Classification Sémantique des Macro-Blocs Mpeg dans le Domaine Compressé.

F. Souvannavong B. Merialdo B. Huet

Département Multimédia
Institut Eurécom
2229 route des Crêtes, 06904 Sophia-Antipolis FRANCE

{souvanna, merialdo, huet}@eurecom.fr

Résumé

Ce papier présente les premiers résultats obtenus par notre méthode de classification sémantique des macro-blocs mpeg. Ces travaux sont motivés par la récente introduction dans la série de conférences TREC d'une tâche dont l'objectif est de détecter les sémantiques présentes dans des plans vidéos. Pour atteindre ce but, nous proposons une méthode de classification qui opère directement dans le domaine compressé sur l'information fournie par les macro-blocs. Conscient des limitations imposées sur la précision du traitement des images, cette approche permet d'éviter l'étape fastidieuse de la décompression du flux Mpeg. Dans ce cadre, notre objectif est d'évaluer la classification sémantique obtenue à partir d'une approche directe qui sera le point de départ de méthodes plus évoluées.

Mots clefs

Classification sémantique, Mixture de Gaussiennes, Transformée DCT.

1 Introduction

L'importante quantité d'information visuelle, fournie aussi bien par les vidéos que les images fixes, nécessite des outils fiables et efficaces d'indexation et de navigation [1, 2]. L'institut américain des standards et technologies (NIST) finance une série de conférences, TREC vidéo 2002 [1], pour promouvoir le développement des méthodes de recherche par le contenu dans des bases de données vidéo. Notre travail se place dans ce contexte, et nos recherches se focalisent sur l'extraction de caractéristiques sémantiques ; les plans d'une vidéo doivent être classifiés selon les catégories suivantes : *scène intérieure, scène extérieure, paysage urbain, paysage rural, texte, visage* ou *groupe de personnes*.

Dans un premier temps le contenu doit être analysé, afin d'extraire des caractéristiques pertinentes. Cette opération est particulièrement fastidieuse, surtout lorsqu'une base de données complète doit être traitée. L'analyse fiable de l'information dans le domaine compressé se révèle donc fort intéressante. De nombreux travaux ont été conduits dans le domaine de la segmentation de l'image et de la vidéo, cependant peu de chercheurs ont étudié la tâche particulière de la segmentation en objets ou régions *sans* décodage complet [3, 4, 5, 6].

Dans ce papier, nous proposons d'extraire les caractéristiques sémantiques d'une séquence vidéo en classifiant les macro-blocs DCT de taille 16 par 16 pixels composant une image. Nous avons dissocié, dans l'ensemble proposé par TREC, la sémantique étendue présente au niveau de l'image comme *scène intérieure, scène extérieure, paysage urbain, paysage rural, groupe de personnes* de la sémantique élémentaire présente au niveau de la région comme *texte, visage*. Nous allons étudier la capacité des macro-blocs à fournir une information sémantique fiable, en particulier nous verrons comment de nouveaux concepts élémentaires comme *verdure, ciel, eau, bâtiment* seront introduits pour décrire correctement la sémantique étendue.

La prochaine section détaille le procédé de classification retenu. Ensuite nous donnerons les résultats obtenus par la classification sur les deux types de sémantiques. Enfin, nous conclurons sur une critique des résultats et les futurs travaux dans la dernière section.

2 La Classification des Macro-blocs

Trois étapes majeures interviennent dans les procédés de classification. D'une part l'extraction des vecteurs décrivant l'information, ensuite la modélisation des classes avec l'estimation des paramètres introduits et enfin la classification selon des règles de décision appropriées. Dans notre approche, nous étudions uniquement les images intra-codées dont les macro-blocs sont représentés par leurs 384 coefficients DCT (format d'image 4 :2 :0) mis sous la forme d'un vecteur tel que les basses fréquences soient dans les premiers éléments et les hautes fréquence dans les derniers. Notons que les premiers coefficients DCT sont alors les plus importants (moindre sensibilité visuelle, robustesse au bruit) et la dimension du vecteur caractéristique peut alors être simplement réduite par troncation.

Nous supposons que la distribution des macro-blocs au sein

[1] TREC est une série de conférences destinées à promouvoir le développement des techniques de recherche de l'information à partir de grandes quantités de données.
Voir http ://www-nlpir.nist.gov/projects/t2002v/t2002v.html

d'une classe peut être décrite par un modèle de mixture et en particulier des mixtures de Gaussiennes [7, 8, 9, 10]. Les Gaussiennes permettent de capturer les traits principaux des macro-blocs, mais également leur évolution due au mouvement aussi bien qu'au changement de luminosité. De plus dans [11], E.Y. Lam and J.W. Goodman ont prouvé que la distribution des coefficients DCT peut être correctement approximée par une Gaussienne *lorsque la variance est constante* ; dans notre situation, l'utilisation de mixtures permet de s'écarter légèrement de cette hypothèse. La densitée de probabilité conditionnelle d'un vecteur X sachant Φ_i peut donc être formulée comme suit :

$$\text{Pour } X \in C_i, P(X \mid \Phi_i) = \sum_j \alpha_j p_j(X)$$

$$\text{où } \alpha_i \in \Re, \Phi_i = (\mu_j, \sigma_j) \text{ et } p_j(X) \sim \mathcal{N}(\mu_j, \sigma_j)$$

Les paramètres des GMM α_j, μ_j et σ_j sont estimés en utilisant l'algorithme classique Expectation-Maximization [12] qui est initialisé par un algorithme de type k-means. Malheureusement le rang de la matrice de l'ensemble d'entrainement est moindre que la dimension de l'espace, aboutissant à une matrice de covariance σ_i non définie positive. Pour éviter de rencontrer cette situation, un bruit blanc est ajouté aux données. Nous considérons également dans les expériences que les composantes des vecteurs caractéristiques sont indépendantes, ainsi σ_i est une matrice diagonale estimée avec plus de fiabilité. Finalement, le choix du nombre de mixtures est simplement obtenu en analysant l'évolution de la vraisemblance de l'ensemble de test durant la phase d'apprentissage pour plusieurs nombres de mixtures.
Étant donné un macro-bloc, le maximum a posteriori

$$\hat{C} = arg \max_i P(\Phi_i \mid X)$$

nous donne une estimation de la classe auquel il appartient. La probabilité a posteriori peut être développée de la manière suivante en utilisant l'égalité de Bayes :

$$P(\Phi_i \mid X) = \frac{P(X \mid \Phi_i)P(\Phi_i)}{P(X)}$$

finalement,

$$P(\Phi_i \mid X) \propto P(X \mid \Phi_i)$$

puisque nous supposons les classes et les vecteurs distribués *uniformément*.
Toutefois, l'ensemble des classes retenues ne forme pas une partition et certains vecteurs peuvent n'appartenir à aucune classe. Pour cela un seuil minimum $-sm_i$ est fixé pour le log de vraisemblance de chaque classe i. Il est choisi sur l'ensemble d'entraînement de tel façon que seuls 10% des vecteurs soient non classifiés. Le seuil choisi de manière heuristique fait l'objet d'un compromis entre précision et performance. Pour finir, la règle de décision s'écrit :

$$\hat{C} = arg \max_i \{P(X \mid \Phi_i) \mid -\log(P(X \mid \Phi_i)) \leq mb_i\}$$

3 Expériences et Résultats

L'ensemble des expériences qui suivent, est destiné à étudier l'adéquation des macro-blocs de taille 16x16 pixels à fournir une information sémantique à deux différents niveaux. Pour cela les concepts de vidéo TREC ont été retenus : *scène intérieure, scène extérieure, paysage urbain, paysage rural, texte, visage* ou *groupe de personnes*.
Dans un premier temps, l'étude sera menée sur des concepts à l'échelle de l'image ; ensuite nous étudierons les concepts élémentaires, c'est à dire pertinent à l'échelle du macro-bloc. Dans ce dernier cas nous introduirons des concepts supplémentaires qui permettront d'induire ceux à l'échelle de l'image. Les expériences ont été conduites sur les 2000 premières images intra-codées du documentaire "Histoire d'Eau" [2] que nous avons manuellement annotées pour évaluer notre méthode.

Modèles	Visage	Personnes	Ext.	Int.	Autre
Visage	64.53	15.20	4.38	14.04	1.87
Visage	64.01	15.74	4.44	14.09	1.72
Gens	28.25	46.03	7.69	16.63	1.40
Gens	28.57	44.07	8.04	17.50	1.81
Ext.	18.95	12.05	52.41	14.31	2.28
Ext.	18.72	12.70	52.36	13.91	2.31
Int.	30.73	26.08	10.56	31.24	1.39
Int.	31.95	25.89	10.51	30.09	1.56

Tableau 1 – *Taux de reconnaissance au niveau de l'image.* Les colonnes correspondent aux classes et les lignes aux ensembles d'entraînement et de test.

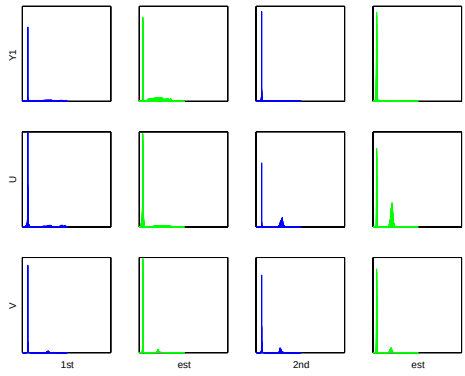
Modèles	Pays. rural	Pays. urbain	Autre Ext.	Autre
Pays. rural	72.78	13.80	11.67	1.75
Pays. rural	71.95	13.87	12.58	1.60
Pays. urbain	10.83	81.49	7.37	0.31
Pays. urbain	10.99	81.16	7.51	0.35

Tableau 2 – *Taux de reconnaissance au niveau de l'image.* Les colonnes correspondent aux classes et les lignes aux ensembles d'entraînement et de test.

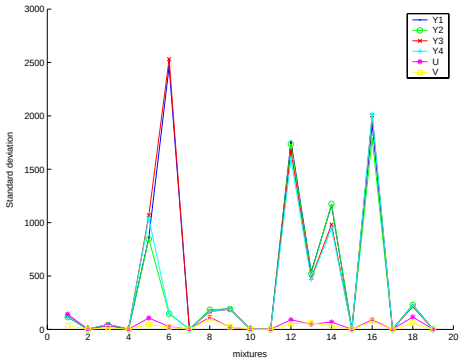
Nous supposons que les images formant un plan d'une séquence vidéo, sont qualifiées par un unique concept et que les macro-blocs correspondant contiennent cette information. Une première classification permet d'établir la distinction entre *scène intérieure, scène extérieure, groupe de personnes* ou *visage (portrait)*, ensuite les macro-blocs composants les *scènes extérieures* sont étudiées plus précisément et séparées dans les catégories *paysage urbain, paysage rural* ou *autre*. Les tableaux 1 et 2 fournissent les résultats obtenus et comme attendu ils sont mitigés. Même si les mixtures de Gaussiennes approximent correctement la distribution des macro-blocs d'une classe comme illustré par la figure 1, les variances associées aux

[2] Document issu des archives de l'Institut National de l'Audiovisuel

coefficients DC sont très élevées. La répartition des macro-blocs est donc très hétérogène et non adaptée à la classification.



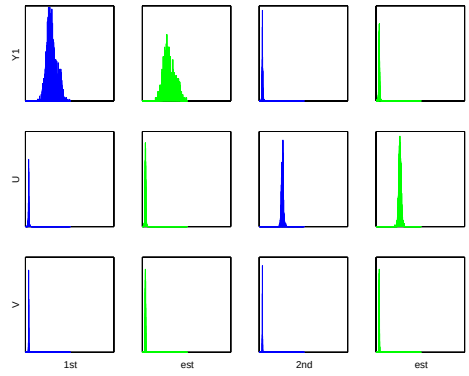
(a) Comparaison de la distribution des premiers coefficients DCT.



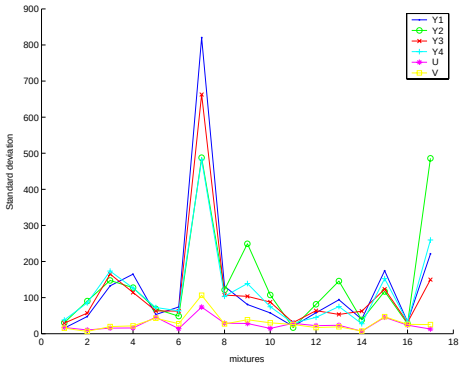
(b) Variance des coefficients DCT pour chaque mixture.

Figure 1 – *Estimation des paramètres de la classe scène extérieure*

Puisque les macro-blocs ne peuvent pas convenablement exprimer un concept au niveau de l'image, l'idée naturelle fut de décrire ces derniers par de nouveaux concepts en formant une hiérarchie simple. Les concepts *eau, bâtiment, verdure, forêt* et *ciel* ont donc été introduits pour caractériser les *paysages ruraux* et *urbains*. Le tableau 3 illustre les résultats de cette nouvelle classification.
Les résultats sont bien plus satisfaisants avec une grande partie des classes atteignant un taux de reconnaissance de 90%. Les mixtures de Gaussiennes sont donc bien adaptées à ce problème, comme le montre la comparaison des distributions réelles et estimées ainsi que quelques valeurs de la matrice de covariance dans la figure 2.
Nous avons introduit de manière heuristique des concepts élémentaires qui sont identifiables dans un macro-bloc. Mais comme le montre le tableau 4, les relations entre ces concepts et ceux de plus haut niveau ne sont pas triviales [13].



(a) Comparaison de la distribution des premiers coefficients DCT.



(b) Variance des coefficients DCT pour chaque mixture.

Figure 2 – *Estimation des paramètres de la classe visages*

Modèles	Eau	Ciel	Bâtiment	Verdure	Forêt	Autre
Scène Ext.	16.28	6.39	25.58	1.35	6.40	44.01
Scène Int.	27.07	2.47	43.41	5.00	6.83	15.22

Modèles	Eau	Ciel	Bâtiment	Verdure	Forêt	Autre
Pays. urbain	17.26	12.53	48.05	0.24	11.50	10.44
Pays. rurale	25.52	16.93	21.85	9.20	18.00	8.49

Tableau 4 – *Composition des images.* Les lignes correspondent aux classes et les colonnes aux ensembles d'entraînement et de test.

4 Conclusion

Les résultats présentés montrent la difficulté d'obtenir une classification directe au niveau de l'image avec le vecteur caractéristique décrit. Toutefois une classification prometteuse a été obtenue au niveau de la région. Les expériences devront être approfondies, notamment en explorant d'autres concepts sur différentes séquences. Nous projetons également de construire automatiquement la hiérarchie, cela permettra de réduire considérablement la tâche d'annotation manuelle particulièrement fastidieuse, tout en se plaçant dans un cadre probabilistique complet.

Modèles	Visage	Texte	Eau	Ciel	Bâtiment	Verdure	Forêt	Autre
Visage	95.60	1.00	0.13	0.00	1.67	1.20	0.40	0.00
Visage	95.07	1.07	0.40	0.00	1.80	1.40	0.27	0.00
Texte	3.18	90.78	1.24	0.08	3.18	0.00	0.77	0.77
Texte	1.94	90.08	1.63	0.00	3.95	0.00	0.54	1.86
Eau	0.20	0.07	94.87	1.53	0.67	0.00	2.67	0.00
Eau	0.53	0.00	94.60	1.53	0.73	0.00	2.60	0.00
Ciel	0.00	0.07	0.53	99.40	0.00	0.00	0.00	0.00
Ciel	0.00	0.40	0.67	98.93	0.00	0.00	0.00	0.00
Bâtiment	6.07	1.40	12.07	0.27	75.18	0.07	4.87	0.07
Bâtiment	7.81	2.74	13.34	0.20	68.58	0.00	7.27	0.07
Verdure	0.50	0.00	0.00	0.00	0.00	99.50	0.00	0.00
Verdure	1.49	0.00	0.00	0.00	0.00	98.51	0.00	0.00
Forêt	6.40	0.58	5.81	0.00	7.27	0.58	79.36	0.00
Forêt	7.56	1.45	9.01	0.00	8.14	1.16	72.67	0.00

Tableau 3 – *Taux de reconnaissance au niveau de la région.* Les colonnes correspondent aux classes et les lignes aux ensembles d'entraînement et de test.

Concepts	Nb de Mix.	Taille entraînement	Taille test
Visage	10	1500	1500
Texte	14	1250	1250
Eau	14	150	1500
Ciel	12	1500	1500
Bâtiment	14	1500	1500
Verdure	8	600	600
Forêt	6	350	350
Viage (portrait)	20	10000	10000
Gens	20	9500	9500
Scne extèrieurs	20	10000	10000
Scène intèrieure	20	10000	10000
Paysage urbain	20	6000	6000
Paysage rural	20	9500	9500

Tableau 5 – *Conditions de classification.*

Ce travail a été réalisé dans le cadre du projet européen SPATION.

Références

[1] Shih-Fu Chang, W. Chen, H.J. Meng, H. Sundaram, et Di Zhong. A fully automated content-based video search engine supporting spatiotemporal queries. Dans *IEEE Transactions on Circuits and Systems for Video Technology*, volume 8, pages 602– 615, 1998.

[2] A. Pentland, R. Picard, et S. Sclaroff. Photobook : Content-based manipulation of image databases. Dans *SPIE Storage and Retrieval for Image and Video Databases*, february 1994.

[3] Hualu Wang et Shih-Fu Chang. A highly efficient system for automatic face region detection in mpeg video. Dans *IEEE Transactions on Circuits and Systems for Video Technology*, volume 7, pages 615–628, August 1997.

[4] O. Sukmarg et K. Rao. Fast object detection and segmentation in mpeg compressed domain. TENCON, 2000.

[5] Yu Zhong, Hongjiang Zhang, et Anil K. Jain. Automatic caption localization in compressed video. Dans *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 22, pages 385–392, April 2000.

[6] A. Girgensohn et J. Foote. Video classification using transform coefficients. Dans *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3045–3048, 1999.

[7] J. Verbeek, N. Vlassis, et B. Kr. Greedy gaussian mixture learning for texture segmentation. Dans *Workshop on Kernel and Subspace Methods for Computer Vision*, 2001.

[8] Michael E. Tipping et Christopher M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2) :443–482, 1999.

[9] Rui Zhang et Xiaoqing Ding. Offline handwritten numeral recognition using orthogonal gaussian mixture model. Dans *International Conference on Image Processing*, volume 1, pages 1126–1129, 2001.

[10] M. Saeed, W.C. Karl, T.Q. Nguyen, et H.R. Rabiee. A new multiresolution algorithm for image segmentation. Dans *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 2753–2756, 1998.

[11] Edmund Y. Lam et Joseph W. Goodman. A mathematical analysis of the dct coefficient distribution for images. Dans *IEEE Transactions on Image Processing*, volume 9, pages 1661–1666, October 2000.

[12] Jeff A. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Rapport technique, International Computer Science Institute, Berkeley, 1997.

[13] A. Benitez, J. Smith, et S. Chang. Medianet : A multimedia information network for knowledge representation. Dans *Conference on Internet Multimedia Management Systems*, volume 4210, November 2000.