Schéma de compression hybride de séquences d'images avec régions d'intérêt orientées mouvement

E. Nguyen et C. Labit

IRISA, Campus de Beaulieu 35042 Rennes Cedex, France

Dans cet article on propose une méthode de compression sélective de séquence d'images basée région pour la transmission à très bas débit. La méthode repose sur la notion de niveau d'intérêt affecté à chaque région de la scène. La qualité de reconstruction à débit fixé est alors distribuée de manière inhomogène sur les différentes régions. Le codeur de structure hybride usuelle utilise une représentation sous-bandes globale. On opère une allocation de ressources basée région en considérant explicitement la localisation espace-fréquences des fonctions de base. On décrit particulièrement le cadre applicatif où la sélection des régions d'intérêt et la compression sélective sont définies sur des critères de mouvement. L'évaluation de la méthode est illustrée sur des séquences réelles de trafic routier.

1. INTRODUCTION

Dans le cadre de la compression de séquences d'images les techniques d'estimation/compensation de mouvement associées aux techniques de codage par transformées ou sous-bandes sont étudiées de manière intensive pour la conception de codeurs à structure "hybride" [1]. Cette structure est à la base des recommandations ou normes de compression comme H261 (CCITT) ou MPEG-1,2 (ISO). L'approche usuelle de compression consiste à distribuer les ressources en débit R ou en qualité (distorsion D) de manière globale sur l'ensemble du support spatio-temporel du signal sans aucun *a priori* sur le contenu de la scène à coder. Notre approche se situe dans le cadre des techniques récentes de codage basées sur une représentation "objets" pour la compression à très bas débit. On se propose d'opérer la compression conditionnellement à la définition de *régions d'intérêt*. On entend habituellement par région d'intérêt les zones de l'image comportant l'information visuelle essentielle nécessaire à l'interprétation des images décodées au récepteur. La définition de ces régions d'intérêt diffère selon le type d'application visée (exemples : détection d'objets mobiles présents dans la scène ou ayant un déplacement particulier au cours de la séquence dans le cas de la télé-surveillance, zones de type tête et épaules en visiophonie, structures d'intérêt pathologique pour la consultation de base de données d'imagerie médicale à distance etc ...). Du point de vue de la compression, l'objectif est alors de concentrer les ressources $R - D$ sur les différentes régions d'intérêt au dépend des autres zones de l'image. L'étude est donc fondée sur l'existence de *région(s) d'intérêt* (ROI) désignée(s) interactivement ou sélectionnées automatiquement à partir d'une segmentation $\{\mathcal{R}_k\}$ basée sur des critères dépendant de l'application (texture, mouvement). Une étape de sélection classe les régions par niveaux de priorité :

$$\{O_k \ : \ O_{ROI=0} > O_1 > \ldots > O_{R-1}\} \qquad (1)$$

Pour une transmission à débit fixé l'opération de compression sélective consiste à distribuer la qualité de reconstruction sur les différentes régions définies selon la hiérarchie d'intérêt (1). On se définit alors deux critères *a priori* reliés aux deux modules algorithmiques distincts du codeur : le premier critère concerne la segmentation et l'attribution des niveaux d'intérêt pour chaque région dans le module d'analyse, le second critère est relié à la stratégie d'allocation des ressources dans le module de compression.

On s'intéresse ici à des critères de mouvement. Une étape générale d'analyse/segmentation au sens du mouvement est opérée avant la phase de compression. L'information de mouvement est utilisée de manière qualitative pour la sélection des régions d'intérêt et de manière quantitative pour la compensation dans un schéma de codage hybride. La compression sélective est spécifié en utilisant un *a priori* psychovisuel relatif à l'activité visuelle de l'observateur.

In this paper we present an approach of selective compression for very low bit rate sequence coding. The method is based on the concept of Region Of Interest (ROI). Under rate constraint, spatial reconstruction qualities are distributed among regions according to their priority levels. The problem of selective compression is considered in the context of usual hybrid DPCM subband coding. Region-based rate-distortion allocation is performed taking into account both spatial and frequency localization of the subband representation. We emphasize on applications where ROI selection and distortion allocation are based on motion criteria. The performance of the described scheme is illustrated for a real scene of traffic control.

2. CODAGE HYBRIDE AVEC RÉGIONS D'INTÉRÊT

2.1. Approche orientée mouvement

Dans de nombreux cas pratiques, le mouvement est une primitive essentielle pour l'analyse de scènes dynamiques. Par ailleurs l'estimation du mouvement permet la réduction de la redondance temporelle du signal vidéo et l'application de critères psychovisuels à des fins de compression. Cependant les techniques usuelles d'estimation de mouvement en codage vidéo n'offrent qu'une analyse sommaire de la scène (partition arbitraire en blocs, modèles translationnels) et génèrent des effets de blocs à très bas débit. Une segmentation au sens du mouvement permet alors d'utiliser des

données plus complexes pour l'interprétation et la sélection des régions d'intérêt [3]. L'étape d'analyse fournit une information compacte $\{\mathcal{R}_k, \Theta_k^{init}\}$ constituée d'une carte de régions $\{\mathcal{R}_k\}$ et des descripteurs de mouvement associés $\{\Theta_k^{init}\}$ (typiquement des modèles linéaires en fonction des coordonnées spatiales). Dans un schéma de codage hybride une estimation/segmentation jointe doit être effectuée dans la boucle MICD de prédiction au sens du mouvement afin de minimiser l'erreur de prédiction transmise. Cependant, du fait des erreurs de quantification introduites, l'estimation / segmentation de mouvement peut être biaisée et peut conduire à un mauvais positionnement des frontières de mouvement. La segmentation et les descripteurs de mouvement $\{\Theta_k^{init}\}$ estimés en boucle ouverte sont alors utilisés directement pour la compensation. Ce choix conduit à privilégier l'information de mouvement cohérente et contribue à réduire la perception visuelle des erreurs de prédiction à la reconstruction. L'information d'analyse $\{\mathcal{R}_k, \Theta_k\}$ nécessaire pour l'interprétation de la scène et la décompression du signal est codée et transmise au récepteur. En première approche l'étape d'analyse est indépendante de l'étape de compression : le débit R_d de transmission de l'information de mouvement est imposé par l'algorithme de segmentation. Ce débit peut être régulé en agissant sur la précision de la segmentation. En particulier dans l'approche Markovienne proposée dans [2], on définit l'appartenance d'un pixel à une région de mouvement homogène à une erreur de déplacement résiduel δ près. On observe qu'en pratique le débit R_d est dominé par le coût de transmission de la segmentation $\{\mathcal{R}_k\}$. La segmentation est codée sans perte sur la base d'une représentation contour inter-pixel par un chainage de Freeman différentiel avec prise en compte des "points triples" [4]. L'utilisation d'un codeur arithmétique adaptatif utilisant un modèle Markovien d'ordre 3 permet d'obtenir un compactage efficace (de l'ordre de 1.3 bit par élément de contour).

L'opération de compression *sélective* proprement dite, pour un débit de transmission global donné R_g, conduit à distribuer la qualité de reconstruction sur les différentes régions classées selon une hiérarchie d'intérêt $\{O_k(\Theta_k)\}$. A débit d'analyse fixé R_d on définit un problème d'allocation des ressources basée région sous contrainte de débit $R_e = R_g - R_d$.

2.2. Codage hybride basé région

2.2.1. Représentation sous-bandes/région

L'approche de compression sélective nécessite de définir une représentation basée région efficace du point de vue des performances de codage. Dans le cadre d'un codage hybride, on propose d'adapter la quantification selon les niveaux d'intérêt sur la base d'une représentation sous-bandes ou par transformées du signal d'erreur de prédiction. La représentation en sous-bandes choisie est une représentation orthogonale séparable stationnaire sur le support spatial de l'image. La stationnarité de la décomposition en sous-bandes permet de définir les distorsions relatives entre régions pour les mêmes bases espace-fréquence.

D'un point de vue de la structure de calcul, le codage par transformée en sous-bandes ou ondelettes discrètes est basé généralement sur la donnée de bancs de filtres RIF d'analyse $\{h_i\}$ et de synthèse $\{g_i\}$ avec facteurs de sous-échantillonnage critiques $\{N_i\}$ permettant de représenter le signal de manière exacte et non-redondante. La représentation est obtenue par projection du signal sur des bases $\tilde{h}_i(j - N_i n)$ localisées en espace (support compact $\tilde{L}_i(n)$) et en "fréquence". Dans le cas orthogonal, les bases d'analyse et de synthèse sont les mêmes. On propose ici d'utiliser explicitement la propriété de bonne localisation espace-fréquence des fonctions de base associées. Cette localisation espace-fréquence permet naturellement le codage sélectif des zones spatiales d'intérêt en prenant en compte l'analyse en fréquence pour le rendu du signal reconstruit après synthèse. On utilise explicitement la localisation spatiale en définissant des entités sous-bandes/régions $\{e_{i_k}\}$ par "projection" de la segmentation dans le domaine transformé. Cette projection s'effectue de manière hiérarchique selon les $\{O_k\}$ afin de lever l'ambiguïté aux frontières des régions tout en préservant la hiérarchie des informations spatiales (labels k) dans la phase de décimation :

$$label\,[e_i(n)] = Arg \max_{label \in \tilde{L}_i(n)} O_{label} \qquad (2)$$

On doit remarquer que du fait du recouvrement des fonctions de base cette représentation ne considère plus l'indépendance des informations spatiales de part et d'autre des frontières des régions définies dans l'espace pixel. Cette limitation est cependant compensée par la réduction des "effets de bloc" aux frontières des régions que l'on observe à bas débit dans le cas d'un traitement spatial indépendant (par l'utilisation d'extensions [5]) et par une complexité algorithmique inférieure.

2.2.2. Allocation sélective des ressources $R - D$

La compression proprement dite est obtenue par *quantification* de la représentation $\{e_{i_k}\}$. De manière classique la norme quadratique l^2 (EQM) est utilisée comme mesure de distorsion pour des opérateurs de projection orthogonaux. Des pondérations $\{W_i\}$ agissant sur les sous-bandes peuvent être introduites de manière à prendre en compte la sensibilité relative en fréquence du Système Visuel Humain (SVH). On définit alors une norme quadratique pondérée qui dépend de la décomposition sous-bandes choisie. En supposant un codage indépendant de chaque source e_{i_k} on étend la notion de distorsion pondérée en fréquence à une distorsion pondérée en sous-bandes/région D_w associée à un débit de transmission R en bits par pixels (bpp) :

$$D_w = \sum_k \sum_i \Omega_{i_k}\, \eta_{i_k}\, D_{i_k}\, /N_i \;;\; R = \sum_k \sum_i \eta_{i_k}\, R_{i_k}\, /\, N_i \qquad (3)$$

où les D_{i_k} et R_{i_k} sont respectivement les distorsions quadratiques et les débits associés aux entités $\{\hat{e}_{i_k}\}$ de facteurs d'occupation $\{\eta_{i_k}\}$. Les pondérations "espace-fréquence" $\{\Omega_{i_k}\}$ sont introduites de manière à définir un critère de qualité de reconstruction relatif aux niveaux d'intérêt $\{O_k\}$ associés à chaque région.

Les pondérations sous-bandes/régions peuvent se définir comme le produit de 2 contributions :

$$\Omega_{i_k} = F_k\, W_{i_k} \qquad (4)$$

"spatiale" F_k et "fréquentielle" W_{i_k}. Ces pondérations permettent de définir *a priori* les qualité de reconstruction

relative des régions de manière *quantitative* (en erreur quadratique moyenne par symbôle) et/ou de manière *perceptuelle* en prenant en compte dans une certaine mesure la sensibilité en fréquence du SVH. D'un point de vue quantitatif on se place sous l'hypothèse de modèles de fonctions débit-distorsion convexes $D_{i_k}(R_{i_k}) = \alpha_{i_k} exp(-\beta R_{i_k})$. Ces modèles dits "haute résolution" supposent un même type de quantificateur pour chaque source sous-bande/région e_{i_k}; les paramètres α_{i_k} dépendent des statistiques de chaque source. Avec cette modélisation il est aisé de montrer que l'allocation en distorsion optimale pour chaque source sous-bande/région est donnée par : $D^*_{i_k} \simeq D^*_w/\Omega_{i_k}$. Les pondérations Ω_{i_k} peuvent être alors interprétées comme des facteurs d'échelle en erreur quadratique sur les contributions des sous-bandes /régions dans la mesure pondérée D^*_w globale. Pour $W_{i_k} = 1 \; \forall \; i_k$ les coefficients F_k sont simplement des facteurs d' échelle en erreur quadratique sur les contributions des régions. Les facteurs de pondération W_{i_k} sont calculés comme des facteurs de forme spectraux de bruit de quantification [6]. Ils dépendent des fonctions de base choisies et d'une fonction de transfert de modulation (FTM) du SVH approximée au premier ordre : $W_k(f_x, f_y)$. On peut alors définir selon une *stratégie fixée* une échelle de fonctions de sensibilité relative pour les régions $\{\mathcal{R}_k\}$. Notons que l'utilisation des facteurs F_k permet également dans ce contexte de prendre en compte des effets non-linéaires comme la dépendance en luminance L_k des FTM pour chaque région avec $W_k(L_k, f_x, f_y) \simeq C(L_k).W_k(L_0, f_x, f_y)$ (modèle simplifié de Watson).

2.2.3. Un a priori de sélection psychovisuel: suivi du regard

Le critère d'allocation des ressources dans le module de compression est reporté sur le choix des facteurs de pondération spatiale F_k et fréquentielle W_{i_k}. Les facteurs spatiaux F_k peuvent être utilisés pour la compression en mode *intra* de la première image sur la base d'une segmentation spatiale. Dans le cadre d'un codeur avec régions d'intérêt orientées mouvement l'idée est d'adapter la compression selon la perception des objets en mouvement dans la scène. Il est cependant connu que la perception d'un objet en mouvement (et donc des erreurs de reconstruction associées) dépend de l'activité visuelle de l'observateur (i.e. si l'objet est ou n'est pas suivi du regard) [7]. Dans l'hypothèse naturelle où l'observateur ne s'intéresse qu'à une seule région d'intérêt à un instant donné, on définit les *vitesses visuelles* relatives i.e. les vitesses relatives au point de fixation : $\tilde{v}_k = v_k - v_{ROI}$ où les $\{v_k\}$ sont les vitesses moyennes apparentes des régions $\{\mathcal{R}_k\}$ calculées à partir des descripteurs de mouvement $\{\Theta_k\}$ (paramètres translationnels). Les considérations psychovisuelles sont ici limitées à la sensibilité en fréquence avec le choix des fonctions de sensibilité $W_k(f_x, f_y)$. On considère alors la dépendance de la sensibilité en fréquence en fonction des vitesses relatives $\tilde{v}_k$: $W_k(f_x, f_y, \tilde{v}_k)$. En supposant l'isotropie de la réponse en vitesse et en fréquence, on utilise ici un modèle dérivé de mesures expérimentales décrit dans [8] dans le cadre de la vision fovéale.

L'hypothèse de suivi du regard ne s'applique que dans les limites des vitesses admissibles du phénomène de poursuite visuelle en mode SPEM (*smooth pursuit eyes movement*). Des donnés expérimentales montrent toutefois que la majorité des objets en déplacement dans les scènes réelles ont une vitesse inférieure à la limite admissible [9].

2.2.4. Optimisation R-D_w

Le problème classique de l'optimisation sous contrainte des ressources débit-distorsion (R-D_w) est résolu dynamiquement à partir d'un choix discret et fini de conditions de quantification $\{q_{i_k}\}$ admissibles pour chaque région dans chaque sous-bande. On se limite dans le cas présent à une simple quantification scalaire uniforme avec encodage arithmétique des indices de quantification. On utilise une technique de programmation convexe basée sur une formulation par multiplicateur de Lagrange [10] qui fournit la solution optimale $\{q^*_{i_k}\}$ du problème sur l'enveloppe convexe des points de fonctionnement R-D_w admissibles. Cette technique nécessite la construction des fonctions débit-distorsion de quantification $\{R_{i_k}(q_{i_k}), D_{i_k}(q_{i_k})\}$ calculées en ligne sur la représentation sous-bandes/régions. Du fait de la séparabilité de la formulation (3), les calculs nécessaires sont locaux à chaque sous-bande/région ce qui les rend potentiellement parallélisables.

Des solutions sous-optimales évitant le calcul des fonctions de quantification peuvent être également obtenues sur la base de modèles paramétriques des densités de probabilités des sous-bandes tels que des modèles de gaussiennes généralisées [11].

3. RÉSULTATS EXPÉRIMENTAUX

L'évaluation d'un tel schéma a été réalisée sur des séquences d'images réelles pour des scènes de trafic routier où la sélection des régions d'intérêt est effectuée sur des critères simples (mouvements latéraux directionnels, mouvements de rapprochement ou d'éloignement) [3]. On donne l'exemple d'une scène typique de télé-surveillance. Les régions d'intérêt sont désignées comme les régions ayant un déplacement translationnel de la gauche vers la droite. Une représentation ondelette multirésolution sur 3 niveaux est utilisée sur à partir de filtres de Daubechies orthogonaux de longueur 4 [12].

Dans cet exemple on a utilisé uniquement la pondération en fréquence selon l'*a priori* psychovisuel de suivi du regard. L'optimisation R-D_w est définie sous une contrainte de débit $R_e = 0.1bpp$ sur l'erreur de prédiction. Dans cet exemple simple le coût de l'information de mouvement R_d est très faible : de l'ordre de 6.10^{-3} bpp en moyenne. Les images de la FIG. 1 montrent l'image originale avec surimposition de la segmentation, l'image reconstruite à $t = 46$ et l'image d'erreur associée (multipliée par un facteur 5 et translatée de 128). On observe un effet de lissage sur les contours des objets en déplacement relatif par rapport à la région d'intérêt. Ceci est en accord avec la sensibilité visuelle qui décroît dans les hautes fréquences en fonction de la vitesse visuelle. Lorsque l'objet d'intérêt est suivi du regard la visibilité des artefacts est réduite. On donne également l'image d'erreur équivalente pour la même compression sans avoir tenu compte de l'approche avec région d'intérêt : d'un point de vue de l'allocation des ressources débit-distorsion l'approche avec *a priori* visuel concentre naturellement l'information et la qualité sur la région d'intérêt.

4. CONCLUSION ET PERSPECTIVES

L'approche de compression avec "régions d'intérêt" constitue une alternative pour le codage vidéo à très bas débits . Dans cet article nous avons décrit un schéma de compression basé sur une analyse au sens du mouvement et sur une structure de codeur hybride usuelle. La localisation spatiale de la représentation en sous-bandes permet naturellement une compression sélective des zones spatiales d'intérêt. La localisation en fréquence permet de prendre en compte dans une certaine mesure des critères de perception dynamique. Pour des débits de l'ordre de 0.1 bpp le gain en qualité visuelle est notable. Dans notre approche, du fait de la représentation choisie, seuls des critères de perception relative en fréquence ont été envisagés sous l'hypothèse de conditions de vision fovéale. Des représentations se rapprochant d'un modèle plus complet de la vision intégrant par exemple l'échantillonnage irrégulier opéré par la rétine doivent être explorées.

5. RÉFÉRENCES

[1] GHARAVI H. – Subband coding of video signals. – *Subband image coding* (Kluwer Academic Press), pp. 229–271, 1991.

[2] ODOBEZ J.M – Estimation, détection et segmentation du mouvement: une approche robuste et markovienne. – *Thèse de l'Université de Rennes I.*, Décembre 1994.

[3] BOUTHEMY P. et FRANCOIS E. – Motion Segmentation and Qualitative Dynamic Scene Analysis from an Image Sequence. – *Int. J. Comp. Vision*, Vol. 10, No 2: pp. 157–182, 1993.

[4] MARQUES F. et al. – Shape and location coding for contour images. – *Actes PCS*, pp 18-6, 1993.

[5] BARNARD H.J, WEBER J.H et BIEMOND J.– A region-based discrete wavelet transform. – *Actes EUSIPCO*, pp. 1234–1237, 1994.

[6] VANDENDORPE L. – Optimized quantization for image subband coding. – *Signal Process.: Image Comm.*, Vol. 4, No 1: pp. 65–79, 1991.

[7] GIROD B. – Eyes Movements and Coding of Video sequences. – *Actes SPIE-VCIP*, Vol. 1001, pp. 398–405, 1988.

[8] KELLY D.H. – Motion and vision II. Stabilized spatio-temporal threshold surface – *J. Opt. Soc. Am*, Vol. 69, No 2: pp. 1340–1349, 1979.

[9] BONSE T. – Visually adapted temporal subsampling of motion information. – *Signal Process.: Image Comm.*, 6 : pp. 253–266, 1994.

[10] RAMCHANDRAN K. – Joint optimization techniques in image and video coding with applications to multiresolution digital broadcast. – *Thèse de l'Université de Columbia, New York*, Juin 1993.

[11] NGUYEN E. ,LABIT C. – Définition quantitative des matrices de pondération psychovisuelles pour la quantification adaptée en codage sous-bandes d'images – *Actes GRETSI-TDSI*, pp. 419–422, 1993.

[12] DAUBECHIES I. – Orthonormal bases of compactly supported wavelets – *Comm. on Pure Appl. Math.*, Vol. XLI, pp. 909–996, 1988.

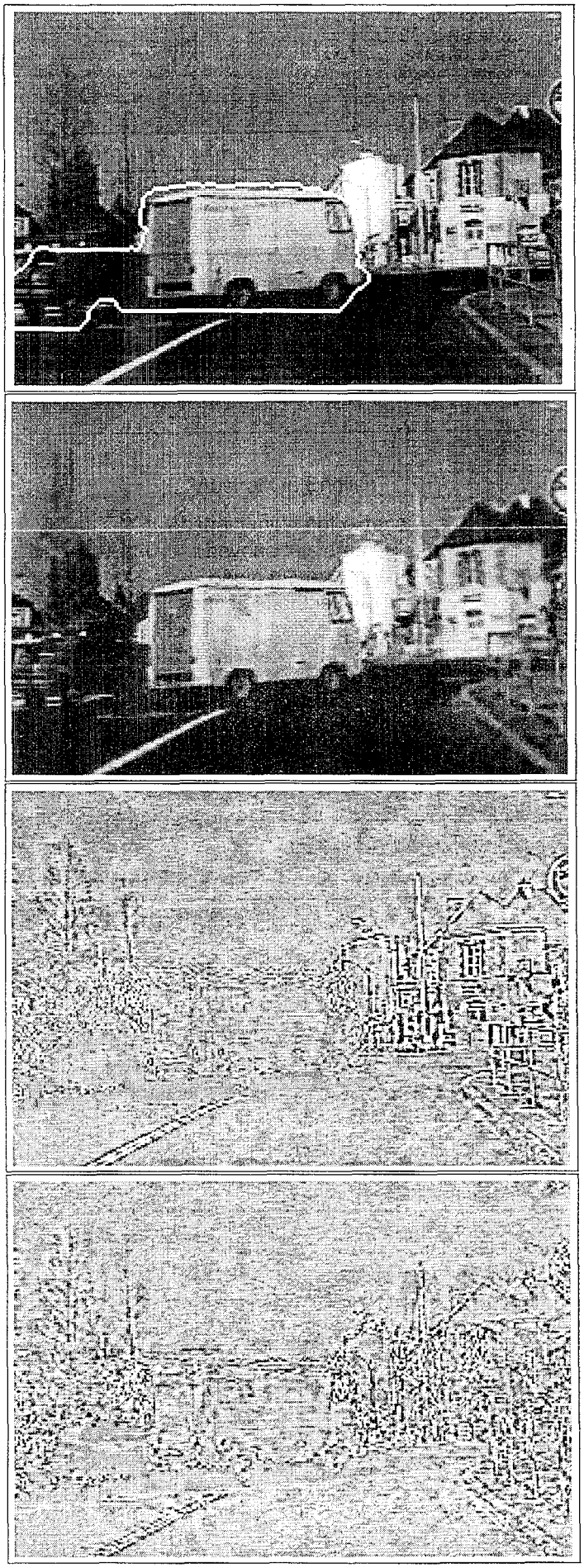

FIG. 1 - Image originale et région d'intérêt, image reconstruite, images d'erreurs.