Utilisation de connaissances a priori pour le paramétrage d'un algorithme de détection de textes dans les documents audiovisuels. Application à un corpus de journaux télévisés.

A priori knowledges using for the parametring of a text detection in audiovisual documents algorithm. Application on a corpus of tv news.

Rémi Landais[1,2], Christian Wolf[2], Laurent Vinet[1] & Jean-Michel Jolion[2]

[1] Institut National de l'Audiovisuel Direction de la recherche
4, Avenue de l'Europe
94366 Bry-sur-Marne Cedex
{rlandais,lvinet}@ina.fr

[2] LIRIS INSA Lyon
Bat. J. Verne
17, Avenue Capelle
69621 Villeurbanne Cedex
{jolion,wolf}@liris.cnrs.fr

Résumé

De tous les vecteurs d'informations qu'il est possible d'extraire automatiquement d'un document audiovisuel en vue de l'exploiter (le décrire, l'archiver ou le classer dans un corpus thématique), le texte figure parmi les plus intéressants d'un point de vue descriptif. Nous proposons dans cet article une nouvelle approche de la problématique de la détection du texte artificiel et ceci sur un corpus constitué de six journaux télévisés de 20H. A la différence de la majorité des méthodes déjà proposées, nous essayons ici de construire des modèles statistiques de la "physique" du texte qui reposent sur des connaissances a priori des documents étudiés : notre but est moins de parvenir à extraire tous les textes que de réduire le nombre de fausses alarmes dû à une modélisation trop lâche d'autant plus que ce nombre, lorsqu'il est trop élevé, empêche toute utilisation de l'algorithme de détection à des fins documentaires. Nous présentons ici notre méthode : construction d'un outil de vérité terrain efficace dont le modèle de données est basé sur le schéma de description MPEG7 VideoText ; mise au point, à partir des données issues des vérités terrains, d'un modèle spatio-temporel du texte de nature statistique ; paramétrage d'un algorithme de détection en fonction de ce modèle et évaluation de l'incidence de cet ajustement sur les résultats de l'algorithme.

Mots Clef

Document audiovisuel, detection de texte, vérité terrain, paramétrisation, modélisation physique.

Abstract

Among all the vectors of information which can be automatically extracted from an audiovisual document in order to exploit it (i.e : to describe it, to archive it or to classify it into a thematic corpus), text appears as one of the most interesting in a descriptive point of view. In this article we propose a new approach to artificial text detection problematic on a corpus composed of six news programs. Unlike most of the methods which have already been proposed we try here to construct physical statistical models of text which are based on a priori knowledge on the documents we study : our goal is less managing to extract all the texts than reducing the number of false alarms due to a too weak modelisation, all the more so since this number makes impossible any documentary using of the detection algorithm. Our method can be summed up as follows : conception of an effective ground truth editor whose data model is based on the MPEG7 description scheme VideoText ; adjustment of a statistical spatio-temporal model based on the data extracted from the ground truths ; parametrization of a detection algorithm according to this model and evaluation of its incidence on the results of the algorithm.

Keywords

Audiovisual document, text detection, ground truth, parametrization, physical modelisation.

1 Introduction

La mission de conservation du patrimoine audiovisuel français confiée à l'Institut National de l'Audiovisuel est une tâche d'une ampleur considérable : au delà du volume sans cesse croissant de documents à traiter, il existe de nombreuses difficultés liées à leur sauvegarde et à leur documentation. La campagne de numérisation a certes amélioré l'accessibilité à certains documents mais l'automatisation des processus mis en oeuvre dans la chaîne de l'archivage en est encore à ses prémices : à l'heure actuelle, la clé de voûte de l'archivage demeure la rédaction de notices descriptives par les documentalistes.

C'est dans ce contexte que le thème "Description des Contenus Audiovisuels" de la direction de la recherche de l'INA concentre ses travaux sur une utilisation de la numérisation en vue d'extraire de la vidéo des informations exploitables pour la description et l'exploitation des documents audiovisuels. Parmi ces informations, nous nous intéressons au texte inclus dans les vidéos qui est une source fiable de données descriptives, puisqu'il évite une phase d'interprétation subjective de l'information collectée.

Bien que la problématique de la détection de texte ait déjà été abordée, notre méthode présente une originalité dans le sens où nous travaillons à une méthode de pilotage d'un algorithme de détection. Notre hypohèse de départ est qu'en acquérant une modélisation suffisamment fine de l'objet "texte vidéo", on parvient à améliorer les résultats d'un algorithme de détection. Plus généralement, nous supposons que cet apprentissage supervisé évitera les écueils dûs à une modélisation trop lâche généralement basée sur des heuristiques.

Nous présenterons dans cet article les différentes étapes de notre méthode qui se limite pour l'instant aux seuls textes artificiels. La première partie motivera, d'un point de vue industriel, l'intérêt que nous portons au texte : quelles sont les applications possibles ? La seconde partie posera le problème de la constitution des vérités terrains indispensables à notre étude : comment caractériser le texte ? La troisième partie abordera la modélisation physique des textes à partir des vérités terrains. Enfin, après avoir présenté un état de l'art du domaine de la détection ainsi que détaillé l'algorithme que nous paramétrerons, nous aborderons la paramétrisation en elle-même ainsi que l'évaluation de son incidence sur les résultats de la détection.

2 Quelle utilisation du texte extrait ?

Parmi les utilisations du texte les plus évidentes, on compte l'aide à la rédaction des notices descriptives, à la segmentation plateaux/reportages dans les journaux télévisés, à l'indexation ou encore à la mise à jour de modèles de langages utilisés pour la transcription de la bande sonore.

En ce qui concerne l'aide à la rédaction des notices, elle est généralement motivée par une volonté d'accélérer le travail de documentation manuelle. Une documentation exclusivement basée sur les textes extraits est évidemment utopique : les textes sont soit trop peu nombreux dans le cas des textes artificiels ; soit trop souvent décorélés du thème du reportage dans le cas des textes de scène[1]. Une documentation partielle est alors possible mais n'accélère pas significativement le processus de documentation.

On peut aussi envisager d'utiliser le texte en support à un autre processus descriptif comme la segmentation ou la transcription de la bande sonore. Pour ce qui est de la segmentation, il apparait que les textes, essentiellement les textes artificiels, peuvent être utilisés dans le cadre de l'étude des journaux télévisés pour délimiter les plateaux des reportages. Les noms des journalistes interviennent par exemple systématiquement au début ou à la fin d'un reportage. Certains textes présentés dans des encarts lors d'un plateau, annoncent le reportage suivant. Pour ce qui est de l'aide à la transcription de la bande sonore, les textes extraits peuvent aider à résoudre le problème des entités nommées qui mettent en échec les systèmes de transcription.

L'application nous semblant la plus viable tant d'un point de vue scientifique que d'un point de vue industriel est la thématisation. Le programme de thématisation est né de la volonté de l'INA de valoriser ses fonds et de les structurer pour faire face à la demande, voire l'anticiper. Ce programme consiste à regrouper sous une étiquette commune des documents vidéo provenant de différentes sources, diffusés pendant différentes périodes, pour finalement former des corpus cohérents qui constituent des offres commerciales de type " prêt-à-diffuser ". On distingue trois types de corpus : les évènements, les thèmes et les personnalités dont voici quelques exemples :

- Thèmes : " Festival de Cannes ", " Football. L'équipe de France en coupe du monde : 1950-1986 "
- Evènements : " Mai 68 ", " La guerre du Vietnam : 1966-1977 "
- Personnalités : " Jacques Dutronc ", " Yves Montand ", " Robert Badinter "

Parmi toutes les tâches liées à la thématisation, la plus intéressante est celle de la complétion de corpus thématiques partiellement formés : le thème étant connu, il est possible de construire un modèle du vocabulaire thématique employé et même des "textes vidéos" présents dans les documents audiovisuels en relation avec ce thème. Il est alors envisageable de rechercher les documents en relation avec le thème en s'appuyant sur les textes extraits et sur un processus de catégorisation. A titre plus prospectif, il est envisageable de rechercher les nouveaux thèmes autour desquels il serait intéressant de former un corpus en s'appuyant sur la fréquence des termes relevés dans les vidéos.

[1] le texte de scène correspond au texte qui, au contraire du texte artificiel, ne résulte pas d'un traitement de post-production

3 La construction du modèle

3.1 Constitution du corpus

A la vue de notre problématique générale : détecter le texte, ainsi que des domaines applicatifs mentionnés précédemment, les journaux télévisés s'imposent comme les meilleurs candidats lors de la constitution de notre corpus. En effet, ceux-ci présentent de nombreux textes, autant artificiels que "de scène". L'impératif de quantité de texte, important dans le cadre de la constitution de modèles, est donc respecté.
En outre, il existe aussi dans le contenu des journaux télévisés, des particularités qui nous poussent à les privilégier :

- un fond important : le premier journal télévisé remonte à Juin 1949 ;
- une structure similaire selon les chaînes : les chaînes s'accordent autour d'une présentation typique axée sur un découpage "plateau/reportage". Seule la charte graphique différencie la forme des différents journaux ;
- une pluralité des contenus : les contenus sont représentatifs de la majorité des émissions de la télévision : un journal télévisé diffuse des extraits de film, de spectacle ou de retransmission sportive ; les sujets peuvent être comparés à des documentaires de format court et les interviews sur le plateau à des débats ;
- le journal télévisé est au coeur de la thématisation : il est fréquent que des extraits de journal télévisé soient retenus dans les corpus thématiques de l'INA ;

Notre corpus est formé de six journaux télévisés de 20h diffusés durant Août 1997. Nous travaillons sur des vidéos au format MPEG1, de résolution 352*288. Nous scindons notre corpus en un corpus d'apprentissage et un corpus de test contenant chacun trois documents.

3.2 La constitution des vérités terrains

Les données retenues. Pour modéliser correctement les textes artificiels, il convient de relever manuellement, les informations pertinentes qui leur sont relatives. En ce qui nous concerne, nous choisissons de conserver lors de l'établissement de ces vérités terrains quatre informations :

- la transcription du texte en conservant les retours à la ligne (ce qui permet de connaître le nombre de lignes de texte),
- les différentes positions de la zone du texte (délimitée par un rectangle) au cours du temps (une seule si le texte est fixe),
- les différents temps de références délimitant chaque mouvement d'une zone de texte (qui ne sont que deux si le texte est fixe),
- la classe du texte.

La classe du texte est choisie parmi les classes suivantes : "bandereau bleu", "blanc", "encart", "carte", "divers". Ces classes correspondent à des catégories "physique" et sémantique du texte : par exemple les bandereaux sont de couleur bleu, toujours situés en bas de l'écran et contiennent systématiquement le nom d'une personne s'exprimant à l'écran ou de journalistes ; les textes encart, situés en haut à droite de l'écran lors d'un plateau, annoncent le thème du reportage à venir ; les textes "cartes" correspondent aux noms de lieux stipulés sur des cartes ; les textes blanc, apparaissant en bas de l'écran, donnent des informations sur le lieu du reportage, sur le mode de diffusion (en direct, par téléphone ...) ou correspondent à des sous-titres ; la classe divers est quant à elle constituée de tous les textes dont ni le contenu, ni la forme ne leur permet de former une classe distincte.

Les modalités de réalisation. Chaque journal télévisé contient en moyenne 80 zones de texte artificiel différentes. Il convient donc de mettre en place au préalable une méthodologie précise que nous utiliserons pour réaliser les vérités terrains des six journaux de notre corpus.
Si le repérage temporel des zones de textes ne pose pas de problèmes, il n'en est pas de même pour le repérage spatial. Il est courant par exemple, qu'une zone de l'image présente différents textes regroupé dans un même encart ou un même bandereau. L'exemple le plus courant est celui de différents noms de journalistes présentés dans un bandereau. Dans ce cas, nous avons deux possibilités : privilégier la physique des textes ou leur sémantique. Dans le cas où nous privilégions la physique des textes, chaque zone suffisamment éloignée des autres zones d'un même type se révèle isolée dans la vérité terrain. Nous retenons ici cette solution bien que dans le cas de l'exemple des noms des journalistes, tous les noms forment une unité sémantique. Nous illustrons ce choix dans la figure 5 : on voit dans l'image (a) que l'on choisit d'isoler les deux groupes de noms.
Un dernier problème est celui de la catégorisation en classes. Ainsi, certaines classes laissent une part d'ambiguïté ne facilitant pas le classement. Par exemple, nous avons choisi de réunir dans une même classe "blanc", des textes ayant une physique et une sémantique relativement différente. Ce choix, discutable, repose sur la nécessité de réduire au maximum le nombre de classes. En effet, dans un cadre plus général, il n'est pas envisageable d'extraire lors de l'établissement des vérités terrains, toutes les informations relatives à chaque texte. Non seulement, cela entrainerait un allongement non raisonnable du temps passé à établir les vérités terrains mais cela aboutirait à terme à un non sens : le but est d'avoir la modélisation la plus générique possible pour pouvoir l'appliquer à un grand ensemble de documents.
A titre indicatif, nous avons choisi de stocker les données issues des vérités terrains au format XML. La structure de données que nous avons utilisée se rapproche du schéma de description MPEG7 : VideoTextType DS.

3.3 Une première exploitation des résultats : le calcul de statistiques

Préalablement à la qualification du texte par l'entremise de primitives complexes, nous choisissons de calculer des données statistiques de bas-niveau (moyennes, écarts types

et extrema) permettant de mieux nous représenter la nature spatio-temporelle du texte. Ces données sont résumées dans les tableaux ci-dessous (les données relatives au temps ont été mesurées en nombre de frames et celles relatives à la localisation spatiale en nombre de pixels) :

	Global	*Bandereau*	*Blanc*	*Carte*
Nombre de zones	340	196	39	33
Durée moyenne	183.44	114.1	104.46	72.060
Durée minimale	24	42	24	39
Durée maximale	1558	224	199	166
Ecart-type	171.21	**62.12**	**53.835**	**50.912**
			Encart	*Divers*
Nombre de zones			58	14
Durée moyenne			485.72	384.35
Durée minimale			179	38
Durée maximale			869	1558
Ecart-type			262.04	383.30

TAB. 1 – Statistiques temporelles des textes dans le corpus d'apprentissage

	Global		*Bandereau*	
	Moyenne	Ecart-type	Moyenne	Ecart-type
X	104.90	86.0	60.36	56.3
Y	194.45	62.8	232.0969	**6.10**
Hauteur	30	9.89	31	**7.20**
Largeur	105	56	106	56.3
L/H	3.48	**1.55**	3.29	**1.40**
	Blanc		*Carte*	
	Moyenne	Ecart-type	Moyenne	Ecart-type
X	105.76	52.8	145.57	69.8
Y	231.94	**7.74**	130.69	48.5
Hauteur	32	**13.2**	22	**8.34**
Largeur	149	**4.98**	70	62.3
L/H	4.85	**2.27**	3.16	**0.89**
	Encart		*Divers*	
	Moyenne	Ecart-type	Moyenne	Ecart-type
X	244.53	**1.83**	51.78	27
Y	94.36	39.9	128	61.5
Hauteur	26	**4.14**	48	**10.7**
Largeur	74	20.6	173	54.7
L/H	3.30	**1.06**	3.88	**1.96**

TAB. 2 – Statistiques spatiales des textes dans le corpus d'apprentissage

Parmi toutes les données présentées dans ces deux tableaux, seuls les écarts types peuvent être utilisés pour rechercher une quelconque stabilité intra-classe, tant d'un point de vue spatial que d'un point de vue temporel.
Il apparait alors que, d'un point de vue temporel, il ne semble pas exister de réelle stabilité intra-classe quant à la durée d'apparition. Ces résultats ont pour principal cause le fait que les textes sont souvent ajoutés à l'écran en direct et que leur durée varie donc en fonction de la personne chargée de les faire apparaitre. Cependant, d'un point de vue statistique, et par un test de comparaison de moyennes à 5% de risques, on peut montrer que seules les classes "Blanc-Bandereau" et "Encart-Divers" ont des moyennes statistiquement équivalentes. Bien que nous ne validions pas que ces lois soient normales, cette équivalence demeure valide de par la robustesse de ce test pour les petits échantillons [15].
En ce qui concerne les données spatiales, il apparait au contraire une certaine stabilité au niveau de la hauteur de la zone de texte. En ce qui concerne la localisation, elle dépend grandement de la classe considérée : les bandereaux bleus, qui apparaissent toujours en bas de l'écran sont par exemple stables relativement à leur position en Y. Une étude statistique similaire à celle effectuée sur les données temporelles permet de montrer par exemple que, si les localisations spatiales des classes "Bandereau" et "Blanc" sont équivalentes, leurs caractéristiques Hauteur et Largeur sont statistiquement différentes. Sur l'ensemble des caractéristiques, le paramètre Hauteur est le plus discriminant en suivant le critère de comparaison statistique des moyennes déjà utilisé.

Remarques et conclusion. Il convient de noter ici l'incidence des choix effectués quant à la réalisation des vérités terrains. Pour plus de clarté, nous nous nous baserons sur des exemples comme celui, déjà mentionné, des noms des journalistes présents dans un même bandereau. Nous avons fait le choix de regrouper les zones proches spatialement. Ceci a pour conséquance directe un manque de stabilité en ce qui concerne la position en X ainsi que la largeur des zones de texte appartenant à la classe bandereau. De la même façon, l'hétérogénéité de la classe "blanc" aboutit à une certaine instabilité tant temporelle que spatiale. Nous verrons en conclusion de cet article comment il est envisageable de remédier à ces problèmes.
Pour conclure cette partie, et avant d'expliquer plus précisément quels ont été nos choix en ce qui concerne la modélisation des textes et le paramétrage de notre algorithme présenté plus loin, nous pouvons remarquer que la réalisation des vérités terrains ainsi que le calcul de certaines données statistiques de bas niveau nous ont permis de déceler une certaine stabilité dans les textes observés, tant d'un point de vue spatial que temporel.
La suite de l'article portera successivement sur l'état de l'art du domaine de la détection de texte dans les documents audiovisuels, sur notre algorithme que nous utilisons pour notre étude et sur le paramétrage en lui-même.

4 État de l'art

L'extraction du texte présent dans les documents vidéo est un domaine de recherche très récent mais cependant très fécond. Les premières méthodes de détection du texte à avoir été proposées peuvent être considérées comme des généralisations de techniques du domaine de l'OCR dans le cadre de l'analyse automatique des documents. La problématique

de la détection est en effet proche de celle de l'analyse de la structure d'un document contenant du graphisme (journaux, page web etc, ...). Les algorithmes conçus pendant cette période segmentent les caractères du texte avant de les regrouper afin de détecter le texte comme des ensembles de caractères alignés. Une analyse detaillée de cette bibliographie peut être trouvée dans [19]. Les méthodes basées sur la segmentation apportent de très bons résultats quand ils sont appliqués aux images de très haute résolution, comme les journaux. Par contre, dans le cas de la vidéo, caractérisée par des polices de très petite taille et par des caractères connectés, on adopte généralement des méthodes basées sur la détection de contours ou l'analyse de texture.

La méthode de Sato, Kanade et al. [14] repose sur le fait que le texte peut être assimilé à une zone présentant de nombreux contours verticaux. L'algorithme de détection s'appuie donc sur un détecteur de contours et regroupe en rectangles les zones présentant une forte densité de contours verticaux. Les auteurs reconnaissent la nécessité d'augmenter la qualité de l'image avant la phase d'OCR. Par conséquent, ils effectuent une interpolation pour augmenter la résolution de chaque boîte de texte suivie par une intégration de plusieurs *frames* (en prenant les valeurs minimale/maximale des niveaux de gris). La reconnaissance est réalisée par une mesure de corrélation. Une méthode similaire est proposée par Agnihotri et Dimitrova [1]. Wu, Manmatha et Riseman [20] combinent quant à eux la recherche des contours verticaux avec un filtre de texture. Malheureusement ces algorithmes dependent fortement du résultat de la binarisation des contours. Un algorithme comparable est présenté par Myers et al [12]. Bien que les auteurs se concentrent surtout sur la reparation des effets de la distortion perspective, l'algorithme aborde aussi le problème de la détection de texte. Ceux-ci se limitent aux textes d'une longueur précise, ce qui facilite la recherche des points de fuite et la détection des lignes de base pour faire recouvrir au texte une vue frontale.

Le travail de LeBourgeois [9] est basé sur l'accumulation des gradients horizontaux. La phase de reconnaissance est basée sur des règles statistiques sur les caractéristiques de projections des niveaux de gris.

Bien que parfois employée uniquement durant la phase de post-traitement de la détection, la morphologie mathématique peut être aussi utilisée comme base de la phase même de détection. Hori cherche ainsi les regions de texte en dilatant les contours de l'image [7]. La plus grande partie de son travail est consacrée à la binarisation et la suppression du fond complexe. Hasan et Karam dilatent les contours détectés par un détecteur morphologique[6]. Néanmoins, les contraintes fortes concernant la géométrie et la taille du texte nuisent à la robustesse de leur méthode.

En ce qui concerne la discrimination entre le texte et le non-texte, les méthodes que nous venons de présenter s'appuient sur des caractéristiques simples, ce qui peut s'expliquer par le fait que dans un voisinage de petite taille, le texte ne posséde pas d'attribut distinctif robuste. La plupart des algorithmes se servent de la densité des contours ou de caractéristiques similaires comme les coefficients haute fréquence d'une décomposition en ondelette. L'emploi de caractéristiques traditionellement utilisées pour la détection de texture est limité au cas du texte court. En ce qui concerne les "longs textes", Sin et al présentent un algorithme destiné à la détection de texte écrit sur des panneaux [16]. Les caractéristiques utilisées pour la discrimination sont extraites de la fonction d'autocorrelation d'une ligne de l'image. Plusieurs lignes sont concatenées, ce qui peut provoquer des problèmes dûs au non-alignement de la phase de la tranformée de Fourier des différentes lignes. Cette méthode ne peut être appliquée aux textes courts.

Par ailleurs il existe plusieurs algorithmes basés sur l'apprentissage. Li et Doermann laissent glisser une fenêtre sur l'image, en utilisant un réseau de neurones de type MLP pour classifier chaque pixel comme "texte" ou "non texte" [10]. Les caractéristiques sont extraites des échelles de hautes fréquences d'une ondelette de Haar. Clark et Mirmehdi utilisent aussi un réseau de neurones pour la détection du texte [3]. Ils extraient des caractéristiques diverses comme la variance de l'histogramme, la densité des contours etc.

D'une façon similaire, Wernike and Lienhart fournissent la densité moyenne et la densité directionelle de contour à un classificateur de type MLP [18]. Dans un article plus récent [11], les mêmes auteurs proposent un vecteur de caractéristiques composé d'un carte 20×10 de contour de couleur. Jung apprend directement les niveaux de gris de la vidéo avec un réseau de neurones [8]. Dans un approche similaire, Kim et al. apprennent également les niveaux de gris du texte avec des Support Vector Machine. Tang et al. employent un algorithme d'apprentissage non supervisé pour la détection [17]. Leurs caractéristiques étant basées sur la dynamique des niveaux de gris uniquement, ils détectent l'apparition et la disparition de texte. Par conséquent, le texte présent du début à la fin d'un plan n'est pas détecté. La méthode proposée par Chen et al [2] comporte deux étapes . La première étape est consacrée à la détection de texte par détection de contour et morphologie mathématique. La deuxième étape consiste à classer les zones obtenues en deux classes : texte et non-texte. Ils utilisent pour cela les SVM en se basant pour ce qui est du mode de décision sur une carte contentant pour chaque pixel la distance minimum au contour la plus proche. On remarquera que cette méthode souffre de sa première phase, trop simple, et qui ne prend ainsi pas en compte suffisamment d'aspects du texte.

Il convient ici de rappeler que les méthodes d'apprentissage dépendent grandement de la qualité et de la quantité des données d'apprentissage. Or le texte apparaissant dans les vidéos peut avoir une multitude de polices, de styles, de tailles, d' orientations etc., ce qui rend la généralisation d'un système d'apprentissage très difficile. Les caractéristiques très simples, comme les niveaux de gris etc., ne permettent sans doute pas d'aboutir à un système d'apprentis-

sage qui soit robuste à la diversité des textes possible.
Une méthode qui agit directement dans le domaine de la vidéo compressée est proposée par Zhong, Zhang et Jain [22]. Leurs caractéristiques sont calculées directement à partir des coefficients de DCT des données MPEG. La détection est basée sur la recherche des variations horizontales d'intensité, suivie par une étape de morphologie mathématique. Randall et Kasturi [4] calculent une énergie de texture à partir des coefficients de DCT. Malheureusement leurs bons résultats se limitent au cas des textes de très grande taille. La méthode proposée par Gu est conçue pour le texte statique horizontal [5]. Mis à part les coefficients de DCT il utilise aussi des vecteurs de mouvement pour la suppression des fausses alarmes. On remarquera pour conclure que le grand nombre de formats vidéo (MPEG1, MPEG2, MPEG4, Sorenson, RealVideo) empêche toute pérennisation de ces méthodes, d'autant que la recherche dans le domaine de la compression vidéo est extrêmement fertile.

5 Le système de détection

Nous présentons ici dans les grandes lignes l'algorithme de détection que nous avons utilisé pour notre étude. Celui-ci est détaillé plus amplement dans [19] [2]. La détection du texte est réalisée dans chaque frame de la vidéo. Les rectangles figurant la localisation du texte sont suivis pendant leur période d'apparition pour associer les rectangles se correspondant dans les différents frames. Cette information est nécessaire pour améliorer le contenu de l'image, ce qui peut être atteint par l'intégration de plusieurs rectangles contenant le même texte. Cette phase doit produire des sous-images d'une qualité en accord avec les pré-requis d'un processus OCR.
Notre approche s'appuie sur le fait que les caractères du texte forment une texture régulière contenant des contours verticaux alignés horizontalement. L'algorithme de détection reprend la méthode de LeBourgeois [9], qui utilise l'accumulation des gradients horizontaux pour détecter le texte :

$$A(x,y) = \sum_{i=-t}^{t} \frac{\partial I}{\partial x}(x+t, y)$$

Les paramètres de ce filtre sont l'implémentation de l'opérateur dérivatif (nous avons choisi la version horizontale du filtre Sobel, qui a obtenu les meilleurs résultats) et la taille de la fenêtre de l'accumulation, qui correspond à la taille des caractères. Comme les résultats ne dépendent pas trop de ce paramètre, nous l'avons fixé à $2t+1 = 13$ pixels. Le résultat de cette phase consiste en une image contenant une mesure de la probabilité de chaque pixel d'être un pixel de texte.
La binarisation des gradients cumulés est poursuivie par une version deux-seuils de la méthode d'Otsu [13]. Cette méthode calcule un seuil global à partir de l'histogramme des niveaux de gris. Nous avons changé la décision de binarisation pour chaque pixel comme suit :

$$\begin{array}{lcl} I_{x,y} < k_{bas} & \Rightarrow & B_{x,y} = 0 \\ I_{x,y} > k_{haut} & \Rightarrow & B_{x,y} = 255 \\ k_{bas} > I_{x,y} > k_{haut} & \Rightarrow & B_{x,y} = \end{array}$$

$$\begin{cases} 255 & \text{s'il existe un chemin} \\ & \text{à un pixel } I_{u,v} > k_{haut} \\ 0 & \text{sinon} \end{cases}$$

où le seuil k_{haut} correspond au seuil calculé par la méthode d'Otsu, et le seuil k_{bas} est calculé à partir du seuil k_{haut} et le premier mode m_0 de l'histogramme : $k_{bas} = m_0 + 0.87 \cdot (k_{haut} - m_0)$.
Avant de passer des pixels à des zones compactes, nous utilisons un post-traitement morphologique afin d'extraire les rectangles englobant des zones de texte. Cette dernière étape permet d'atteindre plusieurs buts :
– Réduire le bruit.
– Corriger des erreurs de classification à partir de l'information du voisinage.
– Connecter des caractères afin de former des mots complets.

La phase morphologique est composée des étapes suivantes :
– Fermeture (1 itération).
– Suppression des ponts (Suppression de tous les pixels qui font partie d'une colonne de hauteur inférieure à 2 pixels).
– Dilatation conditionnelle (16 itérations) suivie par une érosion conditionnelle (16 itérations).
– Érosion horizontale (12 itérations).
– Dilatation horizontale (6 itérations).

Nous avons conçu un algorithme de dilatation conditionnelle pour connecter les caractères d'un mot ou d'une phrase, qui forment différentes composantes connexes dans l'image binarisée. Cela peut arriver si la taille de la police ou si les distances entre les caractères sont relativement grandes. L'algorithme est basé sur une dilatation "standard" ayant pour élément structurant $B_H = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$ La nouveauté consiste à restreindre cette dilatation aux seuls pixels satisfaisant les conditions suivantes : un pixel P est dilaté seulement, si
– la différence de hauteur de la composante C_a incluant le pixel P et la composante C_b voisin à droite ne dépasse pas un seuil t_1.
– la différence de positions y de ces deux composantes ne dépasse pas un seuil t_2.
– la hauteur de la boite englobante incluant ces deux composantes ne dépasse pas un seuil t_3.

Les pixels dilatés sont marqués avec une couleur spéciale. L'érosion conditionnelle qui suit la dilatation conditionnelle utilise le même élément structurant B_H. Seuls les pixels marqués au cours de la dilatation peuvent être supprimés durant cette phase d'érosion. Après les deux opéra-

[2] Cet algorithme a donné lieu à un brevet FR 01 06776 déposé par France Télécom le 23 Mai 2001

tions tous les pixels marqués qui n'ont pas été érodés sont considérés comme étant des pixels de texte.

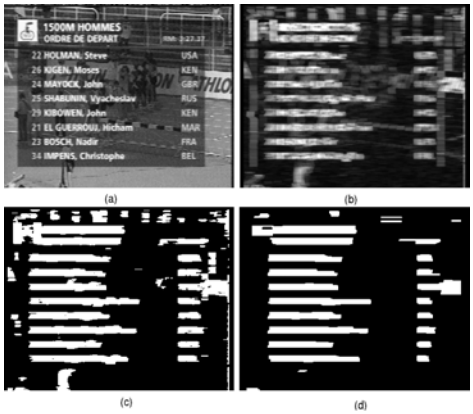


FIG. 1 – Les résultats intermédiaires pendant la détection : l'image d'entrée (a), le gradient (b), l'image binarisée (c), l'image après le post-traitement (d).

La figure 1 montre les résultats pendant la détection. Les figures 1a et 1b affichent l'image originale et l'image des gradients accumulés. Les régions de texte sont visiblement marquées en blanc. La figure 1c montre l'image binarisée, portant encore du bruit, bruit qui est ensuite supprimé dans la figure 1d.

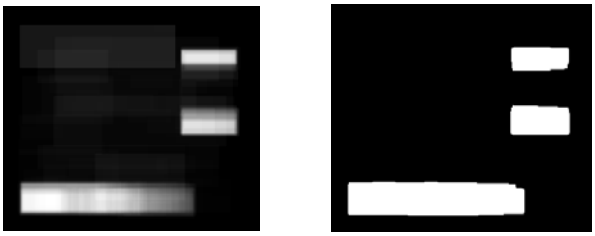
Après le traitement morphologique, on applique à chaque région (ou composante connexe) des contraintes géométriques afin de réduire le nombre des fausses alarmes. Enfin, une recherche des cas particuliers est appliquée pour considérer aussi les traits détachés de la zone de chaque composante (les hauts de casse et les bas de casse).

Le but du module de suivi est d'associer les rectangles correspondant à un même texte apparaissant dans plusieurs images successives dans le but final de réduire encore une fois le nombre de fausses alarmes. Pour parvenir à associer correctement les différentes instances d'un même texte, nous utilisons des mesures du recouvrement calculées entre chaque rectangle de l'image courante et les rectangles de la liste des apparitions, liste qui se forme au fur et à mesure de la détection. La durée d'apparition d'un texte, c.à.d le nombre de frames où celui-ci apparaît à l'écran, nous sert alors de mesure de stabilité. Nous considérons les apparitions de longueur plus courte qu'un seuil fixe comme fausses alarmes.

6 Le paramétrage de l'algorithme

L'algorithme de détection utilise un grand nombre de paramètres. Tous ne sont pas explicites ou accessibles facilement. De plus, notre but est de parvenir à exploiter des données de bas-niveau construites à partir des vérités terrains pour le paramétrage. Dès lors, les paramètres sur lesquels nous devons influer se doivent d'être explicites.

Le paramétrage spatial. Nous avons vu lors de l'étude des données statistiques de bas-niveau, qu'il existait une certaine stabilité intra-classe, observée essentiellement dans les données relatives à la position des textes ainsi que dans celles relatives à leur taille.
Nous décidons alors de limiter la détection aux zones de l'image dans lesquels on relève le plus de texte au cours du temps dans les vérités terrains. Nous illustrons ce choix dans la figure 2.



(a) Image originale (b) Image seuillée

FIG. 2 – Le choix des zones de détection

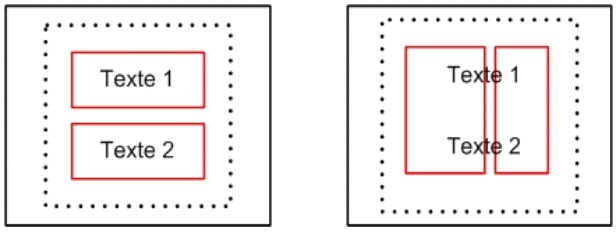
Dans l'image (a), la valeur de chaque pixel est relative au nombre de zones de texte auxquelles celui-ci appartient au cours du temps en comptant pour chacune de ces zones le nombre de frames correspondant à la durée d'apparition.
L'image (b) correspond à la première image seuillée. Ce seuil est déterminé manuellement, de telle sorte que les trois zones principales déjà relevées dans la première image soient isolées. Ce choix repose sur notre classification : ces trois zones contiennent tous les "bandereaux", la majorité des textes "blancs" ainsi que les "encarts". Les textes "cartes" ainsi que les textes "divers", dont la répartition spatiale n'est pas stable sont donc en grande majorité écartés de notre étude.
Nous utilisons aussi comme paramètre le ratio $\frac{Largeur}{Hauteur}$ qui constitue dans l'algorithme un critère de sélection géométrique des zones de texte en vue de réduire le nombre de fausses alarmes. Nous choisissons ici de limiter les zones de texte valides aux seules zones dont le ratio est compris entre le minimum et le maximum constatés lors de l'établissement de la vérité terrain.
Enfin, une fois la phase de détection terminée, nous appliquons un autre filtrage géométrique, qui consiste à supprimer toutes les zones dont les caractéristiques de taille (hauteur et largeur) ne sont pas comprises entre les extrema constatés lors de l'établissement des vérités terrains.

Les paramètres temporels. Il n'apparait pas, au vue des données calculées, de stabilité relative aux durées des textes selon leur type.
Néanmoins, nous pouvons utiliser facilement la donnée relative à la durée minimale globale. Il existe dans l'algorithme un paramètre relatif à la durée minimale d'apparition permise pour une zone de texte : toute zone de texte

qui n'est pas détectée durant un nombre minimal d'images n'est pas validée en tant que zone de texte. Nous fixons alors ce paramètre à la durée minimale calculée à partir des vérités terrains.

7 Résultats

Les modalités d'évaluation. Pour évaluer l'incidence de notre paramétrage, nous comparons les résultats obtenus dans chaque "mode" de détection (paramétré ou non) avec les vérités terrains. Dans le cadre de cette évaluation, nous ne considérons que les aspects spatiaux, nous ne tenons pas compte de la temporalité des zones de texte. Ainsi, nous considérerons qu'un recouvrement temporel d'une image suffit pour que deux zones soient considérées comme équivalentes d'un point de vue temporel. En ce qui concerne l'aspect spatial, l'évaluation est la suivante : une zone détectée est considérée comme une fausse alarme si elle n'intersecte pas suffisamment l'une des zones de la vérité terrain. De la même façon, une zone de la vérité terrain est considérée comme un oubli si elle n'intersecte pas suffisamment l'une des zones détectées par l'algorithme mis en jeu. Nous fixons le seuil concernant le taux de recouvrement à 70%.
Pour autant, il demeure certains cas concernant la fusion ou la scission de zones qui remettent en question notre méthode d'évaluation . En effet, si les zones détectées sont données en entrée à un OCR, certaines d'entre elles, considérées comme des fausses alarmes dans le cadre de la détection (suite à nos choix : "chaque zone suffisamment éloignée des autres zones [...] se révèle isolée dans la vérité terrain"), pourraient déboucher sur des résultats de reconnaissance concluants. C'est la cas illustré dans la figure 3 (les traits en pointillé désigne les zones de la vérité terrain tandis que les traits pleins désignent celles détectées par l'algorithme) :



(a) Cas d'une scission verticale

(b) Cas d'une scission horizontale

FIG. 3 – Le cas de la scission

Dans ces deux exemples les quatres zones détectées sont considérées comme des fausses alarmes et ceci malgré l'apparente réussite de la détection dans une perspective de reconnaissance par un OCR, obtenue dans l'image (a). Néanmoins, le cas de l'image (b) nous a incité à ne pas modifier notre méthode d'évaluation.

Résultats. Nous évaluons notre paramétrage sur trois journaux télévisés. Les deux figures suivantes illustrent le procédé de détection :

(a) Zones de la vérité terrain

(b) Zones détectées par l'algorithme non-paramétré

(c) Zones détectées par l'algorithme paramétré

FIG. 4 – Résultat de la paramétrisation : suppression de fausses alarmes

(a) Zones de la vérité terrain

(b) Zones détectées par l'algorithme non-parametré

(c) Zones détectées par l'algorithme parametré

FIG. 5 – Résultat de la paramétrisation avec scission de zones de la vérité terrain par l'algorithme

La figure 4 montre comment la paramétrisation permet de supprimer deux fausses alarmes (nous choisissons de ne pas compter le logo comme une fausse alarme). En ce qui concerne la figure 5 on constate que les textes sont correctement détectées dans les deux "modes" de détection en vue d'une reconnaissance par un OCR et que la paramétrisation permet de supprimer une fausse alarme. Pour autant, dans cet exemple et selon nos modalités d'évaluation, le fait que l'algorithme scinde la zone retenue pour la vérité terrain, lui est préjudiciable : cela provoque deux fausses alarmes ainsi qu'un oubli. Cet exemple, atypique du reste des résultats, montre ainsi les limites d'une évaluation "automatique". Il serait ainsi envisageable de pondérer l'évaluation en introduisant des poids relatifs à la fusion ou à la scission de zones détectées. Ce type d'évaluation a d'ailleurs déjà été proposé dans [21].
Pour autant, l'évaluation des résultats permet de constater que nous parvenons à augmenter d'un facteur 3 environ le taux de précision calculé ainsi : $\frac{nombre\ de\ zone\ correctement\ détectées}{nombre\ de\ zones\ détectées}$. Cette augmentation, qui correspond à une réduction des fausses alarmes est compensée par une diminution du taux de rappel

($\frac{nombre\ de\ zone\ correctement\ détectées}{nombre\ de\ zones\ dans\ la\ vérité\ terrain}$) d'un facteur 1.8. Il est toutefois possible de qualifier les textes qui n'ont pas été détectés par le "mode" paramétré : on peut dans un premier temps différencier les textes oubliés dans les zones où l'on applique la détection de ceux oubliés à l'extérieur de ces zones. Il s'avère alors que les textes non-détectés dans ces zones appartiennent à 97.3% aux classes "bandereau", "blanc" et "encart" tandis que les textes oubliés à l'extérieur de ces zones appartiennent à 83.7% aux classes "cartes" et "divers". Ces chiffres tendent donc à valider notre modèle spatial : les textes oubliés à l'extérieur des zones appartiennent aux classes dont nous avons montré qu'elles n'étaient pas stable spatialement, du moins du point de vue de leur position. En outre, il est envisageable, puisque les textes de carte apparaissent à l'écran dans un contexte visuel particulier (celui d'une carte) de repérer les cartes dans les journaux télévisés pour appliquer ensuite à ces images un mode de détection particulier. Il conviendra ensuite de trouver un "mode" de paramétrage qui soit adapté à la détection des textes divers, minoritaires.

8 Conclusions et perspectives

Cette étude a montré qu'il était possible d'améliorer principalement le taux de rappel des résultats d'un algorithme de détection en s'appuyant sur des connaissances a priori acquises en établissant des vérités terrains. Il conviendra donc par la suite d'approfondir le travail dans ce sens en développant différents axes de réflexions :

- **L'exploitation des caractéristiques temporelles des textes** : S'il est vrai qu'il ne semble pas exister de stabilité quant à la durée des apparitions des textes, il n'en demeure pas moins qu'il est envisageable d'exploiter leur date d'apparition. En effet, la structure des journaux télévisés est extrèmement codifiée de sorte que certains textes apparaissent systématiquement à des dates clés du journal : lors de l'annonce d'un sujet pour les encarts, en début de sujet pour un texte "blanc" décrivant la localisation du reportage ou encore en fin de sujet pour les noms des journalistes ... Néanmoins, cette corrélation entre l'apparition de certains textes et certaines dates du journal télévisé ne peut se faire qu'au niveau des sujets et non du journal dans son entier : il est difficile de prévoir sur un journal entier les dates d'apparition de chaque type de texte. Nous devrons donc au préalable segmenter la vidéo en plateaux/reportages pour pouvoir exploiter ces données relatives à la répartition temporelle des textes.
- **Le choix des classes** : comme nous l'avons déjà fait remarquer, le choix des classes de texte que nous avons fait peut parfois aboutir à des classes hétérogènes. Une solution, qui permettrait dans un même temps de réduire l'hétérogénéité intra-classe mais aussi d'être robuste à un changement de corpus, serait de pratiquer une classification automatique : une première phase déterminerait le nombre de classes en s'appuyant sur un certain nombre de données calculées d'après les vérités terrains, ensuite, on pratiquerait un clustering sur les zones relevées dans la vérité terrain en appliquant finalement certains critères d'homogénéité intra-classe pour réduire les éventuelles erreurs liées au choix initial du nombre de classes ; si une classe s'avère trop hétérogène, celle-ci est séparée en différentes classes jusqu'à stabilisation du processus.
- **Le schéma général** : nous nous sommes pour l'instant intéressé uniquement à la phase de détection. Pour autant l'algorithme que nous utilisons pratique, à la suite de la phase de détection, une phase de suivi ainsi qu'une phase d'extraction que nous devrons elles aussi paramétrer. En outre, notre objectif final est de parvenir à finaliser la boucle schématisée ci-dessous :

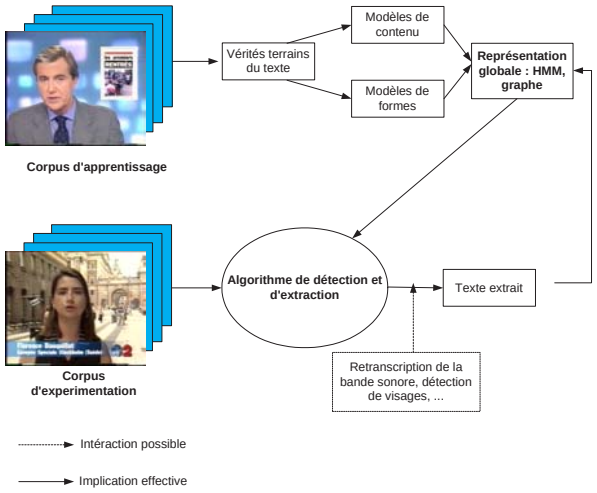


FIG. 6 – Schéma du système global envisagé

Pour l'instant nous n'avons pas finalisé la phase de modélisation au sens où nous n'avons pas intégré les informations de type physique et sémantique dans notre modèle statistique. Nous pouvons déjà mentionner ici certaines des catégories formant la typologie sémantique des textes présents dans les journaux télévisés : on peut tout d'abord distinguer les textes n'ayant aucune valeur descriptive des textes relatifs au contenu informatif du journal mais aussi des textes relatifs au mode de réalisation de l'émission (noms des réalisateurs, textes relatifs au mode d'émission : en direct...). Ensuite, il faudrait différencier les textes selon leur portée temporelle (toute l'émission, un sujet dans son entier, un reportage seul ou encore seulement quelques images dans le cas des noms des intervenants). A ceci peut s'ajouter une distinction relative au lien entre le texte considéré et les autres modalités : le texte est-il autonome, apparait-il plutôt en support de l'image ou de la bande sonore ? Ces différentes catégories nous permettront de catégoriser plus précisément chaque type de texte d'un point de vue sémantique. Cette catégorisation nous permettra ensuite, en associant le sens des textes à leur mode de représenta-

tion physique de n'extraire que certains textes dont nous avons établi qu'ils pouvaient nous être utiles dans l'application visée. Par exemple, dans le cadre de l'aide à la complétion de corpus thématiques ouverts, nous pourrons, en établissant quels sont les types sémantiques et physiques des textes relatifs au thème recherché, cibler au mieux notre recherche.

On remarquera pour conclure qu'il est aussi envisageable, comme cela apparait sur le schéma de notre système, d'utiliser d'autres modules d'analyse pour affiner les résultats de la détection ou de l'extraction des textes : la transcription de la bande sonore peut permettre de valider certains termes pour lesquels la phase de reconnaissance par OCR n'est pas concluante. La détection des visages peut quant à elle permettre de situer dans le document les interventions de personnes lesquelles sont souvent présentées par un texte mentionnant leur nom et leur qualification.

Remerciements

Steffen Lalande pour son aide au développement de l'outil de vérité terrain.

Références

[1] L. Agnihotri et N. Dimitrova. Text detection for video analysis. Dans *IEEE Workshop on Content-based Access of Image and Video Libraries*, pages 109–113, 1999.

[2] D. Chen, H. Bourlard, et J.-P. Thiran. Text identification in complex background using svm. Dans *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 621–626, 2001.

[3] P. Clark et M. Mirmehdi. Combining Statistical Measures to Find Image Text Regions. Dans *Proceedings of the International Conference on Pattern Recognition*, pages 450–453, 2000.

[4] D. Crandall et R. Kasturi. Robust Detection of Stylized Text Events in Digital Video. Dans *Proceedings of the International Conference on Document Analysis and Recognition*, pages 865–869, 2001.

[5] L. Gu. Text Detection and Extraction in MPEG Video Sequences. Dans *Proceedings of the International Workshop on Content-Based Multimedia Indexing*, pages 233–240, 2001.

[6] Y.M.Y. Hasan et L.J. Karam. Morphological text extraction from images. *IEEE Transactions on Image Processing*, 9(11) :1978–1983, 2000.

[7] O. Hori. A video text extraction method for character recognition. Dans *Proceedings of the International Conference on Document Analysis and Recognition*, pages 25–28, 1999.

[8] K. Jung. Neural network-based text location in color images. *Pattern Recognition Letters*, 22 :1503–1515, 2001.

[9] F. LeBourgeois. Robust Multifont OCR System from Gray Level Images. Dans *Proceedings of the 4th International Conference on Document Analysis and Recognition*, pages 1–5, 1997.

[10] H. Li et D. Doermann. Automatic text detection and tracking in digital video. *IEEE Transactions on Image Processing*, 9(1) :147–156, 2000.

[11] R. Lienhart et A. Wernike. Localizing and Segmenting Text in Images, Videos and Web Pages. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(4) :256–268, 2002.

[12] G. Myers, R. Bolles, Q.-T. Luong, et J. Herson. Recognition of Text in 3D Scenes. Dans *Fourth Symposium on Document Image Understanding Technology, Maryland*, pages 85–99, 2001.

[13] N.Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1) :62–66, 1979.

[14] T. Sato, T. Kanade, E.K. Hughes, M.A. Smith, et S. Satoh. Video OCR : Indexing digtal news libraries by recognition of superimposed captions. *ACM Multimedia Systems : Special Issue on Video Libraries*, 7(5) :385–395, 1999.

[15] B. Sherrer. *Biostatistique*. G.Morin Editeur, 1982.

[16] B.-K. Sin, S.-K. Kim, et B.-J. Cho. Locating characters in scene images using frequency features. Dans *Proceedings of the International Conference on Pattern Recognition*, volume 3, pages 489–492, 2002.

[17] X. Tang, X. Gao, J. Liu, et H. Zhang. A Spatial-Temporal Approach for Video Caption Detection and Recognition. *IEEE Transactions on Neural Networks*, 13(4) :961–971, 2002.

[18] A. Wernike et R. Lienhart. On the segmentation of text in videos. Dans *Proc. of the IEEE Int. Conference on Multimedia and Expo (ICME) 2000*, pages 1511–1514, 2000.

[19] C. Wolf et J.-M. Jolion. Extraction and recognition of artificial text in multimedia documents. *Pattern Analysis and Applications (à paraître)*.

[20] V. Wu, R. Manmatha, et E.M. Riseman. Textfinder : An automatic system to detect and recognize text in images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(11) :1224–1229, 1999.

[21] B. Yanikoglu et L. Vincent. Pink Panther : A complete environment for ground-truthing and benchmarking document page segmentation. *Pattern Recognition*, 31(20) :1191–1204, 1998.

[22] Y. Zhong, H. Zhang, et A.K. Jain. Automatic Caption Localization in Compressed Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4) :385–392, 2000.