails or shelling) are detected, as are the singular points such as the welded joints (Wj), fishplated joints (Fj) and switch joints (Sj) that join together rail lengths. The distinction between real defects and singular points is obviously an essential task that must be made by the decision system. The first implementation of the diagnosis system was successfully tested in July 2002 on the Paris metro network (Bentoumi *et al.* 2003). The classifier, based on the sensor data analysis, detected all the major defects, such as split rails, correctly. On the other hand, the minor defect (shelling) detection rate of this local classifier was estimated at about 72% (Oukhellou and Aknin 1999). This is due to the weak signature of such kind of defect in term of energy level. They are mistaken with other classes, particularly with singular points like Welded joints (Wj).

The aim of this paper is the integration of spatial information in this diagnosis in order to increase its reliability. Indeed, the setting rules of the track and the technical constraints for maintenance actions, induce a strong structuring of the track singular point positions. In fact, some spatial patterns are very frequent and might be identified (cf Fig. 1). For example, fishplatted joints, which are used for electrical isolation of track portions,

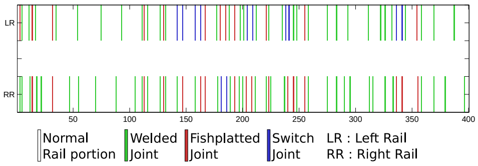

Fig. 1. Image of track label on a 400m portion

are manufactured with two 3 meters rail lengths on both sides and set on the track with welds. The following pattern is therefore very common :

$$Wj \overset{3m}{\longleftrightarrow} Fj \overset{3m}{\longleftrightarrow} Wj$$
$$Wj \overset{3m}{\longleftrightarrow} Fj \overset{3m}{\longleftrightarrow} Wj$$

Such patterns brought useful information for singular point recognition. A model taking them into account would therefore obtain better results. In other word, the track could be compared to a "texture" with stationary statistics. This texture resumes the singular point patterns and its use might be interesting to increase the reliability of the diagnose. Thus, our study aims to built a model using two information sources :

- sensor data
- track label pattern

This article presents a solution based on HMRF to model such a system. Indeed, HMRF are interesting for two main reasons : On the first hand, track label texture might be described by the local conditional probability distributions of the hidden field. On the second hand, sensor data might be integrated in the classification process through the observable field. HMRF needs a discretized lattice. Taking into account the spatial sampling rate used by the track maintenance team, the positions of track singular points will be discrete variables with 1 meter precision. In fact, all the tracks will be segmented into 1 meter length sections.

The two next sections introduce theoretical aspects of HMRF and present the Semi-Parametric model adopted for learning the track texture. Section 3 deals with the chosen experimental settings for testing our solution and section 4 focuses on the results of our post processing tool. Finally, these results are discussed and some concluding remarks are given on the interest of such an approach for the integration of structural information into the diagnostic task.

2. HIDDEN MARKOV FIELDS

Let be $X = \{X_s, \quad s \in S = \{s_1, \dots s_M\}\}$ and $Y = \{Y_s), \quad s \in S\}$ be two random fields defined on the lattice S. Each site of X : X_s takes its value in a frame Ω of possible labels, and each Y_s takes its value in $\mathbb{R}^n$. We proposed to use sensor data as the observable part Y of the HMRF defined by $Z = \{X, Y\}$, and to express the spatial interactions between rail points with the local conditional probability distributions of X. In our case, Ω represents the possible labels of one meter long track portions. For a complete description of Markov Field and Hidden Markov models see (Besag 1986, Geman and Geman 1984, Rabiner and Juang 1986).

2.1 Hypothesis

The classical problem with HMRF is to find $\hat{x}$ the more likely realization of X, knowing $Y = y$. Various models are possible with such an approach, depending on hypothesis that are made on the relations between the two fields. We have adopted the Hidden Markov Random Field with independent noise model. Then, the following hypothesis are made :

$$p(x_s|x_r, r \neq s) = p(x_s|x_r, r \in \mathcal{N}_s), \; \forall s \in S \quad (1)$$
$$p(\mathbf{y}|x) = \prod_{s \in S} p(\mathbf{y}_s|x) \quad (2)$$
$$p(\mathbf{y}_s|x) = p(\mathbf{y}_s|x_s), \; \forall s \in S \quad (3)$$

where $\mathcal{N}_s$ represent the neighborhood of x_s

This model is named "with independent noise" because the random variables Y_s are independent conditionally on X. If X is a Markov field, the marginal field $X|Y = y$ is also a Markov field. With such hypothesis the expression of the local conditional probability distributions that one site X_s of $X|Y = y$ belong to one element of the frame Ω is equal to :

$$p(x_s|\mathbf{y}_t, t \in S, x_r, r \neq s) = p(x_s|\mathbf{y}_s, x_r, r \in \mathcal{N}_s)$$

and if the Bayes rule is applied twice, we obtain :

$$= \frac{p(x_r, r \in \mathcal{N}_s|x_s).p(\mathbf{y}_s|x_s).p(x_s)}{p(\mathbf{y}_s, x_r, r \in \mathcal{N}_s)}$$
$$= \frac{p(x_s|x_r, r \in \mathcal{N}_s).p(\mathbf{y}_s|x_s).p(x_r, r \in \mathcal{N}_s)}{p(\mathbf{y}_s, x_r, r \in \mathcal{N}_s)}$$
$$= \frac{1}{K} p(x_s|x_r, r \in \mathcal{N}_s).p(\mathbf{y}_s|x_s) \quad (4)$$

where K is a normalizing constant which does not depend on the value of x_s.

So, the local posterior distribution over labels of the field $X|Y = y$ is proportional to the product of the X field local conditional probability distributions $p(x_s|x_r, r \in \mathcal{N}_s)$ by the observation likelihood $p(\mathbf{y}_s|x_s)$. These different parts must be defined. They describe the local behavior of our random field, but Besag (Besag 1974) proved that the joint distribution over all the site of a random field is uniquely determined by its local conditional probability distributions. This local conditional probabilities are therefore enough to specify all the model. The Hammersley-Clifford theorem specifies the condition under which the local conditional probability distributions define a valid joint distribution. In that case, the Markov field is equivalent to a Gibbs field and the joint distribution may be defined by :

$$p(x) = \frac{1}{Z} \exp\left(\sum_{c \in \mathcal{C}} -V_c(x)\right) \tag{5}$$

where Z is the partition function assuring normality, $\mathcal{C}$ is the set of all cliques associated with the neighborhood $\mathcal{N}$ and V_c are the associated clique potentials.

The local conditional probability distributions are therefore :

$$p(x_s \mid x_r, r \in \mathcal{N}_s) = \tag{6}$$
$$= \frac{\exp\left(\sum_{c \in \mathcal{C}_s} -V_c(x)\right)}{\sum_{x_s \in \Omega} \exp\left(\sum_{c \in \mathcal{C}_s} -V_c(x)\right)}$$

where $\mathcal{C}_s$ is the set of all cliques containing x_s.

When the local parameters of such model have been learned, an algorithm such as Iterative Conditional Mode (ICM) may be used to produce an interesting estimate $\hat{x} = \{(\hat{x}_s), \quad s \in S\}$, such as the joint probability of the classification $\hat{x}$, knowing the observable process $\mathbf{y}$, is near maximal. Such classification process takes into account the two information sources available here : the spatial interactions between the sites of X are expressed with the help of the local conditional probability distributions. Sensor data are integrated with the help of the observable field Y. The next section will introduce a semi-parametric approach to learn the hidden field local conditional probability distributions and the conditional probability density functions of the observations.

3. A SEMI-PARAMETRIC APPROACH

As already said, the hidden Markov field is not only used, to define a spatial smoothness prior, as it is often done in image segmentation, with the Ising or the Potts model (Wu 1982). In our application, the Markov field is used to learn the repeated singular point patterns of the track. Our approach to learn the local conditional probability distributions of X is thus closer to those found in Texture Synthesis or Recognition (Paget and Longstaff 1984) and is nonparametric. Then, the vector of parameters $(\mathbf{\Psi}, \mathbf{\Phi})$ describing the model has thus two different parts :

- $\mathbf{\Psi}$ (nonparametric part) : contains all the local conditional probability distributions : $p(X_s = \omega | (x_r, r \in \mathcal{N}_s) = \theta)$, $\forall \theta \in \Theta, \quad \forall \omega \in \Omega$, where Θ is the ensemble of possible neighborhood configurations
- $\mathbf{\Phi}$ (parametric part) : contains all the parameters that describe the conditional densities of the observable field knowing each class : $p(\mathbf{y}_s | x_s = \omega) = f(\mathbf{y}_s; \boldsymbol{\lambda}_\omega), \forall \omega \in \Omega$

The local conditional probability distributions of a Markov field can be learned with a nonparametric approach. In that case the probability of one label knowing the neighbors label is simply estimated by its frequency in a homogeneous sample of the random field, as expressed in Eq. 7

$$p(X_s = \omega \mid (x_r, r \in \mathcal{N}_s) = \theta) =$$
$$= \frac{\sum_{s=1}^{M} \delta(x_s, \omega).\delta((x_r, r \in \mathcal{N}_s), \theta)}{\sum_{s=1}^{M} \delta((x_r, r \in \mathcal{N}_s), \theta)} \tag{7}$$
$$, \quad \forall \omega \in \Omega, \quad \forall \theta \in \Theta$$

where δ is the Kronecker function.

When the sample data is sparsely dispersed over the configuration space, this estimation tends to be more reliable than parametric estimation. However, with nonparametric estimation, the local conditional probability distributions may not define a valid joint distribution. Nevertheless, in practice, such an estimation approach has already proved is efficiency in texture recognition (Paget and Longstaff 1984).

To use this scheme some realization of the Hidden Markov field must be available during the training phase. The RATP company (the Paris metro network operator), has recorded in a specific database named SIAM, the positions and natures of all joints of its lines. So, in our case, some realizations of the hidden field were available for the training.

An application particularity must be explained to define Ω the label space of X . The final frame, that we are interested in, is :

$$\Omega^{'} = \{N, Wj, Fj, Sj, D\}$$

where D is the label for defects and N is the label for rail portions without defects nor singular points.

However, track setting rules and maintenance actions constrain the positions of singular points, but they do not constrain the position of minor rail defects. So the label space Ω of X must be modified to take into account this fact :

$$\Omega = \{O, Fj, Wj, Sj\}$$

where O stands for Other.

Ω is related to Ω', by the following relations:

$$\begin{aligned} \Omega \,//\, \Omega' \\ \{O\} \rightarrow \{D \cup ND\} \\ \{Wj\} \rightarrow \{Wj\} \\ \{Fj\} \rightarrow \{Fj\} \\ \{Sj\} \rightarrow \{Sj\} \end{aligned} \tag{8}$$

Using SIAM database, an image of each track was built. Each track meter was labeled by a value in Ω. Fig. 1 show an extract of such a picture. All of these images were considered as various realizations of the same Markov Field X and the local conditional probabilities $p(X_s = \omega | (x_r, r \in \mathcal{N}_s) = \theta)$ were easily estimated, with their frequencies in these samples.

To learn the parameters $\mathbf{\Phi}$, describing the conditional distribution of Y, a parametric model was chosen and the parameters were learned with a maximum likelihood approach. Normal conditional distributions and their mixture were postulated to deal with the differences between, the frame of interest for classification Ω' and the frame of X (Ω). Using 8, the distribution of y_s knowing $X_s = O$ was defined as a mixture of normal distributions:

$$\begin{aligned} p(y_s | x_s = O) = & \pi_{D|O} \phi(y_s; \boldsymbol{\mu}_D, \Sigma_D) \\ & + \pi_{N|O} \phi(y_s, \boldsymbol{\mu}_N, \Sigma_N) \end{aligned} \tag{9}$$

where $\phi(.; \boldsymbol{\mu}, \sigma)$ is the multidimensional normal probability density function with parameters $(\boldsymbol{\mu}, \Sigma)$ and $\pi_{D|O} = p(X_s = D | X_s = O)$, $\pi_{N|O} = p(X_s = N | X_s = O)$.

The last things that must be adapted to fit the model to our application specificities is the neighborhood system. The proposed solution is introduced in the next subsection.

3.1 Neighborhood System

Our application context is quite different from the classical image classification problem. Indeed, we always have to deal with "images" of size $M = 2 * N$, one row for each rail of the track. In our case N is near 10 000. To define a valid joint distribution, a neighborhood system must respect the following criterion:

$$s \in \mathcal{N}_r \Leftrightarrow r \in \mathcal{N}_s \tag{10}$$

So, for a neighborhood $\mathcal{N}^k$, two groups of sites were considered. The first one corresponds to a $2 * k + 1$ meters windows centered on the current site which covers the current rail. The second one uses the same window but on the opposite rail. Such symmetric neighborhood is justified by the track laying rules and maintenance policies which introduce many symmetries between the two rail. Different window length $k = 1 \ldots 6$ have been tested to choose the best solution.

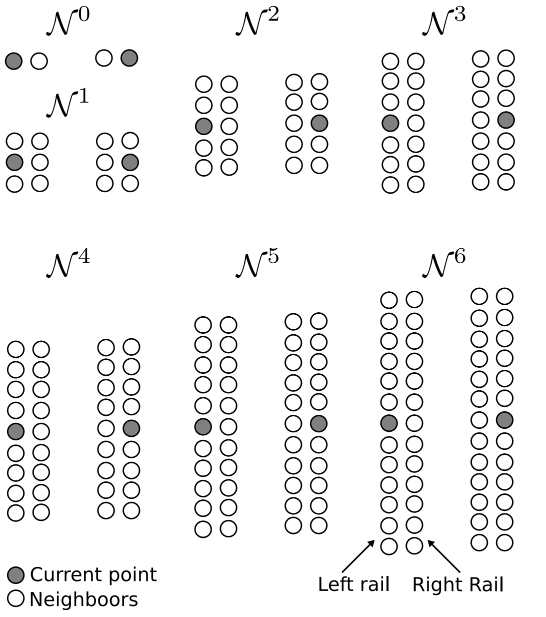

Fig. 2. Diagram of the Neighborhood systems

The last neighborhoods are quite large, compared to classical image applications. But, due to the strong structure of the track, only a small part of the possible configurations are met on our data. In fact during the training we have found 4 964 different configurations for $\mathcal{N}^5$ and 7 879 for $\mathcal{N}^6$, which must be compared to 4^{21} and 4^{25} possible configurations. So, as it was assumed, the sample data is sparsely dispersed over the configuration space. In the next section the ICM algorithm used for the classification task is briefly presented.

3.2 Classification with ICM

The ICM algorithm was first proposed by Besag (Besag 1986), as an iterative approximation of the joint distribution MAP estimator. The algorithm begins with an initialization $\hat{x}^0$, and until convergence, a site is randomly chosen and the most likely label for this site, knowing the current classification and the observable process y, is assigned to it. The following pseudo code resumes this algorithm:

```
initialization
$x_s^0 = rand(\Omega), \quad \forall s \in S$

until convergence
  $s = random(S)$
$\hat{x}_s^{k+1} = arg \max_{\omega \in \Omega} \left( p(\omega | x_r^k, r \in \mathcal{N}_s).p(y_s|\omega) \right)$
```

This algorithm converges to a local maximum of the Gibbs conditional distribution. The convergence of the algorithm to a local maximum is detected when, between two consecutive scans of all sites, the number of changes is low. This algorithm was used because a simulated annealing is more time consuming. Furthermore, the texture learned by the hidden random field is strongly structured, so a simple ICM is interesting to reach a good solution in few time. During our experiments, we observed that the convergence is reached in less than 10 scans. This fact confirmed that the solution found with the ICM algorithm is quite good.

To produce a final decision on the frame Ω' a final processing must be done. When ICM has converged to a good solution, a last scan is made and all sites with the label O are classified in N or D with the help of sensor data only. The decision is made according to the posterior probabilities of the two hypothesis, from Eq. 9 :

$$\omega = \arg \max_{\omega \in \{N,D\}} \left(p(\omega|\mathbf{y}, O) \right) \tag{11}$$

$$p\ (\ x_s = N|\mathbf{y}, x_s = O) = \tag{12}$$

$$= \frac{\pi_{N|O} \phi(y_s; \boldsymbol{\mu}_N, \Sigma_N)}{\pi_{N|O} \phi(y_s; \boldsymbol{\mu}_N, \Sigma_N) + \pi_{D|O} \phi(y_s; \boldsymbol{\mu}_D, \Sigma_D)}$$

$$p\ (\ x_s = D|\mathbf{y}, x_s = O) = \tag{13}$$

$$= \frac{\pi_{D|O} \phi(y_s; \boldsymbol{\mu}_D, \Sigma_D)}{\pi_{N|O} \phi(y_s; \boldsymbol{\mu}_N, \Sigma_N) + \pi_{D|O} \phi(y_s; \boldsymbol{\mu}_D, \Sigma_D)}$$

4. EXPERIMENTAL SETTINGS

To test our solution, we have simulated sensor data and used a specific part of the SIAM database as realizations of the hidden field. But, the SIAM database contains any information on minor defects position. So, to introduce defects some normal rail locations were randomly converted to defects, with a probability p. The model of sensor data is a Gaussian mixture composed of five modes. Fig. 3 presents the four principal modes. The last one $\{N\}$ is far away rejected and not presented on Fig. 3.

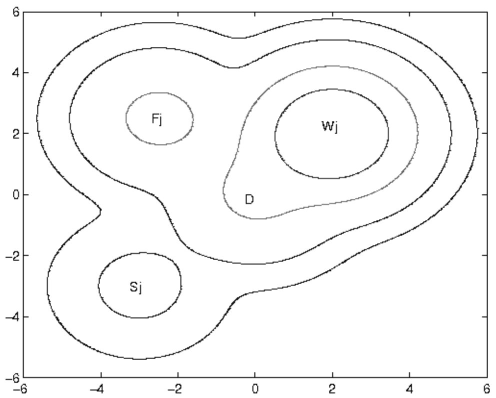

Fig. 3. Diagram of the virtual sensor data

In order to reproduce the real experiment context, we have chosen the parameters of the normal distributions to obtain confusion levels close to those found during the first experimental test of the sensor (Bentoumi et al. 2003). The performances of the optimal Bayes classifier on this simulated sensor data are the following :

ND	WJ	FJ	SJ	D
100	98.33	94.24	94.50	70.99

Table 1. Correct detection rate of the optimal Bayes classifier with simulated data

5. RESULTS

Four metro tracks (92 132 observations) have been used for HMRF training and one metro track (16 062 observations) for the evaluation of model's accuracy. In order to obtain better performances evaluation, the random process which inserts minor defects and generates the sensor data, has been repeated one hundred times. Tab. 2 presents the results of HMRF models with neighborhood $\mathcal{N}^i, i \in \{0 \ldots 6\}$, averaged over these hundred runs.

We may observe different facts in these results. First, even using very few information, HMRF with neighborhood $\mathcal{N}^0$ gives already interesting results and outperforms the optimal Bayes classifier with only sensor data. This may be explained by the fact that the left and right rails of a track are very strongly correlated. Thus, the information from the opposite rail is very informative. We may also notice, that it's possible, for all the

	ND	Wj	Fj	Sj	D
$\mathcal{N}^0$	100	98.42	**97.81**	98.45	79.19
$\mathcal{N}^1$	100	**98.54**	97.63	98.10	82.13
$\mathcal{N}^2$	100	98.51	95.18	98.10	83.46
$\mathcal{N}^3$	100	98.46	95.25	95.65	84.32
$\mathcal{N}^4$	100	98.31	94.46	95.45	84.59
$\mathcal{N}^5$	100	97.70	93.34	99.20	**84.67**
$\mathcal{N}^6$	100	97.38	93.59	**99.35**	84.35

Table 2. Correct detection rate of the HMRF

classes, to find a neighborhood which outperforms the optimal Bayes classifier using only sensor data.

An increase of the neighborhood size must induce an improvement on the detection rates, except if the training set size is too small. If we look closer at the correct detection rate for the D class (cf. Fig 4), a neighborhood size of 5 seems to be a good trade off between complexity and good generalization properties.

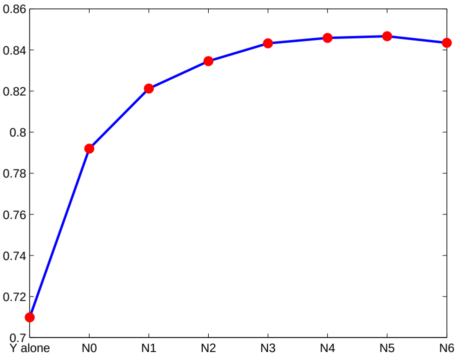

Fig. 4. D Correct detection rate of the different Neighborhood System

Finally, these results are very encouraging for our application. Indeed, as hoped, the non structural point like D, misclassified as joints without our post processing, are now well recognized as non-singular points. And real joints are still recognized. Our post processing is therefore relevant to distinguish between elements of the track structure (singular points) and real defects. Logically, the correct recognition rate improvement for the minor defect is the most important, with an increasing of more than 13%.

6. CONCLUSIONS

This study dealt with the integration of the track label texture for a better railways minor defects classification. Track label texture learning appeared to be useful to improve a classical diagnosis system based only on sensor data. Moreover, HMRF have proved their relevancy to learn, with a labeled database, such texture. Furthermore, the choice of a simple classification algorithm (ICM) seemed to be acceptable with respect to computational load, computational time and results accuracy. Finally, the semi-parametric approach presented here for parameter learning has also proved is efficiency.

7. AKNOWLEDGEMENT

The authors would like to thank the RATP Infrastructure Dpt for providing SIAM data.

REFERENCES

Bentoumi, M., P. Aknin and G. Bloch (2003). Online defect diagnosis with eddy current probes and specific detection processings. European Journal of Applied Physics **23**(3), 227–233.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society B **35**, 192–236.

Besag, J. (1986). On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society B **48**, 259–302.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distribution, and the bayesian restoration of images. IEEE Transaction on Pattern Analysis and Machine Intelligence **6**, 721–741.

Oukhellou, L. and P. Aknin (1999). Hybrid training of radial basis function networks in a partitionning context of classification. Neurocomputing **28**, 165–175.

Oukhellou, L., P. Aknin and J.P. Perrin (1999). Dedicated sensor and classifier of rail head defects for railway systems. Control Engineering Practice **7**, 57–61.

Paget, R. and I.D. Longstaff (1984). Texture synthesis and unsupervised recognition with nonparametric multiscale markov random field models. IEEE Transaction on Pattern Analysis and Machine Intelligence **6**, 721–741.

Rabiner, L.R. and B.H. Juang (1986). An introduction to hidden Markov models. IEEE ASSP Magazine pp. 4–15.

Wu, F.Y. (1982). The potts model. Review of modern physics **54**(1), 235–268.