Machines à vecteur de support 1-classe (OC-SVM) pour la détection non supervisée d'événements sonores anormaux

Sébastien LECOMTE[1,2], Cédric RICHARD[3], Régis LENGELLÉ[2], François CAPMAN[1], Bertrand RAVERA[1]

[1]Laboratoire Multi-Media Processing, Thales Communications, Colombes.

[2]Institut Charles Delaunay - LM2S, UMR STMR, Université de Technologie de Troyes.

[3]Institut Univ. de France, Laboratoire Fizeau, UMR, CNRS 6525, Observatoire de la Côte d'Azur, Univ. Nice Sophia-Antipolis.

sebastien.lecomte@utt.fr, cedric.richard@unice.fr lengelle@utt.fr
francois.capman@fr.thalesgroup.com, bertrand.ravera@fr.thalesgroup.com

Résumé – Cet article introduit une méthode non-supervisée pour la détection en temps réel d'événements sonores anormaux dans le contexte de la surveillance audio. En se basant sur les machines à vecteur de support 1-classe (OC-SVM) pour modéliser une distribution, celle d'une ambiance sonore, nous introduisons une hypothèse complémentaire et proposons de construire des ensembles de familles de règles de décision. On montre qu'il est alors possible de contrôler le compromis entre fausse-alarme et non-détection sans modifier le OC-SVM qui capture le mieux l'ambiance. Nous présentons ensuite une technique adaptative en ligne pour intégrer temporellement la statistique de décision afin d'améliorer les performances et la robustesse de notre approche. Nous introduisons également un système de génération de bases de données dédiées à l'évaluation des systèmes de surveillance. Finalement, nous présentons les résultats de détection obtenus.

Abstract – This paper introduces an unsupervised method for real time detection of abnormal audio events in the context of audio surveillance. Based on One-Class Support Vector Machine (OC-SVM) to model the distribution of an audio ambience, we introduce a complementary hypothesis et propose to build sets of families of decision rules. We show it is then possible to control the trade-off between false-alarm and miss probabilities without modifying the OC-SVM that best captures the ambience. Then we present an adaptive online scheme for temporal integration of the decision statistics in order to improve performance and robustness of our approach. We also introduce a framework to generate databases dedicated to the evaluation of audio surveillance systems. Finally, we present the results obtained on the database.

1 Introduction

Dans un contexte de sécurité publique (transports, milieux urbains, etc.), les systèmes de surveillance dits de troisième génération [13] se basent sur une analyse multimodale, qui inclut l'utilisation de l'audio. L'approche que nous considérons pour la modalité audio consiste à détecter puis identifier (classer) des situations anormales (déviation de l'ambiance, dite « normale », apprise par le système). On s'intéresse dans cet article à l'amélioration de la phase de détection.

La plupart des systèmes de détection audio sont supervisés. Ils sont dédiés à un ensemble d'événements identifiés [17], ou fortement corrélés à une ambiance stationnaire [10]. Ces approches ne sont pas appropriées à notre contexte car nous n'avons pas d'information sur les événements anormaux ; et les ambiances considérées sont des continuums non-stationnaires incluant des événements non anormaux. De plus, [18] considère que construire des systèmes de surveillance automatisés plus efficaces et plus intelligents passe par la nécessité d'un minimum d'informations. Nous étudions donc une approche non supervisée pour modéliser la distribution de la normalité, sans *a priori* sur les événements ou l'ambiance.

La littérature propose de nombreuses approches pour estimer la distribution d'un ensemble d'apprentissage ; mélanges de modèles Gaussiens (GMM) et machines à vecteur de support 1-classe (OC-SVM, [12, 14, 16]) sont les plus populaires. Par la nature du critère optimisé, les OC-SVM présentent de bons résultats en généralisation. Ils sont de plus capables de modéliser des ensembles de formes arbitraires. Nous présentons une modification de la formulation OC-SVM en introduisant une hypothèse complémentaire afin de construire des ensembles de familles de règles de décision. Ceci va nous permettre de contrôler le compromis entre probabilités de non-détection et de fausse-alarme, sans avoir à réaliser un nouvel apprentissage.

Nous nous intéressons également à l'intégration temporelle la statistique de décision. Comme les événements que nous souhaitons détecter sont de taille variable, nous proposons d'utiliser des informations de segmentation du signal audio afin de procéder à l'intégration. A l'instar de [3], la plupart des algorithmes de segmentation proposés dans la littérature reposent sur un critère d'information. Cependant, il est difficile de mettre en oeuvre ce type de système avec de bonnes capacités en généralisation lorsque nous considérons des types de signaux hétérogènes. Nous présentons donc une approche par segmentation en ligne par mesure de similarité entre trames comme cela a été suggéré en reconnaissance de la parole [5, 6].

Dans cet article, nous rappelons la théorie liée aux OC-SVM puis présentons la construction d'ensembles de familles de règles de décisions basée sur cette approche. Nous traitons ensuite de l'intégration temporelle de la statistique de décision. Nous introduisons également une approche originale pour générer des signaux audio pour la détection d'événements anormaux. Nous exposons enfin nos résultats avant de conclure.

2 Familles de détecteurs OC-SVM

Soit $\{\mathbf{x}_1 \dots \mathbf{x}_l\}$, $\mathbf{x}_i \in X \in \mathbb{R}^d$ un ensemble majoritairement issu d'une classe unique C_0 (la normalité), où $l \in \mathbb{N}$ est le nombre d'observations. Les OC-SVM [14] ont pour objectif de définir la frontière de Γ, la région de volume minimum qui englobe les $(1-\nu)l$ observations, appartenant à C_0. L'hyperparamètre $\nu \in [0;1]$, contrôle la portion des observations qui est autorisée à être en dehors de Γ (points aberrants ou *outliers*). Soit $f_X : \mathbb{R}^d \to \mathbb{R}$ une fonction de décision telle que :

$$\begin{cases} f_X(\mathbf{x}) \geq 0 & \text{si } \mathbf{x} \in \Gamma \\ f_X(\mathbf{x}) < 0 & \text{sinon} \end{cases} \quad (1)$$

Dans le contexte des SVM, l'espace des fonctions $f_X(x)$ possibles est réduit à un espace de Hilbert à noyau reproduisant (RKHS) muni d'un noyau $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Celui-ci induit l'espace transformé H (ou « espace des *features* ») via la projection $\phi : \mathbb{R}^d \to H$. Soit $\langle \cdot, \cdot \rangle_H$ le produit scalaire dans H. On considère ici le noyau Gaussien :

$$\kappa(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_H = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2} \right)$$

Remarque : $\kappa(x, x) = 1$, les données sont donc projetées sur une hypersphère de rayon unitaire centrée à l'origine de H.

La phase d'apprentissage d'un OC-SVM consiste à définir l'hyperplan séparateur $W = \{\mathbf{h} \in H : \langle \mathbf{h}, \mathbf{w} \rangle_H - b = 0\}$ dans H tel que la marge $b/\|\mathbf{w}\|_H$ soit maximum (cf. figure 1). Les paramètres $\mathbf{w}$ et b sont solutions du problème d'optimisation suivant [14] :

$$\min_{\mathbf{w}, \xi, b} \frac{1}{2} \|\mathbf{w}\|_H^2 - b + \frac{1}{\nu l} \sum_{i=1}^{l} \xi_i \text{ s.c. } \begin{cases} \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle_H \geq b - \xi_i \\ \xi_i \leq 0, i = 1 \dots l \end{cases}$$

où les ξ_i sont les termes de pénalisation associés à chaque $\mathbf{x}_i$. Les multiplicateurs de Lagrange $\alpha_i, i = 1, \dots, l$, associés à ce problème dans sa formulation duale, déterminent complètement $\mathbf{w}$ et b. On obtient finalement $f_X(\mathbf{x}) = \sum_i \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}) - b$ et $\{\mathbf{x}_j : a_j \neq 0\}$ définit l'ensemble des vecteurs de support (notés VS).

Dans la formulation canonique des OC-SVM, on considère l'unique hypothèse H_0 : « l'observation appartient à la classe C_0 ». A cette hypothèse est associée la fonction de décision suivante : si $f_X(\mathbf{x}) \geq 0$ alors $\mathbf{x}$ vérifie H_0 (décision D_0). Nous proposons de définir une hypothèse H_1 « l'observation n'appartient pas à C_0 » et d'introduire le seuil $\lambda \in \mathbb{R}$ afin de construire une famille de règles de décision :

$$\begin{cases} \text{si } f_X(\mathbf{x}) \geq \lambda, \text{ alors } \mathbf{x} \in C_0 \ (D_0) \\ \text{si } f_X(\mathbf{x}) < \lambda, \text{ alors } \mathbf{x} \notin C_0 \ (D_1) \end{cases} \quad (2)$$

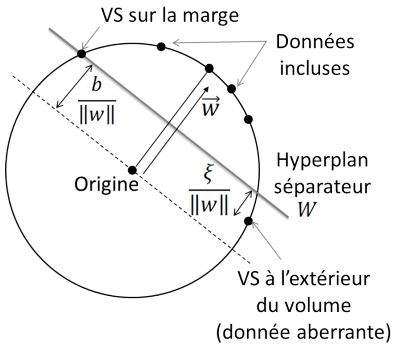

FIGURE 1 – Principe des OC-SVM

Nous pouvons alors définir les probabilités de non-détection et de fausse-alarme, respectivement $P(D_0|H_1)$ et $P(D_1|H_0)$, qui déterminent le point de fonctionnement de notre détecteur.

Dans cette formulation, le seuil λ, qui paramètre une translation de l'hyperplan W dans l'espace des *features* H, permet de contrôler le compromis entre fausse-alarme et non-détection sans avoir à réaliser un nouvel apprentissage. La frontière de Γ qui en résultent dans l'espace d'origine est une ligne de contour de la fonction de décision $f_X(\mathbf{x})$. On ajuste λ expérimentalement en fonction des besoins opérationnels.

Le choix de ν est un problème délicat car pour des valeurs faibles, Γ est estimé dans des régions où la densité de probabilité des données d'apprentissage est très faible (variance d'estimation élevée). A l'inverse, pour des valeurs de ν élevées, un biais important dans l'estimation de Γ pourrait conduire à une représentation sous-optimale de la distribution des données normales. Dans notre formulation le paramètre ν est indépendant des besoins opérationnels et est uniquement conditionné par le signal d'apprentissage : c'est une estimation de la fraction de données qui doivent être exclues du domaine Γ.

La figure 2 illustre les ensembles de fonctions de décision obtenus pour différentes valeurs de ν. Chaque ensemble est représenté par une courbe, ensemble des points de fonctionnement possibles en faisant varier λ après un apprentissage (ν fixé). Sur chaque courbe, un symbole « □ » localise le point de fonctionnement correspondant à la formulation canonique OC-SVM ($\lambda = 0$). Ces premiers résultats montrent que faire varier λ permet de compenser un choix approximatif de ν. De plus, notre approche peut conduire à de meilleures performances (ν fixé, λ varie), plutôt qu'en faisant varier seulement ν.

3 Intégration de la statistique

Le flux audio est traité trame par trame (échantillons audio successifs sur 20 ms avec recouvrement de 50% et fenêtrage de Hamming). De chaque trame est extrait un vecteur de descripteurs « acoustiques » (également appelés paramètres) pour lequel est évaluée la statistique de décision. Cette approche ignore donc les aspects de continuité temporelle inhérents aux signaux audio auxquels nous nous intéressons main-

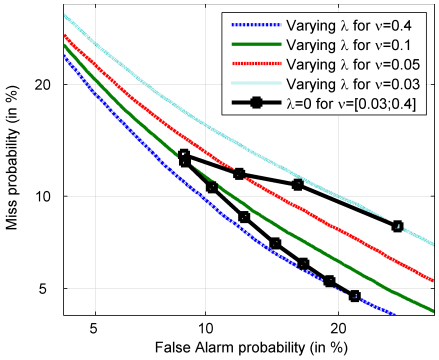

FIGURE 2 – *Détection d'événements sonores anormaux à 10dB.* Influence de ν et λ sur le compromis entre probabilités de fausse-alarme / non-détection. Les symboles □ représentent le point de fonctionnement pour un OC-SVM standard ($\lambda = 0$).

tenant. Dans un premier temps, nous proposons d'appliquer un filtre médian, ce qui conduit à la fonction de décision suivante :

$$\begin{cases} \text{si } \mathcal{M}_M\left(f_X(x_k)\right) \geq \lambda, \text{ alors } x_k \text{ est normal } (D_0) \\ \text{si } \mathcal{M}_M\left(f_X(x_k)\right) < \lambda, \text{ alors } x_k \text{ est anormal } (D_1) \end{cases} \quad (3)$$

où x_k est le vecteur de paramètres mesuré à la trame k, et $\mathcal{M}_M(u_k)$ l'opérateur qui renvoie la valeur médiane de la série $(u_{k-M+1}, \ldots, u_k)$ (M est l'ordre du filtre). Ce filtre permet d'améliorer sensiblement les résultats mais la taille des événements correctement détectés est corrélée à la taille de la fenêtre d'intégration, fixée par M.

Sans *a priori* sur la durée des événements à détecter, nous nous intéressons maintenant à une méthode utilisant des informations de segmentation issues d'un algorithme multi-niveaux en ligne (voir figure 3). L'idée générale est de regrouper hiérarchiquement les trames similaires d'un *buffer* en segments, puis les segments entre eux jusqu'à n'obtenir qu'en seul segment (fusion *bottom-up*). On cherche ensuite dans le dendrogramme ainsi construit le niveau où la segmentation est optimale (recherche *top-down*).

Nous utilisons un *buffer* de 2 secondes et une représentation spectrale comme paramètres de segmentation (énergies moyennes en sortie d'un banc de 24 filtres de largeur de bande constante). Les segments sont représentés par leur vecteur moyenne et on utilise ensuite la distance Euclidienne entre segments pour la fusion. Identifier le niveau de segmentation optimal consiste alors à calculer le coefficient de corrélation intra-segment (entre segments fusionnés), puis retenir le niveau pour lequel tous les coefficients sont au delà d'un seuil fixé (0.98 dans nos évaluations). La segmentation à ce niveau est conservée.

La statistique de décision est intégrée sur chaque segment « homogène », d'où la fonction de décision suivante :

$$\begin{cases} \text{si } \langle f_X(x)\rangle_{S_j} \geq \lambda, \text{ alors } x \text{ est normal } (D_0) \\ \text{si } \langle f_X(x)\rangle_{S_j} < \lambda, \text{ alors } x \text{ est anormal } (D_1) \end{cases} \quad (4)$$

où $\langle u_k\rangle_{S_j}$ est l'opérateur qui renvoie la valeur moyenne de la série $u_k, \forall k \in S_j$ et S_j l'ensemble des indices des trames appartenant au segment j.

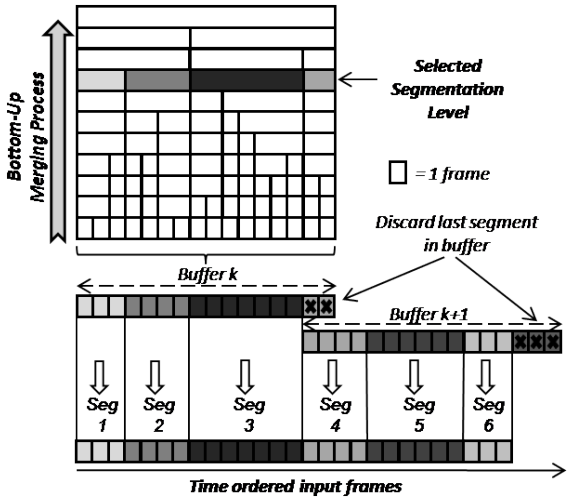

FIGURE 3 – Fonctionnement de la segmentation multi-niveaux.

4 Base de données et évaluation

Nous avons développé une procédure pour combiner des événements anormaux avec des signaux d'ambiance réels issus d'un environnement sous audio-surveillance. Notre approche présente les avantages suivants : contrôle du Rapport Signal-Bruit (SNR), parfaite connaissance de la position des événements ajoutés, génération de grandes quantités de signaux de test. Parmi les méthodes de la littérature pour la mesure de niveau de bruit [1, 4, 8, 9], nous utilisons la norme ITU-R468 [7] dans le calcul des SNRs. Celle-ci cible les fréquences communes aux événements anormaux (signal, généralement haute fréquence) et à l'ambiance normale (bruit, généralement riche en basses fréquences). Dans notre application, le SNR est défini globalement (gain calculé depuis les énergies moyennes sur l'ensemble des signaux ambiance et événement). L'événement est ensuite inséré avec une amplitude constante à différentes positions dans le signal d'ambiance, le SNR local varie donc pour chaque événement inséré.

Pour l'évaluation, nous avons généré une base de données à partir de 18 signaux d'ambiance de 10 minutes chacun, enregistrés dans une station de métro à Rome [2]. Les événements, d'une durée de 1 seconde, sont extraits d'une base de données commerciale [15]. Nous avons considéré 96 événements répartis en 27 catégories. L'ensemble d'apprentissage contient 6 des signaux d'ambiance (1 heure) et l'ensemble de test 5760 signaux, soit les combinaisons parmi 5 valeurs de SNR, 96 événements et les 12 ambiances non utilisées pour l'apprentissage (960 heures, un événement anormal toutes les 12 secondes).

Les paramètres acoustiques utilisés pour la détection sont les énergies moyennes issues d'un banc de 32 filtres de largeur de bande constante. Les hyperparamètres du OC-SVM ont été déterminés expérimentalement (grille de recherche uniforme) à $\nu = 10^{-3}$ et $\sigma = 10$ (paramètre du noyau RBF), conduisant à un choix de 1548 vecteurs de support parmi 360033 points (trames). Nous avons appliqué les fonctions de décision

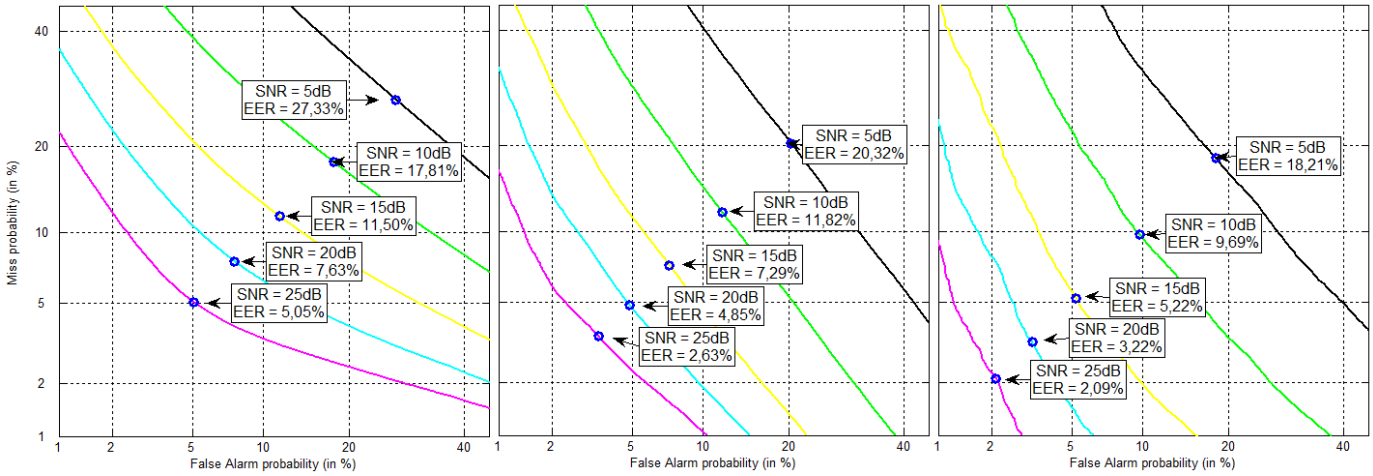

FIGURE 4 – Courbes DET (*Detection Error Trade-off*) pour différents SNR sans intégration temporelle (à gauche), avec un filtrage médian (1 seconde) de la décision (au milieu), et avec intégration à partir des informations de segmentation (à droite).

1- trame par trame, 2- après filtrage médian d'une seconde et 3- en utilisant la segmentation adaptative. Les résultats sont présentés sur la figure 4 [11] ; chaque courbe correspond à un SNR fixé et illustre le compromis entre probabilité de fausse-alarme et probabilité de non-détection (ensemble de points de fonctionnement possible) quand λ varie.

5 Conclusion et perspectives

Nous avons introduit une méthode non supervisée de détection d'événements sonore anormaux pour des applications de surveillance. Basée sur les OC-SVM, notre approche permet de construire des familles de règles de décision. Il est ensuite possible de modifier les caractéristiques du détecteur sans nécessiter de nouvel apprentissage. Nous avons proposé un algorithme de segmentation performant et démontré de l'intérêt d'intégrer temporellement la statistique de décision suivant ces informations de segmentation. Nous avons enfin décrit un processus pour l'élaboration de bases de données audio et présenté des résultats qui illustrent l'efficacité de notre approche non supervisée. Afin de minimiser le risque lié à la variance d'estimation lorsque λ est modifié, nous envisageons d'investiguer des techniques d'évaluation de la pertinence d'un hypervolume appris par OC-SVM. L'étape de détection présentée devra également être intégrée avec un module de classification pour pouvoir être comparée aux approches supervisées de l'état de l'art.

Références

[1] British Broadcasting Corporation. *The Assessment of Noise in Audio Frequency-Circuits - EL17*. Engineering Division Research Report, 1968.

[2] CARETAKER Project : Content Analysis REtrieval Technologies to Apply Knowledge Extraction to massive Recording. *FP6 IST 4-027231*. 2006-2008.

[3] M. Cettolo, M. Vescovi, and R. Rizzi. Evaluation of bic-based algorithms for audio segmentation. *Computer Speech & Language*, 19(2) :147–170, 2005.

[4] H. Fletcher and W. Munson. Loudness, its definition, measurement and calculation. *Journal of Acoustical Society of America*, 5 :82–108, 1933.

[5] J. Glass and V. Zue. Multi-level acoustic segmentation of continuous speech. In *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, volume 1, pages 429–432, 1988.

[6] J.-L. Husson and Y. Laprie. A new search algorithm in segmentation lattices of speech signals. In *Proceedings of the 4th International Conf. on Spoken Language Processing*. IEEE Computer Society, 1996.

[7] Internation Telecommunication Union. *Measurement of Audio-Frequency Noise Voltage Level in Sound Broadcasting*. Recommandation, Broadcasting Service, 1986.

[8] International Electrotechnical Commission. *IEC-61672-2 Sound Level Meters - Part 2 : Pattern Evaluation Tests*. 2003.

[9] International Organization for Standardization. *ISO-226 Acoustics - Normal Equal-Loudness-Level Contours*. ISO Standards, 2003.

[10] D. Istrate, E. Castelli, M. Vacher, L. Besacier, and J.-F. Serignat. Information extraction from sound for medical telemonitoring. *Information Technology in Biomedicine, IEEE Transactions on*, 10(2) :264–274, 2006.

[11] National Institute of Standards and Technology. Det-curve plotting software", information technology laboratory, detware v.2.1.

[12] A. Rabaoui, M. Davy, S. Rossignol, Z. Lachiri, and N. Ellouze. Improved one-class svm classifier for sounds classification. In *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pages 117 –122, 2007.

[13] T. Räty. Survey on contemporary remote surveillance systems for public safety. *Transactions on Systems, Man, and Cybernetics, Part C : Applications and Reviews*, 40(5) :493–515, 2010.

[14] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13 :1443–1471, 2001.

[15] Sound Ideas. *The Series 6000 "The General" Sound Effect Library*.

[16] M. Tohmé and R. Lengellé. Maximum margin one class support vector machines for multiclass problems. *Submitted to Pattern Recognition Letters*, 2011.

[17] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti. Scream and gunshot detection and localization for audio-surveillance systems. In *Proceedings of the 2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 21–26. IEEE Computer Society, 2007.

[18] M. Valera and S. Velastin. Intelligent distributed surveillance systems : a review. *Vision, Image and Signal Processing, IEE Proceedings -*, 152(2) :192–204, 2005.