ESTIMATION DE MOUVEMENT BASÉE SUR DES RÉGIONS POUR LE CODAGE DE SÉQUENCES TÉLÉVISUELLES

Henri SANSON

CCETT

4, rue du Clos Courtel, F 35510 - Cesson Sévigné, France

RÉSUMÉ

Cet article présente une méthode réalisant conjointement l'identification de modèles de mouvement dans les séquences d'images et la segmentation de ces images en régions homogènes vis-à-vis des modèles estimés. Les modèles choisis sont de type polynomial, l'estimateur proprement dit est de type différentiel. Une analyse hiérarchique descendante est utilisée pour réaliser d'une part la segmentation, et d'autre part constituer la partie prédiction de l'estimateur différentiel. Cette méthode intégrant facilement la contrainte d'une quantité d'information de mouvement limitée est particulièrement bien adaptée à des fins de codage. Le but recherché est néanmoins de fournir un mouvement apparent le plus cohérent possible avec le mouvement réel.

ABSTRACT

This paper presents a method realizing jointly the identification of motion models in image sequences and the image segmentation in regions homogeneous with respect to the estimated motion models. The chosen models are of polynomial type, and the estimator itself is of differential type. A top-down hierarchical analysis is used to achieve the segmentation on the one hand, and to provide the differential estimator with initial guesses on the other hand. This method, integrating easily the constraint of a bounded amount of data for the motion is well suited to coding purposes. Nevertheless, the aim is to furnish an apparent motion as consistent as possible with the actual movements.

I. INTRODUCTION

L'analyse de mouvement, et plus généralement la prédiction temporelle, joue un rôle critique en codage de séquences d'images, et cette technique supporte aujourd'hui une grande partie des espoirs de gain en taux de compression. Parmi les différentes approches envisagées, les techniques à base de régions [1], ou d'objets [2], sont quant à elles des candidates de choix pour atteindre des améliorations significatives, ceci pour plusieurs raisons:

- elles conduisent à une représentation compacte de l'information de mouvement, les zones animées de façons différentes étant généralement peu nombreuses dans les séquences d'images.
- le coût de description du mouvement est essentiellement concentré sur la représentation des frontières de régions. On peut donc adapter la modèlisation du mouvement, en particulier sa dynamique et sa précision, pour chaque région. On peut également adapter la complexité du modèle de prédiction, par exemple en prenant en compte les variations d'illumination.
- Même pour une segmentation imparfaite, les erreurs engendrées sont visuellement moins gênantes que les classiques effets de blocs observés à relativement bas débit avec les techniques à base de blocs.
- le fait de posséder une information sur les frontières du mouvement doit permettre une détermination plus aisée des zones d'occlusion. Il est d'ailleurs souhaitable de faire coopérer la segmentation et le traitement des occlusions pour aboutir à des résultats plus fiables.
- une caractéristique importante de l'approche régions est la possibilité de réaliser un suivi temporel des zones mobiles au cours du temps. Ce suivi temporel peut alors avoir des impacts sur la robustesse des segmentations obtenues, mais aussi sur le débit consacré à l'information de mouvement, qui peut encore être réduit.

Dans ce qui suit, nous proposons une méthode réalisant un découpage des images d'une séquence en régions connexes, homogènes vis-à-vis d'un modèle polynomial de mouvement, ainsi que l'estimation des paramètres de ces modèles.

Les points clé de cette méthode sont l'analyse hiérarchique en quadtree avec recouvrements [3] [4], à la fois pour la partie prédiction de l'estimateur et pour affiner les frontières de la segmentation, l'approche différentielle couplée à une analyse multirésolution pour la partie correction de l'estimateur, ainsi que la théorie des graphes pour la partie fusion de régions de la segmentation. Outre ces outils de base, la technique proposée ici repose sur le principe d'une coopération étroite entre estimation et segmentation caractérisée par un rebouclage de l'une sur l'autre et réciproquement.

II. MODÈLES DE MOUVEMENT

Le calcul de la projection sur le plan image du mouvement d'un objet de forme régulière (de surface polynomiale) animé d'un mouvement rigide conduit à une expression polynomiale du champ des vitesses sur la région 2D correspondant à la projection de l'objet. L'exemple le plus courant est celui d'une facette plane [3], conduisant à un modèle quadratique. Le modèle polynomial peut de plus rendre compte de mouvements de déformation, puisqu'il correspond de fait à une déformation 2D dans le plan image en général. Nous considérons donc des champs dont l'expression est donnée par:

$$\begin{cases} d_x(x, y, A^x) = \sum_{i=0}^{n} \sum_{j=0}^{i} a_{ij}^x . x^{i-j} y^j \\ d_y(x, y, A^y) = \sum_{i=0}^{n} \sum_{j=0}^{i} a_{ij}^y . x^{i-j} y^j \end{cases}$$

$d_x(x,y)$, $d_y(x,y)$: composantes en x et y du vecteur de déplacement au point (x,y).

$A^x = (a_{00}^x,\ldots,a_{nn}^x)^t$, $A^y = (a_{00}^y, \ldots,a_{nn}^y)^t$: vecteurs de paramètres pour ces composantes.

Dans la suite, nous noterons $A = (A^{xt}, A^{yt})^t$ le vecteur correspondant au modèle complet.

Le choix de l'ordre du modèle est a priori un paramètre de l'algorithme. Cependant, une stratégie plus fine d'adaptation de l'ordre du modèle par région peut être envisagée.

III. ESTIMATION DES PARAMÈTRES DE MOUVEMENT

Nous considérons ici une région donnée de forme quelconque. Qu'elle ait physiquement un mouvement homogène ou non, il est toujours possible de lui attribuer un modèle de mouvement, au sens d'une fonction de coût basée sur l'énergie de l'erreur de prédiction:

$$E(R, A) = \sum_{(x,y)\in R} DFD^2(x, y, A)$$

où DFD désigne la différence inter-images déplacées, dont l'expression est :

$$DFD(x, y, A) = L_t[x, y] - L_{t-1}[x - d_x(x, y, A_x), y - d_y(x, y, A_y)]$$

et $L_t(x,y)$ est le signal de luminance à la position (x,y) au temps t.

La minimisation de cette fonctionnelle est réalisée par une méthode itérative classique d'analyse numérique basée sur les dérivées: la méthode de Gauss-Newton ou pseudo-Newton. Partant d'un modèle initial A_0, on effectue un développement au 1er ordre de la luminance aux points déplacés par ce modèle, par rapport aux composantes d'une correction δA de ce modèle. Cette approximation reportée dans l'expression de E conduit à une approximation quadratique de celle-ci, donc au quasi 2ème ordre.

Correction différentielle simple

On considère donc le critère à minimiser suivant:

$$E(R, A^0+\delta A) = \sum_{(x,y)\in R} \{DFD(x,y,A^0) - g_x.\sum_{i,j} \delta a^x_{ij}.f_{i,j} - g_y.\sum_{i,j} \delta a^y_{ij}.f_{i,j}\}^2$$

en posant $f_{i,j}(x,y) = x^{j-i}y^i$.

En définissant les éléments de matrice suivants:

$$\begin{cases} R^x_{df}(i,j) = \sum_R g_x.DFD.f_{i,j} \\ R^y_{df}(i,j) = \sum_R g_y.DFD.f_{i,j} \end{cases} \qquad \begin{cases} R^{xx}_{ff}(i,j,k,l) = \sum_R g^2_x.f_{i,j}.f_{k,l} \\ R^{xy}_{ff}(i,j,k,l) = \sum_R g_x.g_y.f_{i,j}.f_{k,l} \\ R^{xx}_{ff}(i,j,k,l) = \sum_R g^2_y.f_{i,j}.f_{k,l} \end{cases}$$

où g_x et g_y sont les composantes du gradient de $L_{t-1}(x-d_x,y-d_y)$. L'annulation des dérivées partielles du critère E par rapport aux paramètres δA conduit au système linéaire suivant:

$$\begin{pmatrix} R^{xx}_{ff} & R^{xy}_{ff} \\ R^{xy}_{ff} & R^{yy}_{ff} \end{pmatrix} \begin{pmatrix} \delta A^x \\ \delta A^y \end{pmatrix} = - \begin{pmatrix} R^x_{df} \\ R^y_{df} \end{pmatrix}$$

On peut simplifier ce système en faisant l'hypothèse d'une distribution isotrope des gradients photométriques dans la région et en prenant pour référence des coordonnées son centre de gravité. Les blocs hors diagonale sont alors nuls et on aboutit à 2 systèmes non couplés, un pour chaque composante x et y:

$$R^x_{ff}.\delta A^x = - R^x_{df} \quad ; \quad R^y_{ff}.\delta A^y = - R^y_{df}$$

L'ordre de chaque système est d = (n+1).(n+2) / 2.

Correction différentielle avec régularisation

L'analyse des systèmes précédents montre que ceux-ci risquent d'être mal conditionnés, voire non inversibles, si la région est trop peu contrastée, donc avec des valeurs quadratiques moyennes faibles pour les composantes du gradient photométrique. Il est donc intéressant d'adjoindre au critère E un terme régularisant fonction croissante de l'amplitude de la correction, de sorte que la correction soit d'autant plus faible que la zone est homogène. On transforme donc le critère E en :

$$E' = E + \mu.(\sum_{i,j} (\delta a^x_{ij})^2.\bar{f^2}_{i,j} + (\delta a^y_{ij})^2.\bar{f^2}_{i,j})$$

où les $\bar{f^2}_{ij}$ désignent les moyennes quadratiques des fonctions des coordonnées $f_{i,j}$.

Ceci a pour effet de transformer les matrices:

R^x_{ff} en $R'^x_{ff} = R^x_{ff} + \mu.I$ et R^y_{ff} en $R'^y_{ff} = R^y_{ff} + \mu.I$,

où I désigne la matrice identité de dimension d.

Les systèmes résultants sont mieux conditionnés et les corrections ne sont significatives que lorsque que le contenu photométrique de la région est suffisamment structuré. Cette méthode est aussi à rapprocher des techniques dites à base de filtrage de Wiener [5]. Notons qu'ici, ce n'est pas l'amplitude du mouvement qui est pénalisée, mais seulement celle de la correction par rapport à un modèle initial. ceci favorise donc des champs de déplacements lisses dans les zones homogènes sur le plan photométrique. Ceci est particulièrement intéressant lorsque l'on réalise la segmentation par fusion de régions adjacentes, car cela diminue les écarts artificiels entre modèles de mouvement de régions voisines.

IV. SEGMENTATION DU MOUVEMENT

On procède par une approche hiérarchique descendante. L'image est analysée selon un quadtree avec recouvrement [3] [4] et à chaque niveau de l'arbre on détermine une segmentation dont les régions sont des ensembles connexes de blocs. Nous avons trouvé que la 4-connexité était plus facile à manier et conduisait à des régions plus compactes. Cette approche présente plusieurs avantages:

- elle est bien adaptée à un contexte où on ne dispose d'aucune information a priori sur le contenu scénique car elle permet d'analyser des objets de taille variable.
- elle permet un contrôle permanent de la quantité d'information nécessaire pour la représentation du mouvement, puisque le nombre de blocs contour dépend de leur taille .
- elle est directement dérivée d'une estimation de mouvement par blocs, et est donc relativement simple à réaliser sur le plan matériel.

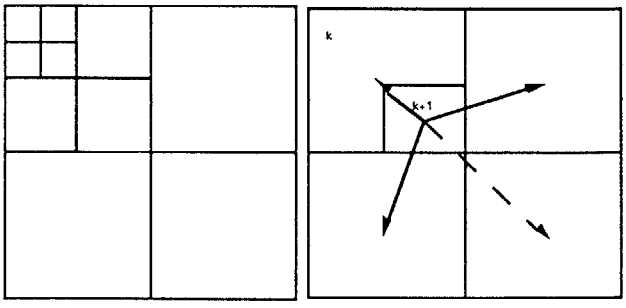

Fig.1 Analyse Quadtree et relations de recouvrement

L'évolution hiérarchique de la segmentation se fait suivant un processus de type division/fusion.

Division

Cette étape comprend 2 phases: une phase de réallocation des blocs du niveau k+1 parmi les régions du niveau k, puis création possible de nouvelles régions sur certains blocs du niveau k+1.

Chaque bloc du niveau k est divisé en 4 sous-blocs de niveau k+1, chacun de ceux-ci est alors attribué à l'une des régions contenant les blocs parents adjacents du niveau k [Fig.1]. L'attribution se fait sur critère d'erreur de prédiction minimale. Cette étape permet donc d'affiner les formes de région, et elle constitue la partie réallocation de la procédure de division.

Une seconde phase permet la céation éventuelle de nouvelles régions, ce qui est indispensable. Elle consiste à considérer chacun des blocs du niveau k+1 comme une région indépendante, et à réaliser l'estimation de leurs mouvements. Pour cela, on utilise la même stratégie multi-prédiction que dans [4]: on choisit comme initialisation les modèles des 4 blocs pères définis par le recouvrement. Pour ceux-ci, l'estimation est réalisée sur les images source pleine résolution, puisque, aux erreurs de segmentation près, les modèles initiaux sont supposés proches des mouvements présents. On y adjoint une 5ème prédiction nulle. Pour cette dernière, on utilise une estimation multirésolution [4] [6]. Contrairement à [4], la taille du bloc est réduite dans les mêmes proportions que celle de l'image pour les différents niveaux de résolution. On choisit le meilleur résultat a posteriori, puis on décide de rendre le bloc indépendant si le gain en erreur de prédiction est suffisant. En termes plus précis, soit E_r l'erreur après réallocation et E_i l'erreur sur le bloc indépendant. Une nouvelle région est créée si $E_i < \alpha.E_r$, $0<\alpha<1$ paramètre de l'algorithme.

L'étape de division déconnecte tout ou partie des régions initiales. Il convient donc de réaliser juste après cette phase un regroupement des blocs en régions connexes. De plus, la création de nouvelles régions n'a de sens que pour une taille de blocs suffisante. Elle n'est donc réalisée que pour les niveaux où la taille des blocs est supérieure à une limite fixée en paramètre (8x8 ou 16x16 par exemple). Au dessous de cette taille, on ne réalise que la réallocation. Il semble alors intéressant de réaliser celle-ci en imposant de conserver la connexité des régions initiales. Pour ce faire, on procède de la manière suivante:

- On calcule les réallocations potentielles et on stocke les diminutions d'erreur dues à la réallocation, ainsi que les étiquettes des régions candidates.
- On trie alors les blocs par ordre de variation d'erreur décroissante.
- On parcours les blocs triés, en validant ou non la réallocation selon qu'elle conduit ou non à une déconnection de la région d'origine.

Fusion

Du fait de l'arbitraire du découpage en quadtree, il est fort probable qu'une région physique donne plusieurs régions détectées. Il convient donc de fusionner les régions voisines dans la mesure où celles-ci correspondent à un même mouvement.
En fait, à cause du biais toujours présent dans les modèles estimés, il n'est pas toujours fiable de comparer les paramètres eux-même, mais il est plus intéressant de calculer l'augmentation d'erreur résultant de la fusion de 2 régions voisines. Cette approche permet un contrôle plus fin de la qualité de la prédiction versus le nombre de régions, ce qui, dans un contexte de codage à débit forcément variable dans ce cas, permet une régulation du débit consacré à l'information de mouvement.
Plusieurs stratégies sont possibles pour effectuer cette fusion. Celle proposée ici repose sur les principes suivants:

- on définit un graphe d'adjacence de régions GAR [7] pondéré et orienté [Fig.2], où le poids d'un arc R1->R2 est égal à l'accroissement de l'erreur quadratique de prédiction de R1 lorsqu'on réalise la compensation de mouvement avec le modèle de R2 au lieu de celui de R1:

$$W(R1,R2)=E(R1,A2) - E(R1,A1)$$

- on détermine ensuite l'arbre minimal recouvrant AMR [8] de ce GAR, après avoir conservé pour chaque couple de sommets l'arc de coût minimum (R1->R2 ou R2->R1).

- On réalise alors la fusion dans l'ordre imposé par l'AMR. Les coûts de fusion étant des coûts locaux, il existe un risque de divergence entre le coût calculé à partir du poids des arcs et le coût réel. On décide alors de recalculer le poids des arcs modifiés par une fusion, lorsque la région extrémité est absorbée par une autre.

- La procédure de fusion est arrêtée lorsque l'accroissement relatif de l'erreur totale atteint un seuil β fixé à l'avance. Dans un schéma de codage, une régulation de débit peut très bien piloter ce processus.

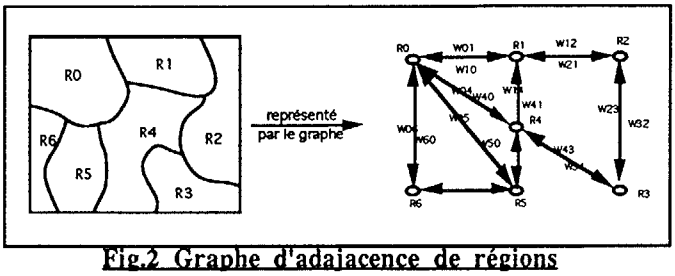

Fig.2 Graphe d'adajacence de régions

La fusion peut-être réalisée à tous les niveaux de la hiérarchie, ou seulement à certains. Nous avons trouvé que la réalisation de la fusion au dernier niveau où la création de nouvelles régions est autorisée donnait de bons résultats en terme de compromis entre le nombre de régions et l'erreur de prédiction. Cependant, ce point reste à étudier de façon plus approfondie, car c'est surtout lui qui détermine la qualité de la segmentation obtenue.

V. CODAGE DES INFORMATIONS DE MOUVEMENT

Si la segmentation du champ des vitesses permet de réduire considérablement le nombre de paramètres cinématiques, elle introduit par contre une nouvelle information à coder, constituée par les frontières de région. D'un codage efficace de ces paramètres dépendra l'efficacité de cette approche, en termes de débit mais aussi de robustesse vis-à-vis des erreurs de transmission.
En ce qui concerne les blocs contour, nous avons choisi de les définir sur une grille décalée, inter-blocs, car cette représentation lève l'ambiguïté de l'appartenance du point contour à l'une ou l'autre des régions voisines. On procède alors à partir de chaque noeud (à la confluence d'au moins 3 régions) au suivi des chaînes de contours [Fig.3], puis à leur représentation sous forme de Freeman différentielle. On peut alors soit procéder à un codage entropique de cette représentation, soit réaliser une approximation, polygonale ou Spline, de celle-ci. L'intérêt de l'une ou l'autre de ces représentations dépend de la taille des blocs contour. Pour des gros blocs, la représentation de Freeman différentielle suffit. Pour des contours de la finesse du pixel, leur nombre pouvant être important, il apparaît plus intéressant de réaliser une approximation Spline des chaînes de contours, et de se ramener ainsi à quelques points de contrôle.

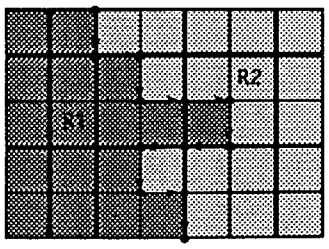

Fig.3 Suivi de contours

En ce qui concerne les paramètres cinématiques, on a choisi de procéder à un changement de représentation, car si la quantification d'un vecteur déplacement est significative en terme de précision de reconstruction, celle de paramètres de déformation ou de rotation l'est moins. On remplace donc par transformation inversible le jeu de paramètres du modèle polynomial par un jeu

de vecteurs déplacement en des points correctement choisis. Un exemple de ces points est donné [Fig.4]. On quantifie alors les vecteurs obtenus avec la précision sub-pixel et la dynamique désirée.

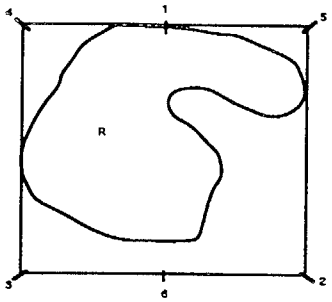

Fig.4 Choix des points de référence pour le codage d'un modèle quadratique

VI. RÉSULTATS EXPÉRIMENTAUX

Cet algorithme a été testé sur séquences d'images naturelles (Mobile & Calendrier et Train) afin d'évaluer ses performances en termes d'erreur de prédiction, de nombre de régions obtenues (donc de points contour engendrés) ainsi que de la signification physique des régions et mouvements déterminés.
Les résultats ci-après ont été obtenus dans les conditions suivantes:
- α = 80% pour la création de nouvelles régions, permise pour une taille de bloc supérieure ou égale à 16x16, β = 20% d'augmentation de l'erreur globale lors de la phase de fusion, modèles affines de mouvement et μ = 0 (pas de régularisation).

En ce qui concerne l'erreur de prédiction, les résultats se situent environ 30 à 50% en dessous de ceux obtenus avec un algorithme de mise en correspondance de blocs, au 1/2 pixel, sur des blocs 8x8, avec une excursion de ± 16 pixels [Fig. 5]. La comparaison est donc très favorable pour la méthode proposée.

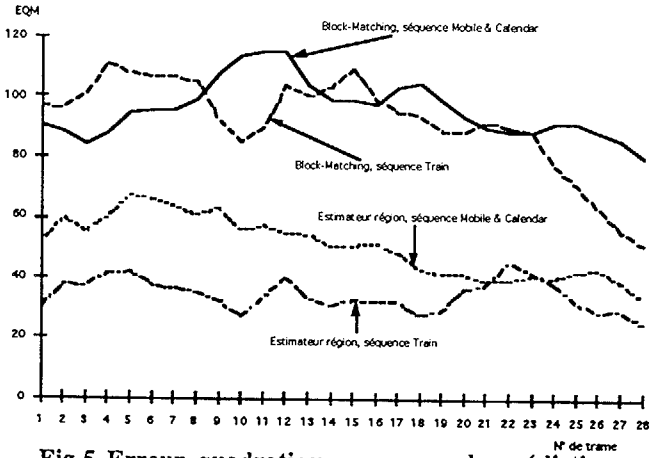

Fig.5 Erreur quadratique moyenne de prédiction

En ce qui concerne le nombre de régions, il est bien sûr variable et fonction de la séquence traitée. Il se situe néanmoins dans un domaine raisonnable, de 10 à 12 pour la séquence Mobile & Calendrier et de 20 à 45 pour la séquence Train.
Globalement, les frontières du mouvement sont assez bien déterminées, avec cependant une sur-segmentation d'autant plus marquée que les zones d'occlusion sont importantes, ce qui est le cas de la séquence Train, où 2 trains se croisent en mouvement rapide (60 pels entre 2 trames consécutives de même parité). Ceci plaide en faveur d'une prise en compte des zones d'occlusion dans le processus de segmentation. Dans les zones prédictibles, le mouvement est généralement bien estimé, y compris dans le cas de zooms ou de rotations.

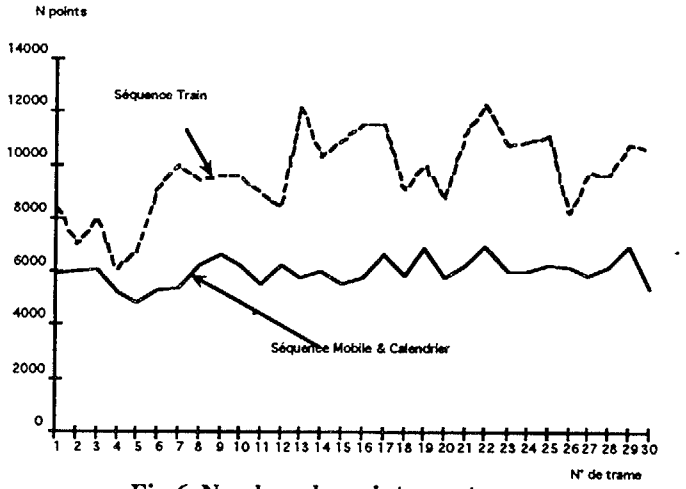

Fig.6 Nombre de points contour

Le nombre élevé de points contour obtenu pour la séquence Train [Fig. 6] montre qu'une représentation efficace de ceux-ci est indispensable pour atteindre une réduction de débit substantielle sur les données de mouvement. L'approximation Spline des chaînes de contours apparaît comme une solution très intéressante.

VII. CONCLUSION

Nous avons présenté une méthode pour la réalisation conjointe de l'estimation et de la segmentation du mouvement dans des séquences d'images. Cette méthode, plus spécifiquement développée pour des applications de codage semble prometteuse du fait qu'elle conduit à des rapports entre l'erreur de prédiction et le débit d'information de mouvement très favorables. De plus, les mouvements calculés sont qualitativement et quantativement cohérents avec les mouvements apparents réels et les masques de région obtenus sont souvent en bon accord avec les zones mobiles réelles, même si certaines images sont sur-segmentées. De ce fait, l'approche proposée ici peut être envisagée pour d'autres applications, après une éventuelle adaptation.

RÉFÉRENCES BIBLIOGRAPHIQUES

[1] D. SHEN, S.A. RAJALA, "Segmentation-based motion estimation and residual coding for packet video: a goal-oriented approach", SPIE VCIP'92, Vol. 1818, Part 3, Nov. 18-20, Boston, USA, pp 825-836.

[2] H .G. MUSMANN, M. HÖTTER and J. OSTERMANN, "Object-oriented analysis-synthesis coding of moving images", Image Com. Vol. 1, No. 2, October 1989 pp 117-138.

[3] J.L. DUGELAY,"Toward an estimation of three-dimensional motion in a 3DTV image sequence", SPIE VCIP'91, Vol. 1605, Part 2, Nov. 11-13, Boston, USA, pp 688-693.

[4] H. SANSON, "Motion Affine Models Identification and Application to Television Image Coding", SPIE VCIP'91, 11-13 November 1991, Boston, Massachusetts, Vol. 1605, Part 2, pp 570 - 581.

[5] J. BIEMOND, J.N. DRIESSEN, A.M. GEURTZ and D.E. BOEKEE, "A pel-recursive Wiener-based algorithm for the simultaneous estimation of rotation and translation", SPIE VCIP'88, Part 3, Nov. 1988, Cambridge MA.

[6] P.J. BURT, "The pyramid as a structure for efficient computation", in: Multiresolution Image Processing and Analysis, ed. Rosenfeld, Springer Verlag, 1984, pp. 6-35.

[7] M. KOCHER, " Adaptive region growing technique using polynomial functions for image approximation", Signal Processing N° 11, 1986, pp 47 - 60.

[8] O.J. MORRIS, M.de J. LEE, A.G. CONSTANTIDINES, "Graph theory for image analysis, an approach based on the shortest spanning tree", IEE Proceedings, Vol. 133, Pt. F, N°2, April 1986, pp 146-152