# Création de résumés vidéos par une approche statistique

Caroline Lacoste, Ronan Fablet, Patrick Bouthemy, J.F. Yao

# Création de résumés de vidéos par une approche statistique

# Video summarization using a statistical approach

C. Lacoste*, R. Fablet**, P. Bouthemy* et J.F. Yao***

*IRISA/INRIA **IRISA/CNRS ***IRMAR
Campus universitaire de Beaulieu, 35042 Rennes Cedex, France
mel: bouthemy@irisa.fr, Jian-Feng.Yao@univ-rennes1.fr

## Résumé

*Nous présentons une approche originale pour la création automatique de résumés de vidéos. L'objectif est de fournir une version très courte d'un document vidéo, représentative de son contenu. Notre étude repose sur l'analyse du contenu dynamique de micro-segments de la vidéo, décrit par les distributions de cooccurrences temporelles de quantités locales de mouvement, distributions représentées par des modèles de Gibbs temporels. Nous proposons, tout d'abord, de réaliser une segmentation temporelle de la vidéo en plages homogènes au sens d'une notion d'activité de mouvement, résultant d'une classification hiérarchique ascendante et conduisant à une sélection d'instants représentant au mieux la vidéo. Une seconde méthode repose sur la sélection directe de plages pertinentes et se base sur une modélisation par chaîne de Markov cachée. Les premiers résultats expérimentaux obtenus permettent de vérifier l'intérêt de l'approche statistique adoptée.*

## Mots-clés

Modèles de Gibbs, cooccurrences, classification, chaîne de Markov cachée, mouvement, vidéo.

## Abstract

*We present in this paper an original approach to video summarization. Our goal is to extract from a video document a brief abstract representative of its content. We rely on the analysis of the dynamic content of micro-sequences of the processed video. Motion characterization is embedded in a statistical Gibbsian framework specified from temporal cooccurrences of local motion quantities. We propose two complementary techniques for video summarization. The first one involves a hierarchical temporal segmentation of the video into homogeneous units according to motion activity information; the second one is concerned with the selection of units of interest and exploits hidden Markov models. We report first experimental results carried out on a real video document.*

## 1 Introduction

Avec les récentes avancées en termes de diffusion et d'utilisation de documents audiovisuels sous des formes numériques, il devient nécessaire de développer de nouveaux outils pour manipuler ces volumes de données multimédia. La création de résumés de vidéos constitue l'une des fonctionnalités importantes à mettre en oeuvre dans cette optique aussi bien pour des applications professionnelles que domestiques. D'une manière générale, notre objectif est d'extraire d'un document vidéo une version condensée très courte qui soit porteuse de l'information "utile" contenue dans la vidéo originale.

De tels résumés de vidéos pourront être exploités dans un contexte professionnel d'archivage et de consultation de bases de documents audiovisuels pour faciliter les opérations de visualisation, de navigation ou de présélection. D'un point de vue domestique, la création de résumés de vidéos pourrait faire partie des nouveaux services liés à la télévision numérique. En effet, dans un avenir proche, les décodeurs numériques associés aux téléviseurs seront munis de capacités très importantes de stockage (disques durs). La fonctionnalité de résumé de vidéo pourra offrir à l'utilisateur un aperçu rapide du contenu de ces enregistrements afin qu'il puisse sélectionner ce qui l'intéresse. Il est à noter que pour ces deux catégories d'applications, l'enjeu n'est pas de construire des résumés qui aient une véritable portée sémantique comme les résumés textuels des magazines annonçant les programmes de télévision. Il s'agit plutôt d'extraire d'une vidéo une représentation visuelle qui fournisse un aperçu global "raisonnable et utile" de son contenu.

En nous appuyant sur nos travaux antérieurs [3, 4, 5] concernant la modélisation statistique du mouvement dans des séquences d'images, nous proposons deux approches originales et complémentaires pour la création automatique de résumés de vidéos. La première repose sur une segmentation temporelle de la vidéo en plages homogènes selon des critères de mouvement et la deuxième est basée sur la sélection directe de plages d'intérêt s'appuyant sur un modèle de Markov caché.

Après une présentation succincte des idées directrices de ces travaux dans le paragraphe 2, nous rappelons brièvement dans le paragraphe 3 les modèles statistiques de mouvement utilisés. Le paragraphe 4 décrit la méthode de segmentation des vidéos en plages homogènes, et le paragraphe 5 celle de sélection de plages d'intérêt. Des résultats expérimentaux sont présentés au paragraphe 6, et le paragraphe 7 conclut cet article.

# 2 Contexte de l'étude

La création de résumés de vidéos a pour objectif de sélectionner ou de collecter des segments temporels significatifs dans le document vidéo original. Deux catégories d'approches peuvent être distinguées. En premier lieu, il est possible de considérer un certain nombre de règles pouvant définir ce qui caractérise un moment d'intérêt dans une vidéo. Dans cette optique, les travaux décrits dans [7] proposent de considérer plus particulièrement les scènes de dialogue, d'action et les génériques dans le cas de fictions. Des techniques de détection de visages, d'extraction d'indicateurs de mouvement et d'extraction de zones de texte dans l'image permettent en pratique de détecter les moments d'intérêt. La définition de ce type de règles nécessite de disposer d'une connaissance *a priori* sur le contenu des documents à traiter. Ceci limite en pratique cette catégorie de techniques à des classes de documents spécifiques.

Nous souhaitons au contraire proposer des méthodes générales sans connaissance *a priori* sur le contenu des documents traités. Les travaux décrits dans [8, 9] s'inscrivent dans cette démarche. Dans [9], les auteurs proposent de réaliser un échantillonnage temporel adapté des documents vidéos en fonction d'un indicateur de mouvement calculé dans chaque image. Dans [8], l'évaluation d'une similarité au sens de la couleur entre images successives est mise à profit pour étudier les variations du contenu des images et sélectionner les instants correspondant à des variations importantes. Nous nous plaçons dans cette catégorie d'approches. Notre objectif est de créer un résumé dont le contenu soit visuellement représentatif du document original. Notre approche repose sur l'analyse de l'évolution temporelle de grandeurs caractéristiques directement extraites du signal vidéo.

De manière analogue à [9], nous pensons que les informations de nature dynamique sont très pertinentes pour qualifier l'intérêt d'un segment temporel de la vidéo. Elles présentent de plus une généricité importante pour appréhender des classes de documents variées. Nous basons ainsi notre étude sur l'analyse des variations temporelles du contenu dynamique. Pour caractériser l'information de mouvement contenue dans une vidéo, nous exploitons nos travaux antérieurs [3, 4, 5] sur la modélisation statistique non paramétrique du mouvement dans des séquences d'images. Ils nous permettent d'évaluer des similarités de mouvement entre séquences d'images dans un cadre probabiliste. Nous proposons alors deux méthodes complémentaires pour la création de résumés de vidéos. La première vise à déterminer des plages homogènes dans une vidéo. Nous exploitons une mesure de similarité au sens du mouvement définie à partir d'une distance entre modèles statistiques de mouvement. La phase de segmentation temporelle proprement dite s'appuie sur une technique de classification hiérarchique ascendante appliquée à des micro-séquences résultant d'une partition initiale fine de la vidéo traitée. La deuxième méthode vise la sélection directe de zones pertinentes. Elle est formulée comme un problème d'inférence bayésienne et exploite une modélisation par chaîne de Markov cachée à deux états, "pertinent" et "non pertinent".

# 3 Modélisation statistique de l'activité de mouvement

Les deux méthodes proposées pour la création de résumés de vidéos s'appuient sur la modélisation du contenu dynamique dans des segments temporels de la vidéo traitée. Afin d'appréhender une grande variété de situations dynamiques (scènes d'extérieur, d'intérieur, scènes d'éléments naturels, scènes de personnages...), nous privilégions une notion générale d'activité de mouvement et nous exploitons les modèles statistiques de mouvement que nous avons introduits dans [3, 4, 5]. Cette approche par modélisation statistique de l'activité de mouvement s'est déjà révélée fructueuse pour des problèmes de reconnaissance du mouvement et de recherche de vidéos par l'exemple en indexation vidéo [4, 5]. Dans ce paragraphe, nous ne donnons que les éléments essentiels de cette modélisation. Nous invitons le lecteur à se référer à [3] pour une présentation détaillée.

Nos modèles statistiques d'activité de mouvement reposent sur l'analyse de la distribution de mesures locales partielles de mouvement. Ces quantités locales de mouvement sont directement évaluées à partir des dérivées spatio-temporelles de l'intensité après compensation du mouvement dominant dans l'image [10] (qui est supposé être dû au déplacement de la caméra). Elles sont ensuite quantifiées sur un ensemble $\Lambda$ de valeurs discrètes. Les modèles exploités dans la suite sont des modèles de Gibbs temporels qui sont spécifiés à partir de distribution de cooccurrences temporelles des quantités locales de mouvement. Plus précisément, à partir d'une séquence d'images, nous calculons une séquence $y$ de cartes de mesures locales de mouvement quantifiées. La distribution de cooccurrences temporelles $\Gamma(y)$ associée à $y$ est une matrice $\{\Gamma(\nu,\nu'|y)\}_{(\nu,\nu')\in\Lambda^2}$ définie par :

$$\Gamma(\nu,\nu'|y) = \sum_{k=1}^{K} \sum_{p\in\mathcal{R}} \delta(\nu - y_k(p)) \cdot \delta(\nu' - y_{k-1}(p)) \quad (1)$$

où $\mathcal{R}$ est le support spatial des cartes de la séquence $y$ et $K+1$ la longueur de la séquence.

Étant donné un modèle $\mathcal{M}$ spécifié par l'ensemble de ses potentiels $\{\Psi_{\mathcal{M}}\}_{(\nu,\nu')\in\Lambda^2}$, la vraisemblance conditionnelle de la séquence $y$ conditionnellement au modèle $\mathcal{M}$ est simplement calculée à partir du produit scalaire $\Psi_{\mathcal{M}} \bullet \Gamma(y)$ entre les potentiels $\Psi_{\mathcal{M}}$ et les cooccurrences temporelles

$\Gamma(y)$ :

$$P_{\mathcal{M}}(y) = \frac{1}{Z} \exp\left[\Psi_{\mathcal{M}} \bullet \Gamma(y)\right] \quad (2)$$

où le produit scalaire $\Psi_{\mathcal{M}} \bullet \Gamma(y)$ est défini par :

$$\Psi_{\mathcal{M}} \bullet \Gamma(y) = \sum_{(\nu,\nu') \in \Lambda^2} \Psi_{\mathcal{M}}(\nu,\nu') \cdot \Gamma(\nu,\nu'|y) \quad (3)$$

L'intérêt de ces modèles réside dans le fait que la constante de normalisation $Z$ est connue explicitement et qu'elle est indépendante du modèle $\mathcal{M}$. Ceci permet donc d'avoir une connaissance exacte et complète de la loi $P_{\mathcal{M}}(y)$.
D'autre part, l'estimation des potentiels du modèle $M$ relatif à une séquence donnée $z$ est réalisée au sens du maximum de vraisemblance et se déduit simplement de la distribution $\Gamma(z)$ de cooccurrences temporelles de la manière suivante :

$$\Psi_{\widehat{\mathcal{M}}}(\nu,\nu') = \ln\left(\frac{\Gamma(\nu,\nu'|z)}{\sum_{\nu'' \in \Lambda} \Gamma(\nu'',\nu'|z)}\right) \quad (4)$$

Par conséquent, l'utilisation de ces modèles statistiques de mouvement se révèle très simple puisqu'il ne requiert que le calcul de distributions de cooccurrences temporelles et de produits scalaires entre l'ensemble des potentiels spécifiant les modèles et ces distributions de cooccurrences temporelles.

# 4 Segmentation en plages homogènes selon un critère d'activité de mouvement

La première méthode proposée pour la création de résumés de vidéos repose sur la segmentation temporelle de la vidéo en plages homogènes au sens d'un critère d'activité de mouvement. Nous considérons initialement un découpage temporel fin de la séquence vidéo en micro-séquences. Nous associons alors un modèle statistique d'activité de mouvement à chaque micro-séquence. La phase de segmentation temporelle s'appuie sur une technique de classification hiérarchique appliquée à cet ensemble de micro-séquences. Elle exploite un calcul de similarité du contenu dynamique entre micro-séquences définie à partir d'une distance entre modèles statistiques. En outre, nous considérons une contrainte temporelle afin de ne fusionner que des segments temporellement connexes.

## 4.1 Exploitation des modèles statistiques d'activité de mouvement

Étant donné une séquence vidéo, nous calculons tout d'abord la séquence correspondante de cartes de mesures locales de mouvement quantifiées. Comme l'illustre la figure 1, nous opérons une subdivision de la séquence vidéo en micro-séquences de deux cartes de mouvement successives, chaque carte étant calculée à partir de deux images successives. Deux cartes successives est la taille minimale requise des micro-séquences, puisque nous évaluons des cooccurrences temporelles pour définir les modèles d'activité de mouvement. De plus, nous mettons à profit une segmentation temporelle de la vidéo en plans [1] pour ne pas prendre en compte des micro-séquences qui chevauchent une transition entre plans. Finalement, nous disposons d'une succession de micro-séquences $\{y_t = (y_t^0, y_t^1)\}_{t\in[1,T]}$ de deux cartes de mouvement. $T$ est le nombre total de micro-séquences.
Nous déterminons alors une représentation du contenu dynamique associée à chaque micro-séquence $y_t$ par le biais d'un modèle statistique d'activité de mouvement. Nous évaluons tout d'abord la distribution de cooccurrences temporelles $\Gamma(y_t)$ puis, à partir de la relation (4), nous estimons le modèle d'activité de mouvement associé $\mathcal{M}_t$. Nous disposons donc finalement d'une suite de distributions de cooccurrences temporelles $\{\Gamma(y_t)\}_{t\in[1,T]}$ et d'une suite de modèles d'activité de mouvement $\{\mathcal{M}_t\}_{t\in[1,T]}$.
L'utilisation de ces modèles d'activité de mouvement nous permet de définir une mesure de similarité du contenu dynamique entre micro-séquences à partir d'une approximation de la divergence de Kullback-Leibler entre modèles statistiques [3]. Nous considérons en pratique une version symétrisée de cette divergence. Pour deux instants $t$ et $t'$, la mesure de similarité $D_{KL}(t,t')$ entre les deux micro-séquences associées $y_t$ et $y_{t'}$ s'exprime finalement simplement à partir du produit scalaire entre les différences des potentiels des modèles statistiques d'activité de mouvement et des distributions de cooccurrences temporelles :

$$D_{KL}(t,t') = \left[\Psi_{\mathcal{M}_t} - \Psi_{\mathcal{M}_{t'}}\right] \bullet \left[\Gamma(y_t) - \Gamma(y_{t'})\right] \quad (5)$$

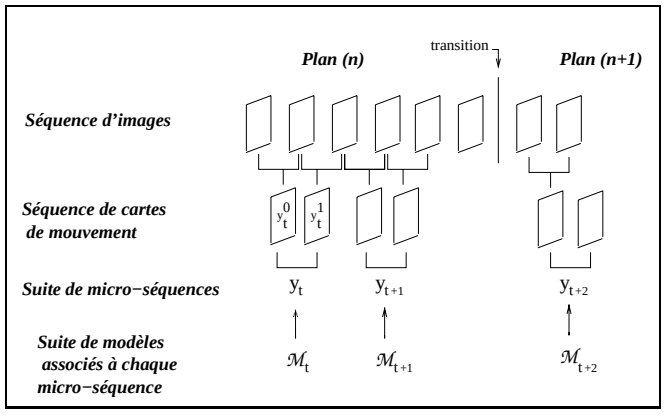


FIG. 1 – *Décomposition de la séquence vidéo en micro-séquences de longueur* 2 *et modélisation statistique de l'activité de mouvement dans chacune d'entre elles.*

## 4.2 Classification temporelle ascendante hiérarchique

Afin d'extraire des plages temporelles homogènes au sens de l'activité de mouvement, nous appliquons une approche de classification hiérarchique ascendante [2] à l'ensemble

des micro-séquences $\{y_t\}_{t\in[1,T]}$. Elle consiste à fusionner itérativement les groupes de micro-séquences les plus similaires au sens de l'activité de mouvement. De plus, nous imposons une contrainte temporelle pour ne fusionner que des groupes de micro-séquences connexes temporellement. De cette manière, nous tenons compte des variations temporelles du contenu dynamique dans le document traité et nous respectons également son organisation temporelle.
Pour appliquer cette technique, il est nécessaire de définir une mesure de similarité entre des groupes de micro-séquences. Pour ce faire, nous exploitons directement la mesure de similarité $D_{KL}$ entre micro-séquences. En fait, la mesure de similarité entre deux plages temporelles successives est définie comme la distance maximale entre les modèles d'activité de mouvement associés aux micro-séquences formant chaque plage. Plus précisément, la distance $D_{KL}(P, P')$ au sens de l'activité de mouvement entre deux plages $P$ et $P'$ successives correspondant respectivement aux groupes de micro-séquences $\{y_t\}_{t\in P}$ et $\{y_{t'}\}_{t'\in P'}$, est évaluée de la manière suivante :

$$D_{KL}(P, P') = \max_{t\in P, t'\in P'} D_{KL}(t, t') \qquad (6)$$

Rappelons que ce calcul de similarité se ramène à la simple évaluation d'un produit scalaire entre les potentiels des modèles statistiques d'activité de mouvement et les distributions de cooccurrences temporelles des quantités locales de mouvement.
En pratique, la procédure de classification hiérarchique opère de la manière suivante. Initialement, chaque micro-séquence de la vidéo traitée forme un groupe distinct. À l'itération $k$, nous disposons d'un ensemble $N^k$ de plages temporelles successives $(P_1, ...., P_{N^k})$. Nous déterminons la paire de plages à regrouper $(P_{i^*}, P_{i^*+1})$ selon le critère suivant :

$$(P_{i^*}, P_{i^*+1}) = \arg \min_{i\in[1,N_k-1]} D_{KL}(P_i, P_{i+1}) \qquad (7)$$

Après la fusion des deux groupes de micro-séquences $P_{i^*}$ et $P_{i^*+1}$, nous obtenons une nouvelle plage provisoirement notée $P_{(i^*,i^*+1)}$. Il est alors nécessaire de mettre à jour deux distances entre plages : celle entre la classe créée et la précédente, soit $D_{KL}(P_{i^*-1}, P_{(i^*,i^*+1)})$, celle entre la classe créée et la suivante, soit $D_{KL}(P_{(i^*,i^*+1)}, P_{i^*+2})$. Ces deux calculs ne requièrent pas en fait d'évaluer toutes les similarités entre modèles concernés. En effet, les distances entre les modèles de la plage précédente et ceux du début de la nouvelle plage (correspondant à $P_{i^*}$) ont déjà été calculées à l'étape précédente, de même vis à vis de la plage suivante.
Cette procédure de classification hiérarchique nous permet finalement de construire une représentation arborescente de la séquence traitée qui fournit une indication sur la variabilité du contenu dynamique dans le document traité. Ainsi, plus les niveaux auxquels des micro-séquences sont regroupées est bas dans cette structure hiérarchique, plus les contenus dynamiques de celles-ci sont proches. La sélection d'un niveau donné dans cet arbre fournit une segmentation temporelle de la vidéo en plages homogènes au sens de l'activité de mouvement. Le choix de ce niveau est fonction du nombre de plages à retenir visé, nombre relié à la durée du résumé à créer.

### 4.3 Représentation du contenu d'une plage

Afin de visualiser les résultats de la segmentation temporelle de la vidéo en plages homogènes ainsi obtenue, nous sélectionnons un instant particulier pour représenter le contenu de chaque plage extraite. L'obtention du résumé vidéo proprement dit peut se faire en prenant un nombre fixé d'images avant et après les instants clés sélectionnés et en les concaténant.
Pour déterminer un instant représentatif de chaque plage temporelle, nous exploitons à nouveau les modèles statistiques d'activité de mouvement. Étant donné une plage $P$ correspondant à un ensemble de micro-séquences $\{y_t\}_{t\in P}$, nous sélectionnons l'instant représentatif $t^*$ correspondant au modèle d'activité de mouvement le plus proche des caractéristiques de mouvement "moyennes" sur la plage $P$. Ceci revient à rechercher le modèle $\mathcal{M}_{t^*}$ qui minimise l'écart à tous les autres modèles dans la plage $P$ :

$$t^* = \arg \min_{t\in P} \sum_{t'\in P} D_{KL}(t, t') \qquad (8)$$

De cette manière, nous visualisons les résultats de la segmentation temporelle de la vidéo en plages homogènes au sens du mouvement sous la forme d'une succession d'images représentatives du contenu de chaque plage extraite.

## 5 Sélection de plages d'intérêt

La deuxième approche proposée peut être vue de deux manières. Elle peut former une alternative à la méthode précédente décrite au paragraphe 4. Elle peut être également présentée comme complémentaire à cette dernière et alors opérer sur les plages homogènes extraites par la première méthode. Cette deuxième approche ne vise pas à représenter l'intégralité de la vidéo numérique, mais à extraire les seuls segments temporels déclarés d'intérêt. À partir du découpage initial de la vidéo en micro-séquences (premier type d'utilisation) ou en plages (second type d'utilisation), il s'agit d'évaluer la pertinence de chaque micro-séquence de relever du résumé de vidéo à créer. Cette sélection de plages d'intérêt est formulée comme un problème d'inférence bayésienne. Nous adoptons une modélisation par chaîne de Markov cachée (CMC) à deux états, "pertinent" et "non pertinent". Nous définirons plus tard ce critère de pertinence.
Nous exploitons à nouveau les modèles statistiques d'activité de mouvement pour définir les probabilités des observations associées à la chaîne de Markov cachée. D'autre part, il est également possible d'intégrer des informations *a priori*, par exemple, sur le pourcentage à retenir d'images ou la durée moyenne d'une plage pertinente.

### 5.1 Modélisation par CMC

De la même manière que dans le paragraphe 4, nous considérons des séquences de cartes de mouvement successives.

Nous cherchons cette fois à attribuer à chaque micro-séquence une étiquette binaire qualifiant la pertinence ou non de son contenu. Cette suite temporelle d'étiquettes peut être considérée comme la réalisation d'une variable temporelle cachée $X$ à deux états représentant l'intérêt ou l'absence d'intérêt de chaque micro-séquence de la vidéo considérée. Nous avons :

$$X = (X_1, ..., X_t, ..., X_T)$$

où $T$ est le nombre de micro-séquences, et où les $X_t$ sont à valeur dans un espace $\mathcal{X}$ à 2 états : $\mathcal{P}$ ("pertinent") ou $\mathcal{N}$ ("non pertinent").
Les observations à travers lesquelles nous évaluons la pertinence d'une micro-séquence sont les informations de nature dynamique captées par les distributions de cooccurrences temporelles des cartes de mesures locales de mouvement $\{y_t\}_{t\in[1,T]}$. Nous définissons donc le processus aléatoire observé $Y$ comme étant :

$$Y = (Y_1, ..., Y_t, ..., Y_T)$$

La modélisation adoptée repose sur plusieurs hypothèses :

1. $(X_t)$ est une chaîne de Markov ergodique du premier ordre sur un espace à deux états, $X_t$ qualifiant la pertinence de la micro-séquence à l'instant $t$. Ceci signifie que la pertinence de la micro-séquence de l'instant $t$ ne dépend que de celle de la micro-séquence précédente, et ce de façon homogène :

$$\begin{cases} P(X_t/X_{t-1}, ...X_1) = P(X_t/X_{t-1}) \\ \\ P(X_t/X_{t-1}) \text{ ne dépend pas de t.} \end{cases} \quad (9)$$

2. Les variables $Y_t$ sont indépendantes conditionnellement aux variables $X_t$, et de lois conditionnelles homogènes dans le temps, c.a.d. :

$$\begin{cases} P(Y_1, ..., Y_t/X_1, ..., X_T) &= \prod_{t=1....T} P(Y_t/X_t) \\ \\ P(Y_t = y/X_t = e_j) &= b_j(y) \end{cases} \quad (10)$$

Ces liens de dépendance sont résumés dans le graphe 1.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\cdots$ | $\rightarrow$ | $X_{t-1}$ | $\rightarrow$ | $X_t$ | $\rightarrow$ | $X_{t+1}$ | $\rightarrow$ | $\cdots$ |
| | | $\downarrow$ | | $\downarrow$ | | $\downarrow$ | | |
| $\cdots$ | | $Y_{t-1}$ | | $Y_t$ | | $Y_{t+1}$ | | $\cdots$ |

TAB. 1 – *Graphe de dépendance entre variables*

Nous définirons ultérieurement les probabilités conditionnelles des observations $b_j(y)$ à partir des lois $P_{\mathcal{M}}(y)$ des modèles statistiques d'activité de mouvement.

## 5.2 Maximisation au sens du MAP

Une approche possible pour estimer $X$, bien fondée statistiquement, est le recours à une technique d'inférence bayésienne. L'intérêt d'une telle approche est la possibilité d'introduire explicitement des connaissances *a priori*. La méthode d'estimation retenue est la maximisation de la probabilité *a posteriori* (MAP), ce qui revient à rechercher l'élément $x^* \in \mathcal{X}^T$ vérifiant :

$$x^* = \arg\max_{x\in\mathcal{X}^T} P(X = x|Y = y) \quad (11)$$

où $y$ représente l'ensemble des observations de mouvement.
D'après la règle de Bayes, cela revient à :

$$x^* = \arg\max_{x\in\mathcal{X}^T} P(Y = y|X = x)P(X = x) \quad (12)$$

où $P(Y|X)$ est la vraisemblance conditionnelle des observations, et $P(X)$ la probabilité *a priori* sur $X$.
Nous utilisons l'algorithme de Viterbi [6, 11], pour résoudre le problème de l'estimation au sens du MAP (équation 11) des états de la chaîne de Markov $(X_t)$. Au temps $t$, pour chaque état $i$, nous avons à considérer tous les chemins dans la séquence d'états qui expliquent les $t$ premières observations et dont le dernier état est $X_t = i$, et nous en déterminons le meilleur chemin, c'est-à-dire la suite des $(\hat{x}_1, ..., \hat{x}_{t-1}) = \hat{S}_t(i)$ qui vérifie :

$$\hat{S}_t(i) = \arg \underbrace{\max_{x_1,...,x_{t-1}} P(y_1, ..., y_t, x_1, ..., x_{t-1}, x_t = i)}_{\delta_t(i)} \quad (13)$$

Cet algorithme récursif repose sur la relation de récurrence sur $\delta_t(i)$ suivante :

$$\delta_t(i) \quad = \quad b_i(y_t) \ \max_j P_{ji} \ \delta_{t-1}(j) \quad (14)$$

La suite optimale aboutissant à $x_t = i$ est donnée par :

$$\hat{S}_t(i) = \arg\max_j P_{ji} \ \delta_{t-1}(j) \quad (15)$$

Ainsi, il nous suffit de définir les probabilités conditionnelles $\{b_i(y_t)\}_{i\in\{\mathcal{P};\mathcal{N}\},t\in[1,T]}$ et les probabilités de transition de la chaîne de Markov pour estimer la suite d'états qui réalise le MAP.

## 5.3 Caractérisation de la pertinence par les modèles d'activité de mouvement

Afin de définir les probabilités conditionnelles des observations, nous associons à chaque état caché considéré des caractéristiques d'activité de mouvement. Nous avons traité pour l'instant le cas le plus simple où chaque état n'est décrit que par un modèle d'activité de mouvement. Nous désignons par $\mathcal{M}_{\mathcal{P}}$ le modèle statistique d'activité de mouvement associé à l'état "pertinent" (étiquette $\mathcal{P}$) et par $\mathcal{M}_{\mathcal{N}}$ le modèle statistique d'activité de mouvement associé à l'état

"non pertinent" (étiquette $\mathcal{N}$). La loi conditionnelle des observations est alors donnée par :

$$\left\{ \begin{array}{lcl} b_{\mathcal{P}}(y_t) & = & P(Y_t = y_t / X_t = \mathcal{P}) \quad = \quad P_{\mathcal{M}_{\mathcal{P}}}(y_t) \\ b_{\mathcal{N}}(y_t) & = & P(Y_t = y_t / X_t = \mathcal{N}) \quad = \quad P_{\mathcal{M}_{\mathcal{N}}}(y_t) \end{array} \right. \tag{16}$$

Nous exploitons la formulation exponentielle des lois $P_{\mathcal{M}_{\mathcal{P}}}(y_t)$ et $P_{\mathcal{M}_{\mathcal{N}}}(y_t)$ donnée par la relation (2). Nous obtenons ainsi :

$$\left\{ \begin{array}{lcl} b_{\mathcal{P}}(y_t) & = & \frac{1}{Z} \exp\left[\Psi_{\mathcal{M}_{\mathcal{P}}} \bullet \Gamma(y_t)\right] \\ b_{\mathcal{N}}(y_t) & = & \frac{1}{Z} \exp\left[\Psi_{\mathcal{M}_{\mathcal{N}}} \bullet \Gamma(y_t)\right] \end{array} \right. \tag{17}$$

Le calcul de ces probabilités conditionnelles se réduit donc à la simple évaluation du produit scalaire entre les potentiels des modèles associés aux étiquettes $\mathcal{N}$ et $\mathcal{P}$ et les distributions de cooccurrences temporelles évaluées sur la micro-séquence $y_t$. En particulier, il n'est pas nécessaire de stocker les cartes de mesures locales de mouvement. Il nous suffit de disposer des informations utiles au calcul des termes $b_{\mathcal{P}}(y_t)$ et $b_{\mathcal{N}}(y_t)$, c.a.d. la distribution de cooccurrences temporelles $\Gamma(y_t)$.

Pour déterminer les modèles $\mathcal{M}_{\mathcal{P}}$ et $\mathcal{M}_{\mathcal{N}}$ représentatifs des propriétés dynamiques devant caractériser les étiquettes "pertinent" et "non pertinent", nous avons opté pour une méthode adaptative fonction du contenu de la séquence vidéo traitée. Nous associons à chaque micro-séquence une mesure d'entropie de son contenu dynamique. Pour une micro-séquence $y_t$, nous estimons le modèle d'activité de mouvement associé $\mathcal{M}_t$ à partir de la relation (4). La mesure d'entropie du contenu dynamique dans la micro-séquence $y_t$ résulte d'une approximation de l'entropie de la loi $P_{\mathcal{M}_t}$ [3]. Cette mesure d'entropie $\mathcal{H}(\mathcal{M}_t)$ est donnée par :

$$\mathcal{H}(\mathcal{M}_t) \quad \approx \quad -\Psi_{\mathcal{M}_t} \bullet \Gamma(y_t) \tag{18}$$

Nous sélectionnons alors comme modèle associé à l'étiquette "pertinent" le modèle $\mathcal{M}_t$ pour lequel la valeur de l'entropie $\mathcal{H}(\mathcal{M}_t)$ est maximale. Inversement, le modèle $\mathcal{M}_t$ pour lequel la valeur de l'entropie $\mathcal{H}(\mathcal{M}_t)$ est minimale est pris comme modèle de référence pour l'étiquette "non pertinent", c.a.d. :

$$\left\{ \begin{array}{l} \mathcal{M}_{\mathcal{P}} = \arg \max\limits_{\{\mathcal{M}_t\}_{t\in[1,T]}} \mathcal{H}(\mathcal{M}_t) \\ \mathcal{M}_{\mathcal{N}} = \arg \min\limits_{\{\mathcal{M}_t\}_{t\in[1,T]}} \mathcal{H}(\mathcal{M}_t) \end{array} \right. \tag{19}$$

Le modèle $\mathcal{M}_{\mathcal{P}}$ correspond alors à une situation de forte activité de mouvement alors que le modèle $\mathcal{M}_{\mathcal{N}}$ traduit un contenu dynamique de faible activité.

## 5.4 Contraintes sur la loi de la chaîne de Markov

À travers la loi *a priori* sur $X$, nous pouvons introduire une contrainte sur le pourcentage $\tau$ désiré d'images sélectionnées pour créer le résumé vidéo, ce qui revient à spécifier la durée du résumé vidéo. Celle-ci peut s'exprimer directement sur la mesure invariante $\pi$ de la chaîne de Markov $(X_t)$ par les équations suivantes :

$$\left\{ \begin{array}{lcl} \pi(\mathcal{P}) & = & \tau \\ \pi(\mathcal{N}) & = & 1-\tau \end{array} \right. \tag{20}$$

Par définition de la mesure invariante, on a :

$$\begin{array}{rcl} \pi\, P_X & = & \pi \\ [\tau\ 1-\tau] \left[ \begin{array}{cc} \alpha & 1-\alpha \\ 1-\beta & \beta \end{array} \right] & = & [\tau\ 1-\tau] \end{array} \tag{21}$$

avec $\alpha = P(X_t = \mathcal{P} / X_{t-1} = \mathcal{P})$ et $\beta = P(X_t = \mathcal{N} / X_{t-1} = \mathcal{N})$.

De (21), nous pouvons déduire la relation suivante entre les paramètres $\alpha$ et $\beta$ :

$$\alpha = 1 - \frac{1-\tau}{\tau}\,(1-\beta) \tag{22}$$

Étant donné un rapport $\frac{1-\tau}{\tau}$, il nous reste alors qu'un des deux paramètres *a priori* de la chaîne de Markov à fixer. Nous obtenons de plus un encadrement de ces paramètres à partir des contraintes $0 < \alpha < 1$ et $0 < \beta < 1$, à savoir :

$$1 - \frac{\tau}{1-\tau} < \beta < 1 \qquad 1 - \frac{1-\tau}{\tau} < \alpha < 1 \tag{23}$$

Les intervalles définis par ces relations sont plus ou moins réduits suivant la valeur de $\tau$ :

- pour $\tau < 0.5$, $\alpha \in [0,1]$ et $\beta$ sera dans un intervalle d'autant plus petit que $\tau$ sera proche de $0$.
- pour $\tau = 0.5$, $\alpha = \beta \in [0,1]$.
- pour $\tau > 0.5$, $\beta \in [0,1]$ et $\alpha$ sera dans un intervalle d'autant plus petit que $\tau$ sera proche de $1$.

À cette contrainte sur le nombre d'images représentatives retenues, nous pouvons ajouter un *a priori* sur la durée d'une plage d'intérêt. En effet, les zones d'intérêt que nous cherchons à extraire correspondent à des situations dynamiques qui s'étendent nécessairement sur plusieurs images successives. Notons $D$ la variable aléatoire relative à la durée d'une plage intéressante. X étant une chaîne de Markov, $D$ suit la loi géométrique suivante :

$$P(D = d) = \alpha^{d-1}(1-\alpha)$$

Son espérance $E[D]$ est donnée par :

$$E[D] = \frac{1}{1-\alpha} \tag{24}$$

Cette équation fournit une relation directe entre le paramètre $\alpha$ de transition de la chaîne de Markov et la durée moyenne $E[D]$ d'une plage pertinente au sens de l'activité de mouvement.

Finalement, à partir de ces deux contraintes (pourcentage d'images retenues et longueur moyenne d'un plage), nous pouvons déterminer les valeurs des paramètres $\alpha$ et $\beta$ de la chaîne de Markov.

FIG. 2 – *Séquence "Avengers" - Visualisation des images médianes de chaque plan de l'extrait considéré.*

# 6 Résultats expérimentaux

## 6.1 Etude de la séquence "Avengers"

Les résultats que nous présentons sont relatifs à un extrait de la série télévisée "*Chapeau Melon et Bottes de Cuir*" ou "*Avengers*"en anglais. Cet extrait est constitué de 3946 images et de 74 plans. La figure 2 fournit un aperçu de cette séquence en donnant l'image médiane de chaque plan.
Cette séquence constitue un document intéressant du point de vue de la création de résumés de vidéos car elle contient des plans de contenus variés en terme d'activité de mouvement. Elle comprend des scènes de faible activité relatives à des plans fixes dans le couloir d'un hôtel, des situations de gros plans sur les personnages qui alternent avec des séquences de poursuite de voitures avec des profondeurs de champ différentes. Enfin, les trois derniers plans de cette séquence correspondent au générique.

## 6.2 Résultats de la segmentation en plages homogènes

La figure 3 fournit une illustration des résultats de la segmentation de la séquence "Avengers" en plages homogènes au sens du mouvement. L'objectif est de retenir au total vingt plages et nous visualisons les images sélectionnées pour représenter chaque plage. Le résultat obtenu est satisfaisant dans le sens où les moments importants de la vidéo sont représentés : personnages en gros plan arrêtés ou en mouvement, scènes statiques, poursuite de voitures, gros-plans sur des voitures en action,... Notons que les deux plans correspondant à un homme en action, dont les activités de mouvement se distinguent de celles des plages précédentes et suivantes (scène dans le couloir caractérisée par une activité de mouvement plus faible, poursuite de voitures caractérisée par une activité de mouvement plus forte), ont été justement assimilés à deux plages homo-

FIG. 3 – *Résultats de la segmentation temporelle en plages homogènes pour la séquence "Avengers". Nous donnons les images sélectionnées pour représenter le contenu des vingt plages extraites.*

gènes et sont représentés par deux images. Finalement, la visualisation des vingt images représentatives des plages temporelles extraites fournit un aperçu concis et informatif sur le contenu de la séquence "Avengers". Ceci montre que le choix de baser l'analyse de la vidéo sur la notion d'activité de mouvement est judicieux, et que la méthode statistique proposée est capable, alors qu'elle exploite des informations de "bas niveau", de fournir un résumé d'une certaine "sémantique", utile et suffisante pour appréhender les éléments essentiels de la vidéo en vue, par exemple, d'une sélection des documents enregistrés à des fins de visualisation ou d'archivage.

### 6.3 Résultats de la sélection de plages pertinentes

Nous avons également appliqué la technique de sélection de plages d'intérêt à la séquence "Avengers". Nous fournissons à la figure 4 les images médianes des deux micro-séquences associées aux modèles d'activité de mouvement sélectionnés pour représenter les états "pertinent" et "non pertinent". Le premier correspond à une micro-séquence montrant une voiture en action au premier plan, le second est une micro-séquence du générique sans mouvement. Pour éviter que les plages retenues ne soient trop longues, nous fixons $\alpha$ à $0.1$. Nous souhaitons obtenir un pourcentage $\tau$ d'images sélectionnées de l'ordre de $10\%$. La contrainte sur le pourcentage d'images sélectionnées nous fournit la valeur de $\beta$: $\beta = 0.9$ de par l'équation (22). Nous obte-


FIG. 4 – *Images médianes des deux micro-séquences associées aux modèles d'activité de mouvement sélectionnés pour représenter les états "pertinent" et "non pertinent" pour la séquence "Avengers".*

nons au final une proportion de micro-séquences retenues d'environ $15\%$ pour la séquence "Avengers" (1913 micro-séquences) ce qui est voisin du pourcentage désiré. La figure 5 présente les résultats obtenus sur cette séquence par affichage d'une image sur vingt dans les plages déclarées pertinentes. Les résultats semblent satisfaisants au regard des micro-séquences correspondant aux modèles "pertinent" et "non pertinent". En effet, la grande majorité des plages sélectionnées correspondent des activités de mouvement élevées, et plus particulièrement à la poursuite en voiture. En particulier, aucune micro-séquence correspondant au générique n'a été sélectionnée et la scène dans le couloir n'est représentée que par une seule image.

La visualisation des résultats fournis par les deux méthodes proposées pour la création de résumés de vidéos est effectuée à partir d'un ensemble d'images. Il ne s'agit toutefois que d'une première étape pour évaluer l'intérêt des deux approches statistiques envisagées. Notre objectif reste de construire une séquence vidéo courte, qui constituera le résumé du document original, en exploitant directement l'extraction de plages homogènes ou la sélection de plages d'intérêt. Nous pouvons envisager différentes alternatives : la plus simple consiste à concaténer des micro-segments du document vidéo original de longueur prédéterminée autour des moments extraits.

## 7 Conclusion

Nous avons présenté dans cet article deux méthodes originales pour la création de résumés de vidéos. Elles sont basées sur l'utilisation d'informations de nature dynamique caractérisées par des modèles statistiques d'activité de mouvement. La première méthode repose sur une segmentation temporelle de la vidéo en plages homogènes au sens du mouvement. Elle exploite une technique de classification hiérarchique appliquée à des micro-séquences de la vidéo traitée. La seconde méthode propose une approche complémentaire par extraction de plages d'intérêt. Il s'agit d'une technique d'étiquetage markovien des micro-séquences de la vidéo traitée sur deux états "pertinent" et "non pertinent". Ces deux états sont décrits par des modèles statistiques d'activité de mouvement. Nous avons appliqué ces deux méthodes sur un document réel. Les résultats obtenus démontrent l'intérêt de l'approche envisagée pour la création de résumés de vidéos.
Différentes perspectives se dégagent de cette étude. En ce qui concerne la sélection de plages d'intérêt, nous envisageons de procéder à une estimation en ligne des représentations des contenus dynamiques associés aux états "pertinent" et "non pertinent" par le biais d'un algorithme EM. Par ailleurs, nous cherchons à combiner les deux méthodes présentées dans cet article, en opérant par exemple une détection de plages d'intérêt à partir des résultats de la segmentation temporelle de la vidéo au sens du mouvement. Des évaluations expérimentales sur un corpus plus étoffé sont également prévues.

### Remerciements

La séquence "Avengers" nous a été fournie par l'INA, Département Innovation, Direction de la Recherche.

## Références

[1] P. Bouthemy, M. Gelgon, et F. Ganansia. A unified approach to shot change detection and camera motion characterization. *IEEE Trans. on Circuits and Systems for Video Technology*, 9(7):1030–1044, 1999.

[2] E. Diday, G. Govaert, Y. Lechevallier, et J. Sidi. Clustering in pattern recognition. In *Digital Image Processing*, pages 19–58. J.-C. Simon, R. Haralick, eds, Kluwer edition, 1981.

[3] R. Fablet. Modélisation statistique non paramétrique et reconnaissance du mouvement dans des séquences d'images ;

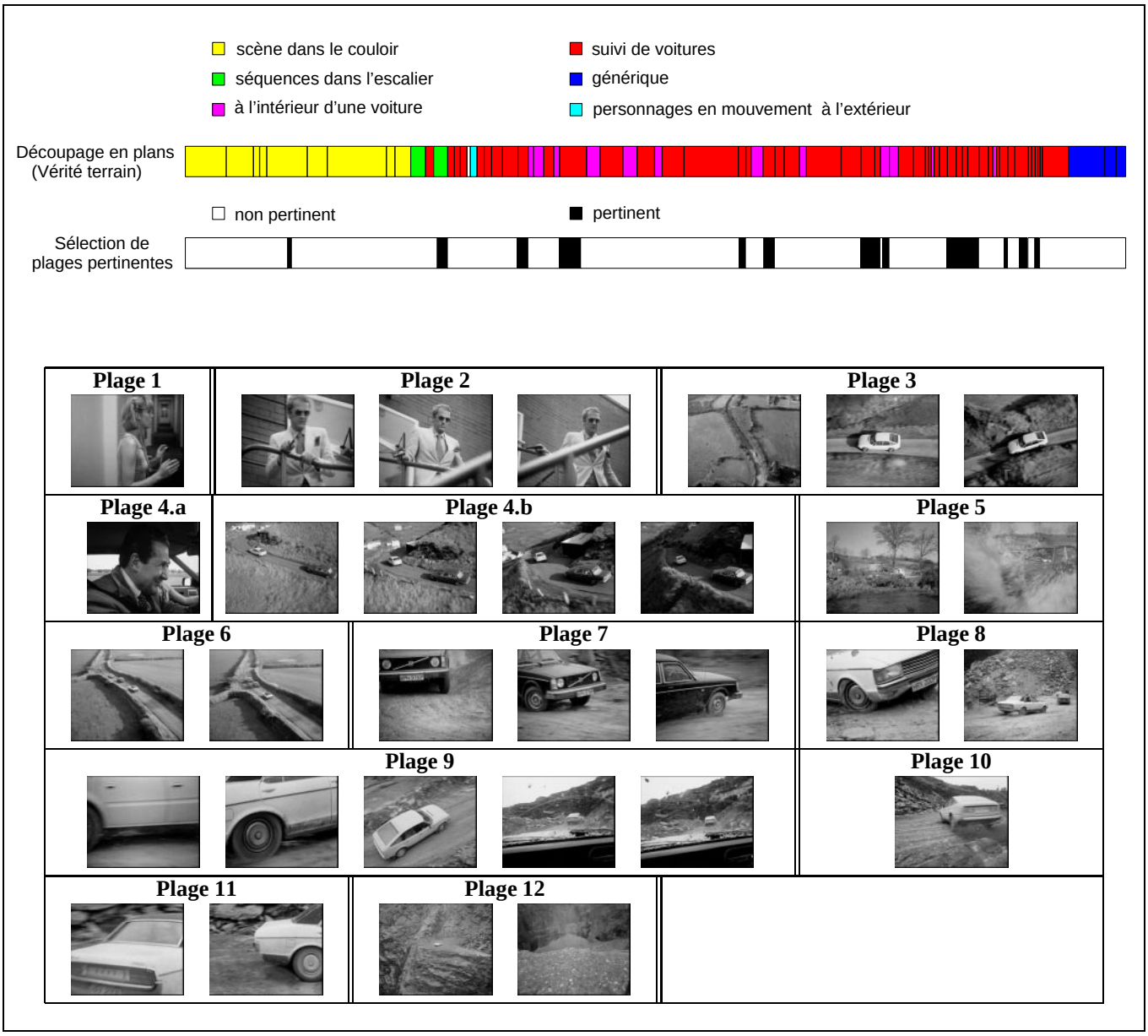


FIG. 5 – *Résultats de la sélection de plages d'intérêt pour la la séquence "Avengers". Les paramètres qui fixent les transitions de la chaîne de Markov sont les suivants :* $\alpha = 0.1$ *et* $\tau = 10\%$ *- Visualisation d'une image sur vingt dans les plages déclarées pertinentes. Le nombre d'images affichées par plage dépend donc de la lonqueur de la plage. Concernant la plage 4, il s'agit en fait de deux plages accolées notées 4.a et 4.b.*

application à l'indexation vidéo. *Thèse Université de Rennes 1, Irisa, No. 2526*, juillet 2001.

[4] R. Fablet and P. Bouthemy. Non-parametric scene activity analysis for statistical retrieval with partial query. *Journal of Mathematical Imaging and Vision*, 14(3):257-270, mai 2001.

[5] R. Fablet, P. Bouthemy, et P. Pérez. Statistical motion-based video indexing and retrieval. In *Proc. of 6th Int. Conf. on Content-Based Multimedia Information Access, RIAO'2000*, pages 602–619, Paris, avril 2000.

[6] G.D. Forney. The Viterbi algorithm. *Proceedings of the IEEE*, 61:268–278, 1973.

[7] R. Lienhart, S. Pfeiffer, et W. Effelsberg. Video abstracting. *Communication of the ACM*, 40(12):55–62, 1997.

[8] D. De Menthon, L.J. Latecki, et A. Rosenfeld. Relevance ranking of video data using hidden Markov model distances and polygon simplification. In *Proc. of 4th Int. Conf. on Visual Information Systems, VISUAL'2000*, pages 49–61, Lyon, novembre 2000.

[9] J. Nam and H. Tewfik. Dynamic video summarization and visualization. In *Proc. of 7th ACM Int. Conf. on Multimedia, ACM Multimedia'99*, pages 53–56, Orlando, novembre 1999.

[10] J.M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4):348–365, 1995.

[11] L.R. Rabiner, C.H. Lee, B.H. Juang, et J.G. Wilpon. HMM clustering for connected word recognition. *Proc. of Int. Conf. on Acoustics, Speech, et Signal Processing, ICASSP'89*, 1:405–408, 1989.