Codeur H.263 amélioré pour la visiophonie mobile

S. Roux[1], E. Petit[2]
[2] eric.petit@rd.francetelecom.com

[1] TIMA
46,avenue Felix Viallet
38031 GRENOBLE

[2] France Telecom R&D, DIH/OCF
28 chemin du vieux Chêne
BP 98, 38243 Meylan cedex.

Résumé

Du fait des capacités actuelles d'intégration et des nouvelles normes de télécommunication mobile (GPRS, EDGE, UMTS,...), le codage vidéo temps réel sur mobile est aujourd'hui envisageable. Cependant, les contraintes sont sévères tant en termes de complexité, de débit, sans oublier de qualité de service. Nous proposons ici un système de codage visiophonique basé H.263 tirant profit de l'information de position du visage du locuteur pour réduire la complexité du codeur et améliorer la qualité de la vidéocommunication ainsi que l'ergonomie du terminal. De plus, ce codage utilise une technique de détection du visage originale basé sur une classification adaptative des pixels.

Mots clefs

Visiophonie mobile, Algorithme adaptatif de détection, Codeur H.263 orienté visage.

1 Introduction

Avec les débits promis par les futures normes de télécommunication mobile, le GPRS et bientôt l'UMTS[1], des services incluant de la vidéo sont envisageables. Ces dernières années ont vues fleurir différents concepts de terminaux de troisième génération plus ou moins réalistes, dont la plupart incluent de la vidéo et plus spécialement un service de visiophonie. Les premières réalisations issue de Kyocera, NEC, Toshiba [2],[3],[4] et bien d'autres montrent la faisabilité des systèmes de codage vidéo pour les applications mobiles mais mettent également en avant les contraintes du terminal mobile. Ces contraintes peuvent être regroupées dans trois catégories. La première, imposée par le mode de transmission, est le faible débit disponible mais également le taux d'erreurs (de qq dizaines de kbit/s jusqu'à qq centaines de kbit/s pour les premiers déploiements, avec un taux d'erreur de 10^{-3}). La deuxième contrainte, liée elle aussi à la mobilité, est celle de l'autonomie du terminal. La troisième, essentielle pour l'expansion des services à base de vidéo sur les réseaux mobiles est la qualité de service fournie. Cette dernière, englobe bien entendu, le rendu de la vidéo, sa fluidité, mais également l'ergonomie du service. De plus, la visiophonie peut améliorer l'intelligibilité de la parole pourvu que la qualité d'image soit maintenue sur le visage, et que la fréquence image soit au moins de 15 Hz. En effet, cette fréquence garantit une bonne fluidité et une bonne synchronisation de la voix et des lèvres. Par ailleurs, le cadrage du visage de l'utilisateur dans le champ de la caméra n'est pas trivial dans un contexte de téléphonie mobile.

Les systèmes de codage vidéo pour les applications embarquées qui ont vu le jour ces dernières années sont généralement basés sur les normes vidéo MPEG-4 [5] ou H.263 [6]. MPEG-4, dernière née des normes de codage vidéo de l'ISO/IEC introduit la notion de codage par objets vidéo. Cependant, l'extraction d'objets vidéo reste un point dure et de complexité incompatible avec les contraintes de l'embarqué. C'est pourquoi, MPEG-4 spécifie, pour les applications mobiles très bas débit, un profil de base, le profil Simple, similaire au codage H.263 (avec des outils de résistance aux erreurs supplémentaires). Cependant, un algorithme basé bloc n'ayant pas de stratégie "d'allocation de bits" orientée objet, le risque est d'obtenir une qualité de vidéo fortement dégradée sur le visage. Tout en restant dans le cadre MPEG-4 SP ou H. 263, nous proposons de mettre en oeuvre une solution prenant en compte la nature de l'image par un codage orientées visage. Ce codeur amélioré respecte les contraintes de la visiophonie mobile : débit, complexité, qualité, ergonomie.

Après une description des principes de codage orienté objet d'intérêt retenus, nous introduirons notre algorithme adaptatif de détection du visage effectuant une partition de l'espace des chrominances en classes. Nous présenteront, ensuite, deux techniques de contrôle du codage, la première influençant la stratégie d'estimation de mouvement, la seconde permettant une meilleure allocation de débit. Nous terminerons sur l'ergonomie du terminal et enfin nous conclurons.

2 Principes

Le système proposé consiste en un codeur H.263 auquel on a adjoint un pré-traitement permettant de détecter un visage dans la scène et de contrôler les stratégies de codage. Le codeur exploite l'information de position du visage dans ses stratégies d'estimation de mouvement (EMOV) et d'allocation de bits (CDOV). La figure 1 présente un diagramme fonctionnel du codeur H.263 hybride

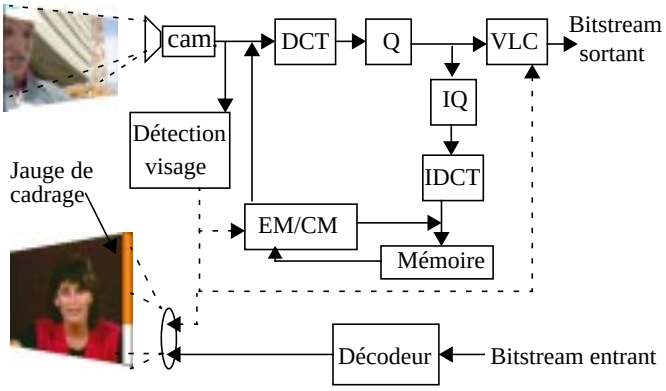


Figure 1. Diagramme fonctionnel du codec H.263 amélioré

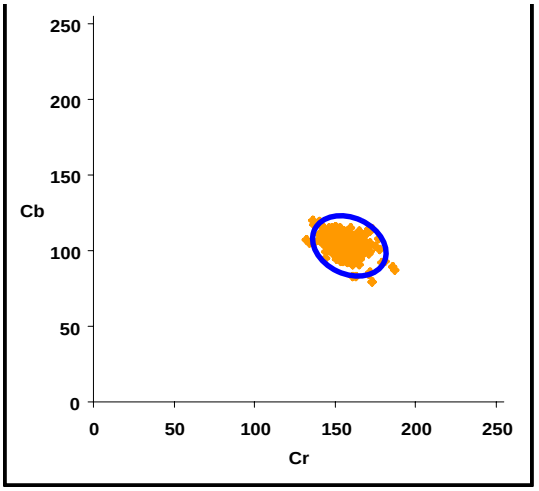


Figure 2: la couleur peau dans l'espace des chrominances

3 Détection du visage

De nombreux algorithmes de détection/segmentation et suivis de visage utilisés ou non dans des systèmes de codage vidéo existent dans la littérature [7][8][9][10], cependant ils sont pour la plupart coûteux ou nécessitent une phase d'apprentissage longue et complexe ou encore ne s'appliquent qu'à des scènes où l'environnement est connu (visiophonie avec arrière plan supposé sans mouvement,...). Dans le cas de la visiophonie mobile, l'arrière plan n'est pas connu a priori et peut contenir beaucoup de mouvement. De plus, la détection de visage doit être de faible complexité et robuste dans le sens où tout visage présent doit être détecté. Ainsi, notre choix s'est porté sur un algorithme adaptatif qui exploite l'hypothèse que la texture de la peau est localisée dans une région étroite de l'espace des chrominances. La figure 2 montre la répartition de la couleur peau dans l'espace des chrominances pour une centaine de visages de groupes ethniques différents (européens, asiatiques, africains,...).

La détection du visage est réalisée en deux étapes (figure 5):

- estimation du masque binaire des régions "peau",
- localisation et estimation du cadre circonscrit au visage.

La première étape est basée sur une partition de l'espace des chrominances en 4 classes dont une correspond à la couleur peau. A chaque classe est affecté un représentant qui est égal à la valeur moyenne des pixels la constituant. L'algorithme itératif de Lloyd-Max [11] fournit les représentants minimisant l'erreur quadratique moyenne. Cet algorithme, également appelé K-mean [12], se déroule comme suit:

1. initialisation des 4 représentants,
2. règle du plus proche voisin vis à vis des 4 représentants pour tous les pixels et ré-estimation des représentants des classes à l'aide de l'estimateur du centroïde,
3. itération de l'étape 2 jusqu'à convergence des classes.

Les représentants sont initialisés en référence à la figure 2, de manière à ce que la partition mette en avant les régions de couleur peau.

Nous proposons de rendre l'algorithme de Lloyd-Max adaptatif au cours du temps. De plus, plutôt que de réaliser une convergence intra-image (pour laquelle le nombre d'itérations n'est pas connue d'avance), nous préconisons une convergence inter-image, exploitant la corrélation temporelle.

Pour lever le doute sur le représentant de la classe "peau", le système effectue des comparaisons entre les représentants et le représentant de référence de la texture peau.

Notre approche converge rapidement (en moyenne au bout de 3 images, figure 3), et réduit considérablement la complexité de la détection car une seule itération est réalisée par image.

image 1 image 2 image 3

séquence Foreman

image 1 image 2 image 3 image 4

séquence Akiyo

Figure 3: convergence de l'algorithme adaptatif

Le masque binaire issu de ce traitement est ensuite sous-échantillonné et filtré afin d'obtenir une information pertinente pour l'étape de localisation. Cette étape est effectuée par un filtrage adapté utilisant des fenêtres rectangulaires approchant une forme de visage. Le résultat est un rectangle circonscrit au visage du locuteur (de taille multiple de 16*16 pixels conformément aux stratégies de codage H.263, figure 4)

.

.

Figure 4. résultat de la localisation (Foreman et Akiyo)

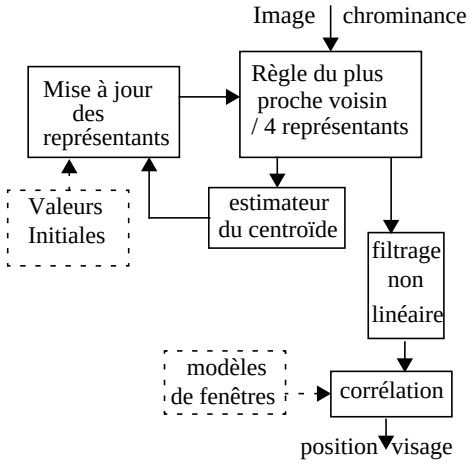


Figure 5. Détection du visage

4 Les stratégies de codage

Elles se situent sur deux plans.

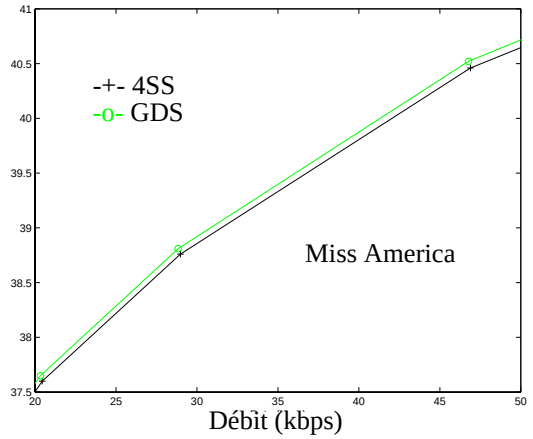
Choix d'un estimateur de mouvement (EM) adapté à la nature du mouvement

La faible complexité requise par l'application ciblée nécessite l'utilisation d'une technique d'EM rapide telle que la recherche selon le gradient (GDS [13]) ou la recherche en quatre pas (4SS [14]). Cependant les performances des algorithmes de recherche rapides dépendent de la nature du mouvement et donc de l'objet à coder. Par exemple, la 4SS est plus efficace que la GDS pour des scènes où les mouvements sont importants et non corrélés, et réciproquement, la GDS est plus efficace pour les séquences où les mouvements inter MBs sont fortement corrélés telle que la visiophonie (figure 5).

Ainsi la stratégie préconisée est de sélectionner l'EM en fonction de l'appartenance ou non du MB au visage, afin d'obtenir un bon compromis entre complexité et qualité des estimations (on utilisera, par exemple, une recherche exhaustive pour les MBs du visage et un EM rapide pour l'arrière plan).

Régulation du débit

La connaissance de la position du visage permet d'avoir une stratégie d'allocation de bits liée à la nature de l'objet à coder comme dans [7]. Cette information permet également de faire varier la fréquence image pour obtenir une qualité subjective plus élevée sur la région d'intérêt. Une stratégie efficace est de coder l'arrière plan à la fréquence image moitié de celle du visage ce qui permet d'allouer plus de bits au codage du visage. De plus, cette stratégie conduit à une plus faible complexité (cf. tableau).



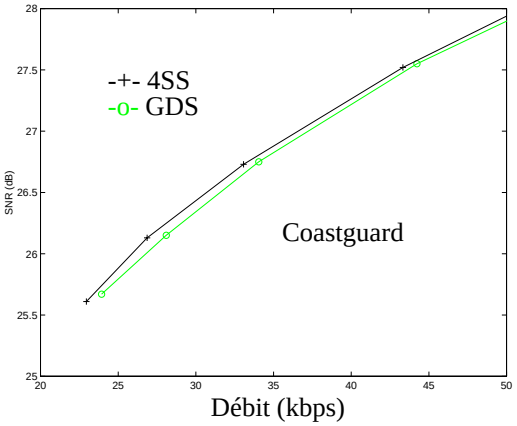


Figure 5. Influence de la technique d'EM

Codeur H.263*		FS	4SS	GDS	amélioré
Complexité ** (MIPS)	Arith.	875	156	153	**112**
	autres	766	216	209	**166**
	Load/ Store	733	116	113	**85**
	Total	2374	488	475	**363**
SNR visage moyen (dB)	Y	29.94	29.87	29.86	**32.37**
	Cr	35.14	35.17	35.2	**37.22**
	Cb	35	34.97	35.05	**37.74**

(*) Séquence Carphone, 15 Hz, 40 kbits/s, VM TMN8
(**) Complexités en MIPS SPARC fournies par l'utilitaire iprof (compilation gcc options mv8 et O3) [15].

Notre approche présente une complexité inférieure de 85% par rapport à une approche utilisant une recherche exhaustive des vecteurs mouvement et inférieure de 25% par rapport à un codage utilisant la technique de recherche rapide 4SS ou GDS. De plus, la qualité subjective est grandement améliorée (figure 6) et les gains en rapport signal à bruit sur le visage sont significatifs : + 2dB en luminance et Chrominances pour la séquence "Carphone" (figure 7).

Codeur H.263, recherche exhaustive

Codeur H.263 orienté visage (EMOV+CDOV)

Figure 7: Comparaison de la qualité subjective
Image 145 de la séquence Carphone codée à 40 kbps et 15 Hz

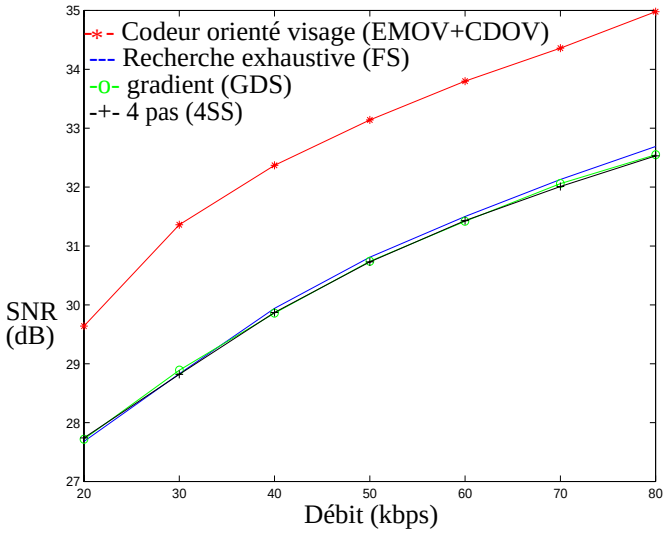


Figure 6.débit vs distortion (séquence Carphone)

5 Assistance au cadrage

Enfin, la détection du visage du locuteur permet par un moyen simple d'informer l'utilisateur de sa position relative par rapport à la caméra. L'utilisation d'une jauge en bordure d'image indiquant le degré de cadrage peut remplacer l'icone de sa propre image (Cf. figure 1).

6 Conclusion

Notre système de codage basé H.263 orienté visage offre un bon compromis entre complexité et qualité. Il met en oeuvre un algorithme de détection du visage adaptatif, simple et robuste, contrôlant les stratégies d'EM et de régulation du débit. Cette détection utilise 1 million d'instructions par image sur machine Sparc. Pour autant, ce traitement supplémentaire n'entraîne pas d'augmentation de complexité, mais au contraire conduit à un coût CPU global plus faible. Ces travaux constituent un premier pas vers une implémentation réaliste d'un codeur MPEG-4 orienté objet pour la visiophonie mobile.

Références

[1] http://www.umts-forum.org/

[2] VisualPhone VP210: http://www.kyocera.co.jp/frame/product/telecom/english/vp210_e/

[3] W-CDMA based mobile Videophone: http://www.nec.co.jp/english/today/newsrel/9909/2201.html

[4] MPEG-4 video codec: http://www.toshiba.com/taec/cgi-bin/display.cgi?table=Family&Family-ID=22

[5] http://cselt.it/mpeg/

[6] http://www.itu.int/home/

[7] H. Luo, A. Eleftheriadis, J. Kouloheris, 'Statistical model-based video segmentation and its application to very low bit-rate video coding', *Signal processing: Image Communication*, Vol. 16, pp 333-352, 2000.

[8] J-C. Terillen, S. Akamatsu, 'Comparative performance of different chrominance spaces for color segmentation and detection of human faces in complex scene images' ,*Vision Interface'99*. pp180-187.

[9] P. Salembier, F. Marqués, M. Pardàs, J.R. Morros, I. Corset, S. Jeannin, L. Bouchard, F. Meyer, B. Marcotegui, 'Segmentation-based video coding system allowing the manipulation of objects', *IEEE Trans. on circuits and syst. for video tech.*, Vol. 7, No. 1, Feb. 1997.

[10] J. Yang, W. Lu, A. Waibel, 'skin-color modeling and adaptation', *ACCV'98*. http://www.is.cs.cmu.edu/ISL.publications.html.

[11] N. Moreau, *'Techniques de compression des signaux'*, MASSON 1995.

[12] C. G. Looney, *'Pattern recognition using neural networks',Theory and algorithms for engineers and scientists*, Oxford University Press, 1997.

[13] L.-K. Liu et E. Feig, 'A block-based gradient descent search algorithm for block motion estimation in video coding', *IEEE Transaction on Circuits and Systems for Video Technology*, Vol. 6, No 4, pp. 419-422, 1996

[14] L-M. Po, W-C. Ma, 'A novel four-step search algorithm for fast block motion estimation', *IEEE trans. Circuits and System for Video Tech.*, Vol. 6, no 3, pp 313-317, 1996.

[15] http://www.lis.e-technik.tu-muenchen.de/people//