Analyse de la vidéo

Chapitre 3.2 - Segmentation sémantique de la vidéo

3 février 2015

Plan de chapitre

Segmentation par primitive

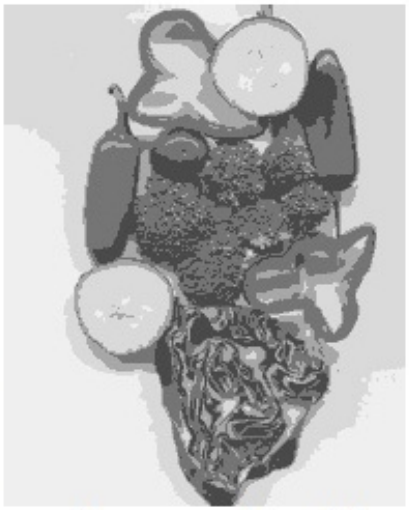
Une **primitive** est définie comme étant une caractéristique ou un pariage de caractéristique propre à l'image ou la vidéo. Par exemple :

- La position ;
- L'intensité ;
- La couleur ;
- La texture ;
- Les vecteurs de mouvement estimé ;
- L'orientation ;
- Fenêtre, bloc ;
- ...

Segmentation par primitive

Le choix de la primitive utilisée influencera le résultat de la segmentation :

Image originale



K-moyenne (intensité)

K-moyenne (couleur)

On peut utiliser un appariement de primitive optimale à la vidéo afin d'avoir la segmentation la plus fidèle possible.

Segmentation par primitive

Définition générale

On incorpore l'information de l'estimation du mouvement à l'intensité pour former un **vecteur de primitive spatio-temporelle**.

Espace de primitive (EP) Représentation selon laquelle chaque primitive est représentée par un point.

Vecteur de primitive (VP) Association de différentes primitives.

Distance de primitive (DP) Distance donnée par la mesure de similarité entre différents VP dans l'EP.

Segmentation par primitive

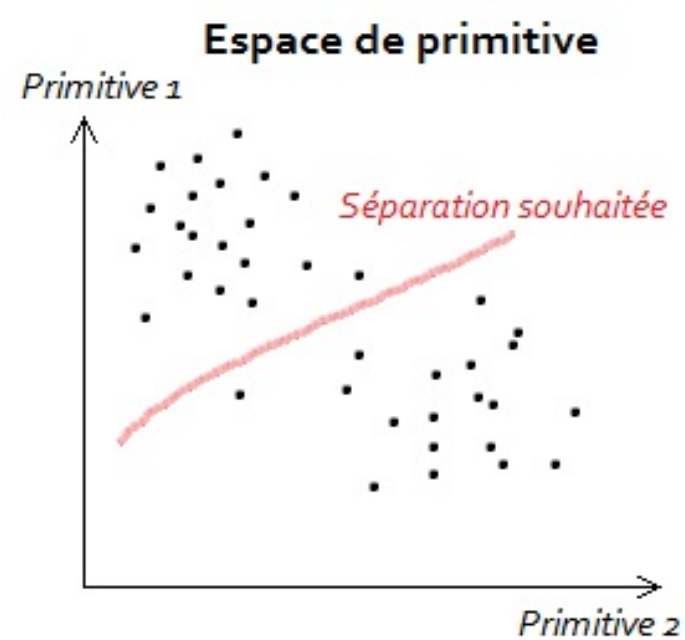
Exemple de cas particulier

Le choix de la primitive à segmenter influencera la **mesure de similarité** et l'**algorithme de partitionnement** choisi.

Vecteur de primitive Angle $\tan^{-1}\left(\frac{V_y}{V_x}\right)$ et norme du vecteur de vitesse $\sqrt{V_x^2 + V_y^2}$.

Mesure de similarité Distance euclidienne

Algorithme de partition K-Moyenne



Vecteur de primitive Norme du vecteur de vitesse $\sqrt{V_x^2 + V_y^2}$.

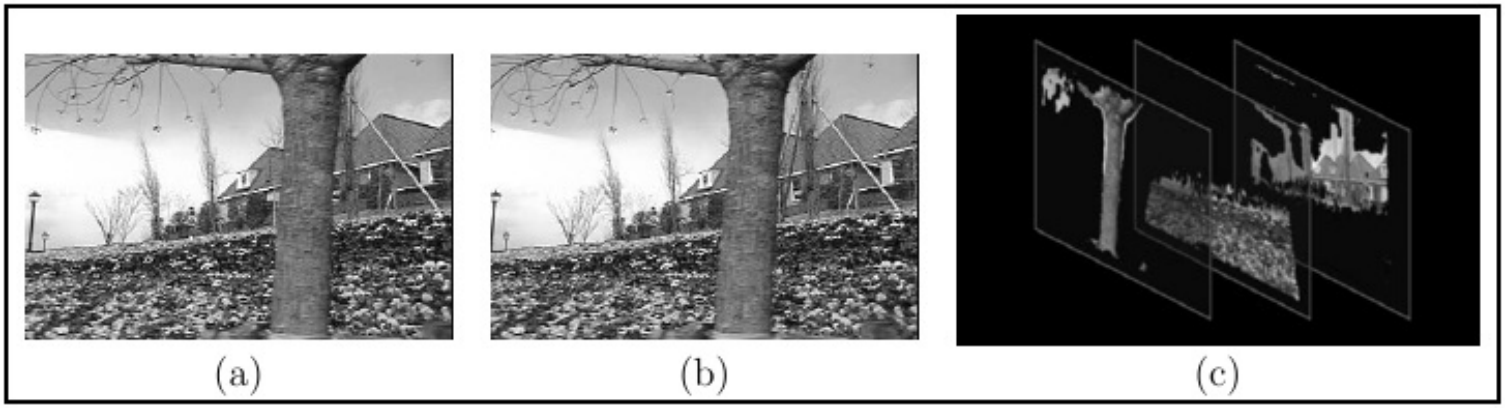
Mesure de similarité Distance euclidienne

Algorithme de partition K-Moyenne

La segmentation par couche

La **segmentation en couches** de la vidéo est définie comme étant **la partition des frames de la vidéo en régions (objets) de profondeurs différentes**.

Une couche correspond à un **ensemble de pixels possédant le même type de mouvement**.



(a) et (b) deux frames successifs de la vidéo, (c) les couches de mouvement estimées.

On suppose alors qu'il n'y a pas ou très peu d'objets en mouvement.

La segmentation par couche

Comment ça fonctionne ?

On estime le mouvement dominant global de la scène (GME) [1].

Suite à l'estimation, on teste la conformité d'un pixel à ce mouvement global suivant l'équation suivante :

$$\begin{array}{rll} \text{Conforme} & si & |I(x,y,t) - I(T((x,y)|p),t-1)| < T \\ \text{Non-Conforme} & sinon & \end{array} \quad (1)$$

On résume les étapes ainsi :

1. Prendre tous les pixels des images comme région ;
2. Estimer le flot global pour toute la région ;
3. Évaluer l'erreur pour tous les pixels de la région par Eq.1 ;
4. Éliminer les pixels ayant une erreur trop élevée ;
5. Répéter les étapes 2-4 pour les pixels exclus.

1. Voir Chapitre 2.3

La segmentation par couche

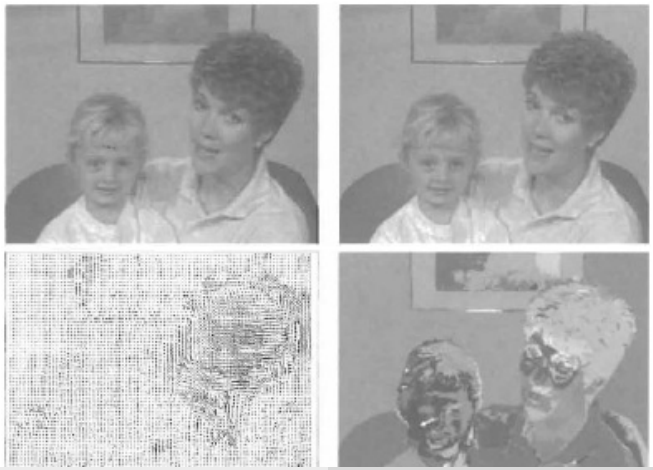
Note :

Couche $\neq$ Un seul objet !

- Plusieurs objets à la même profondeur ;
- Un objet composé de parties ayant leurs propres mouvements.

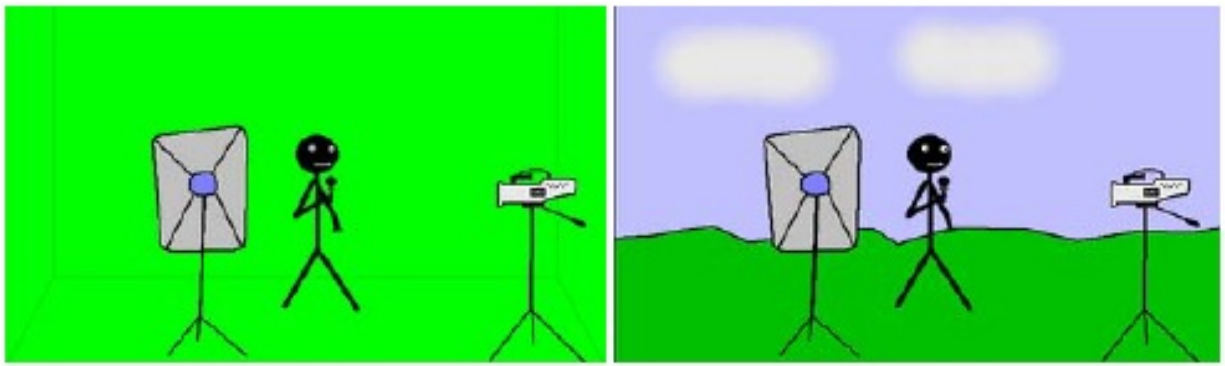
Il n'est pas garanti que la segmentation en couches constitue un seul est unique objet en mouvement.

De plus, un objet peut être composé de plusieurs parties ayant chacune un mouvement différent, et par conséquent appartenant à des couches différentes.



La segmentation par chroma-keying

Le chroma-keying est une acquisition d'une vidéo en face d'une couleur (key color), en général verte ou bleu.



La segmentation par chroma-keying

La couleur est choisie de manière à ce qu'**elle n'apparaisse pas sur l'objet à capter** et **très différente de la tinte de couleur de la peau**.

On veut donc segmenter la couleur d'arrière-plan, en évitant les inconvénients suivants :

1. Ombrage dans l'arrière-plan ;
2. Composition de la couleur de fond dans les objets à extraire.

La segmentation par chroma-keying

Ombrage dans l'arrière-plan

Pour contourner le problème d'**ombrage dans l'arrière-plan**, on fait un **apprentissage** de l'arrière-plan [2] (mixture de gaussienne, k-moyenne, ...).

On observera alors :

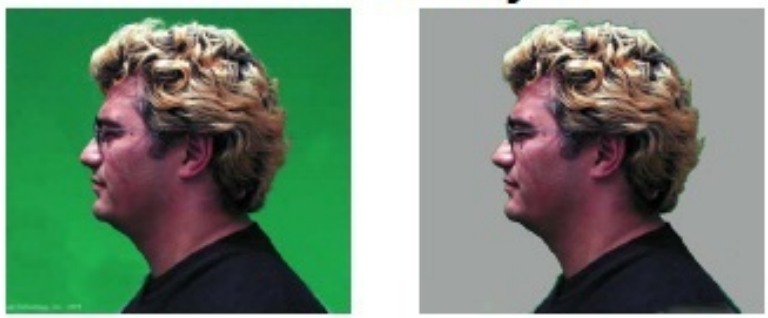
1. Extraction non volontaire de l'ombrage des objets de l'arrière-plan.
2. Halo de la couleur d'arrière-plan autour de l'objet d'avant-plan.

2. Chapitre 3.1

La segmentation par chroma-keying

Cas 1 - Segmentation de l'ombrage

Cas 2 - Halo autour des objets



Segmentation with Invisible Keying Signal. Moshe Ben-Ezra. *School of Computer Science and Engineering*

La segmentation par chroma-keying

Halo de la couleur autour de l'objet

Pour contrer l'halo de couleur, on peut appliquer un *blue-spill removal filter*, qui consiste en :

- Appliquer une couleur **inverse** aux bords de l'objet ;
- Appliquer un masque graduel allégé l'effet aux bords.

La segmentation par chroma-keying

Soft chroma-keying

Afin d'éviter le *blue-spill*, on peut atténuer l'effet en calculant une distance sur la couleur à segmenter. Pour de bons résultats :

1. On passe de l'espace RGB à YCbCr afin d'isoler la luminance (Y) des couleurs chromatiques (CbCr) ;
2. On détermine le code (CbCr) de la couleur à segmenter ;
3. On détermine une région autour de cette couleur (code couleur +/- distance), qui nous donne un rayon ;
4. On met ce qui est en dehors du rayon à 1 ;
5. On calcule un rapport du rayon ($\alpha \in [0, 1]$) pour les couleurs à l'intérieur du rayon ;
6. On met la couleur situé pile sur le code couleur à 0 ;
7. On combine nos deux images à l'aide de ce masque.

La segmentation par chroma-keying

Soft chroma-keying

Afin d'éviter le *blue-spill*, on peut atténuer l'effet en calculant une distance sur la couleur à segmenter. Pour de bons résultats :

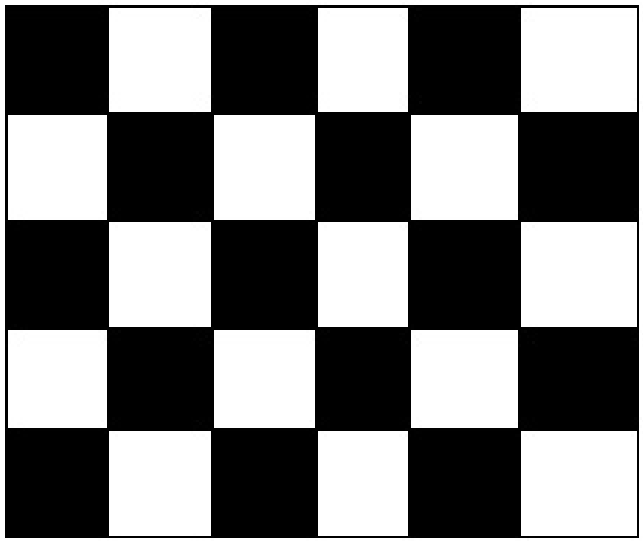
1. On passe de l'espace RGB à YCbCr afin d'isoler la luminance (Y) des couleurs chromatiques (CbCr) ;
2. On détermine le code (CbCr) de la couleur à segmenter ;
3. On détermine une région autour de cette couleur (code couleur +/- distance), qui nous donne un rayon ;
4. On met ce qui est en dehors du rayon à 1 ;
5. On calcule un rapport du rayon ($\alpha \in [0, 1]$) pour les couleurs à l'intérieur du rayon ;
6. On met la couleur situé pile sur le code couleur à 0 ;
7. On combine nos deux images à l'aide de ce masque.

La segmentation par chroma-keying

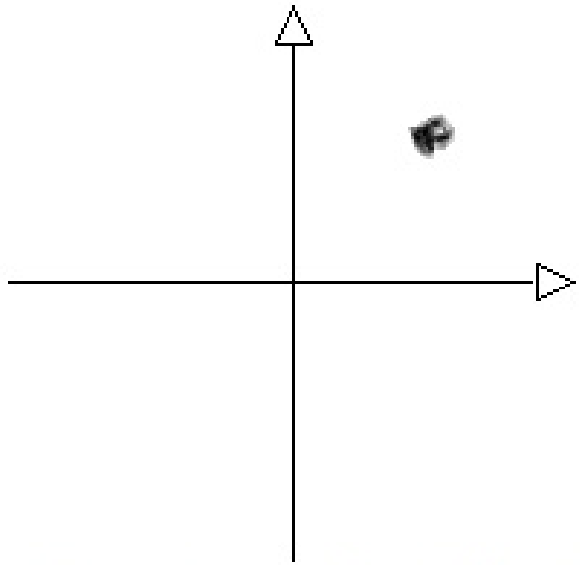
Halo de la couleur autour de l'objet

La segmentation de texture

Il existe plusieurs familles d'approche quant à la **segmentation de texture**. La plus intuitive pour une application du type du chroma-keying utiliserait la tranformée de Fourier.



Texture dans l'espace cartésien



Texture dans l'espace de Fourier (fréquence)

La segmentation de texture

En utilisant une texture **à haute fréquence**, on peut obtenir une bonne segmentation, en supposant que l'on connait le background utilisé :

- On créé un filtre coupe-bande en n'utilisant que la fréquence du background $\mathscr{F}(I_b)$.
- On fait la transformée de Fourier du frame courant $\mathscr{F}(I_t)$.
- On effectue une soustraction dans le domaine de Fourier.
- On fait la transformée de Fourier inverse.

La segmentation de texture

La technique, simpliste, a plusieurs failles :

- On peut éliminer des textures présentes dans les objets ;
- Les contours des objets ne seront pas bien définis ou segmentés (dû à l'obstruction partielle de la texture par l'objet).

On peut éliminer ces artefacts par un post-traitement utilisant le contour des objets comme référence :

- Élimination des petites régions extérieures à l'objet (*floodfill*) ;
- Appliquer un flou pour allégé l'effet aux bords.

La segmentation de texture

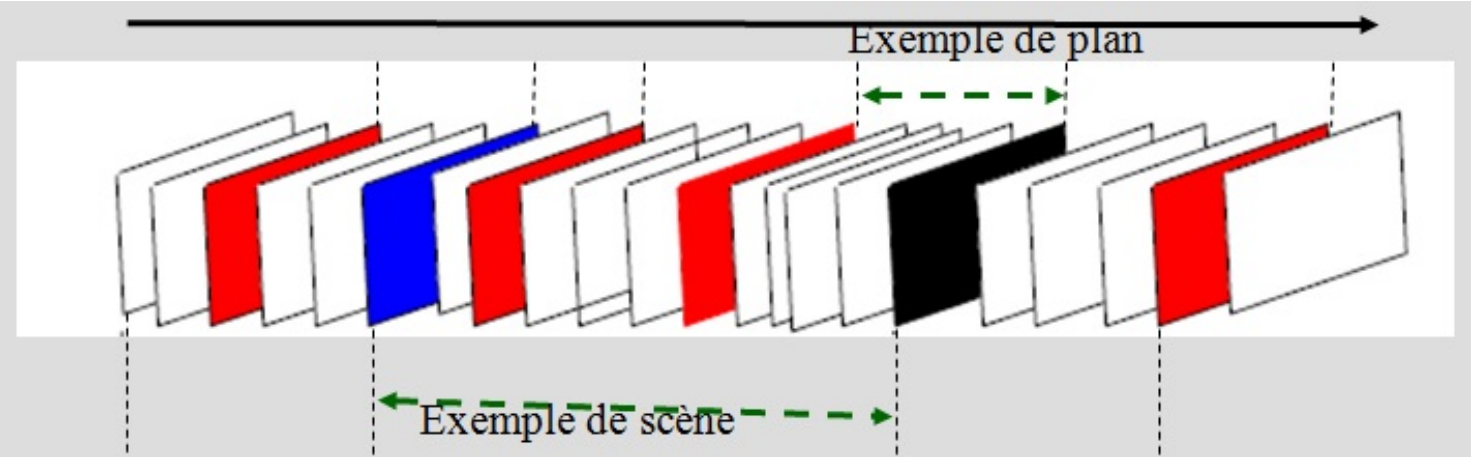
http://ip.hhi.de/imedia_G3/assets/sdk_seq2.jpg

Plan de chapitre

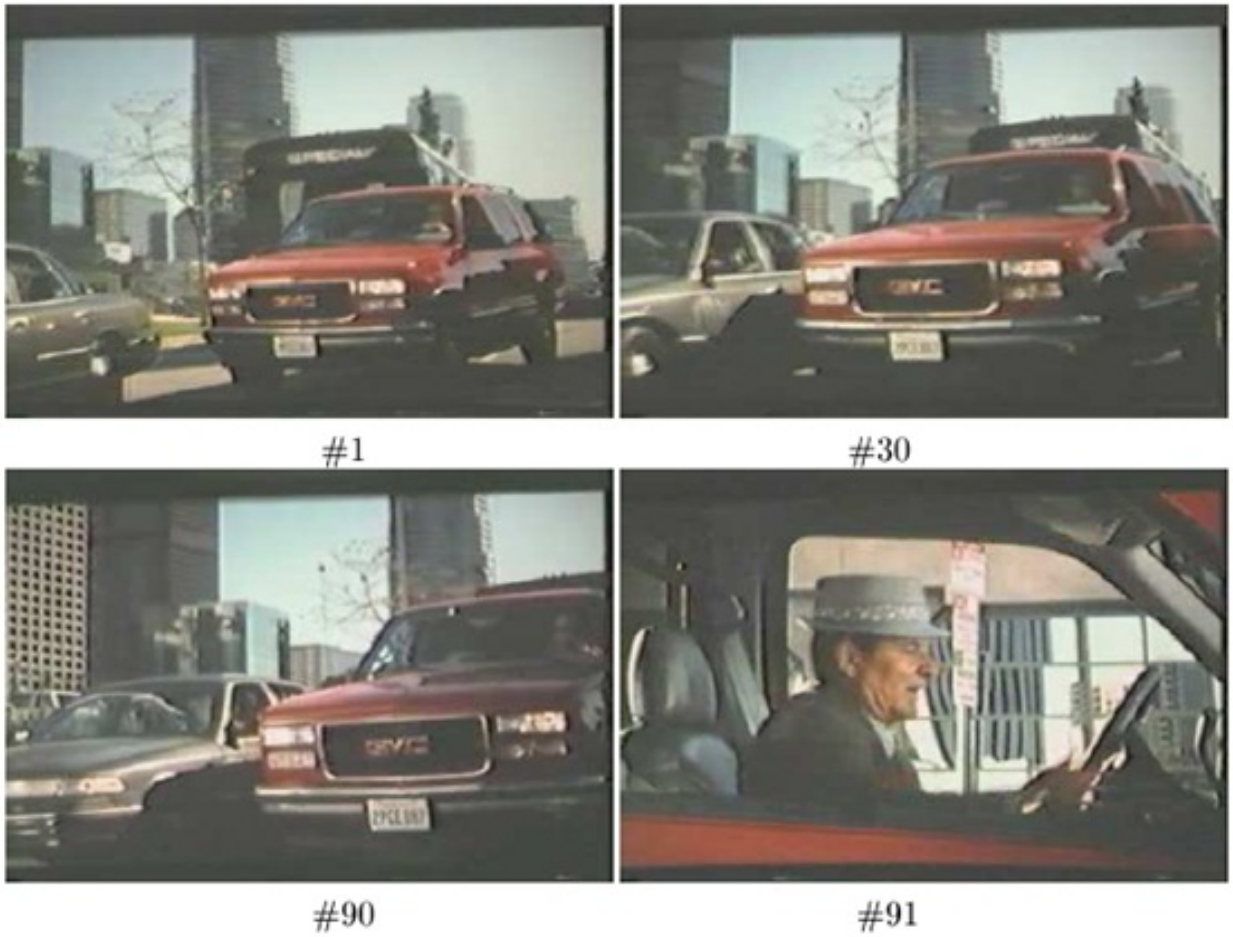
Introduction à la segmentation en plan et en scène

On peut décomposer une vidéo avec une structure hiérarchique, puisque les images qui la compose ont nécessairement un lien sémantique.

Un film peut être décomposé en plusieurs épisodes, et chaque épisode en scènes. Une scène peut être composée d'un ou de plusieurs plans, et un plan de plusieurs frames.



Introduction à la segmentation en plan et en scène



Introduction à la segmentation en plan et en scène

Définitions

- **Plan :** Suite d'images filmées sans interruption temporelle, avec la même caméra
  $\rightarrow$ Unité de base pour le traitement et la manipulation de la vidéo.
- **Scène :** Collection de plans adjacents dans le temps sémantiquement liés.
- **Frame clé (key frame) :** Image de la vidéo représentant grossièrement le contenu visuel du plan.

Introduction à la segmentation en plan et en scène

Les images au voisinage d'une transition sont très différentes
$\rightarrow$ On cherche à repérer les discontinuités dans le flux vidéo.

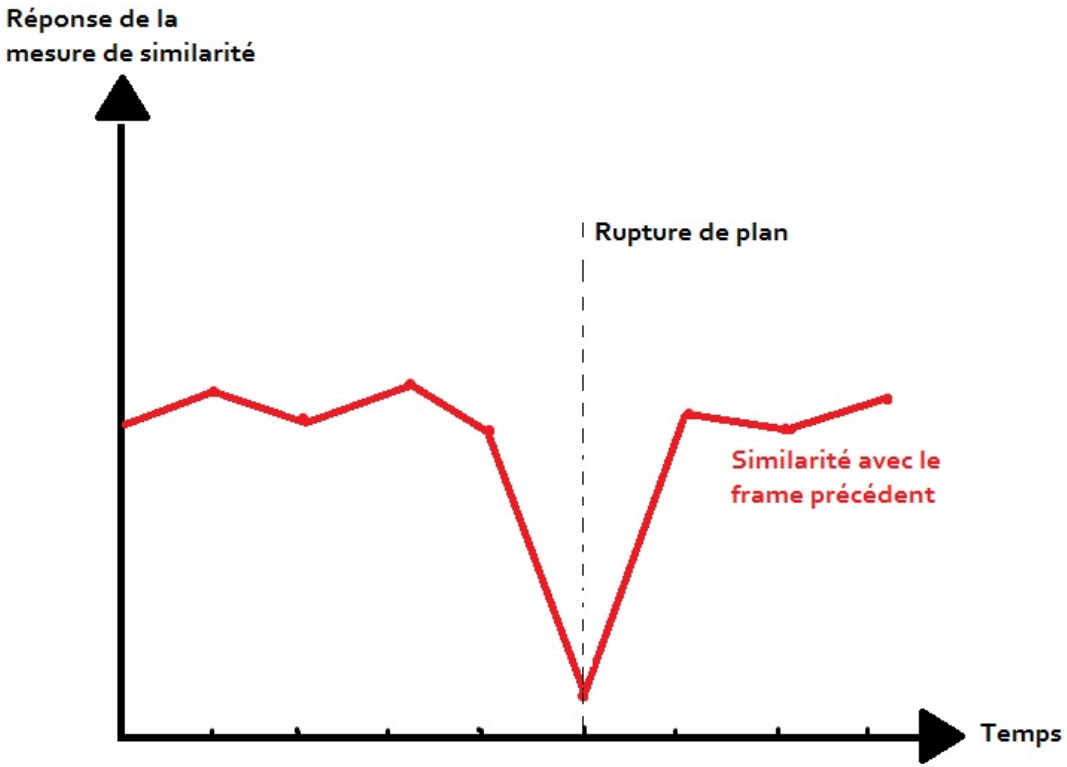
De façon générale, on effectue les étapes suivantes :

1. Extraire une observation (primitive globale) sur chaque image.
2. On définie ensuite une distance (ou mesure de similarité) entre les observations.
3. Entre frame successifs, l'application de cette distance, sur l'ensemble du flux vidéo, produit un signal unidimensionnel.
4. On cherche alors les variations extrêmes (pics du signal) qui correspondent aux instants de faible similarité.

Segmentation en plan et en scène

Détection de changement de plan

La **détection de rupture de plan** est basée sur la segmentation (partitionnement) de la séquence vidéo en **une succession de frames représentant une action spatiale et temporelle continue**.



Segmentation en plan et en scène

Imaginons la technique la plus simple :

Primitive : Les intensités des images.
Mesure de similarité : Les moindres carrés :

$$D(t) = \sum_x \sum_y (I(x,y,t) - I(x,y,t-1))^2 \quad (2)$$

Segmentation en plan et en scène

La technique est très sensible aux :

- Mouvements d'objets ou de caméra ;
- Bruit ;
- Ombrage ;
- Changement d'illumination ;
- ...

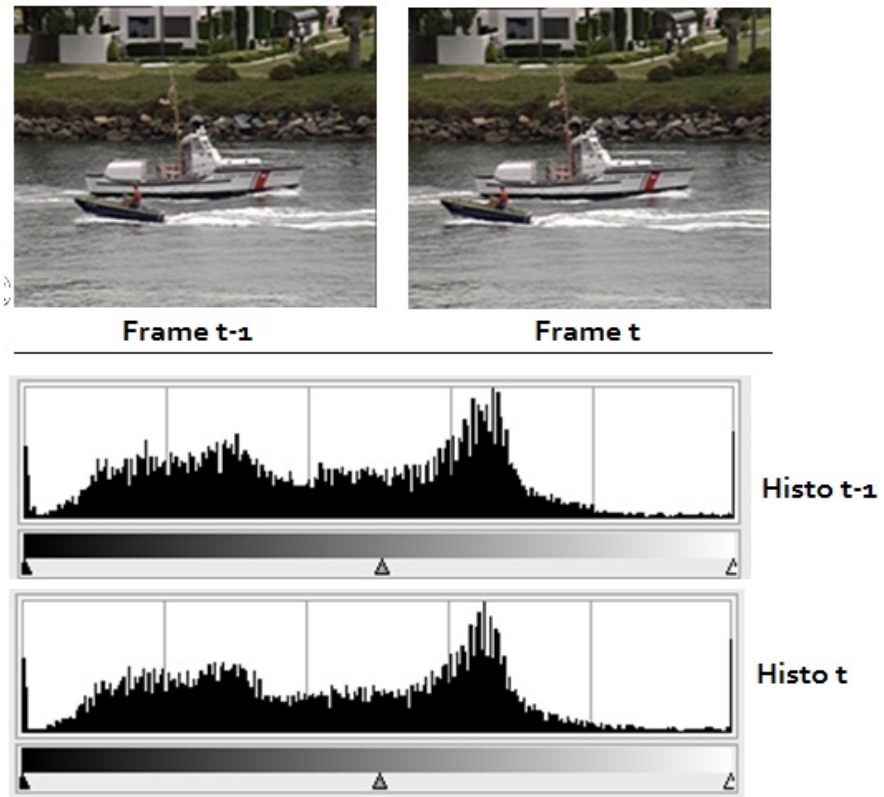
Il faut donc trouver de nouvelles primitives ou mesures plus robustes.

Histogramme

Un plan contient grossièrement les mêmes objets en mouvement sur un fond statique

→ La distribution de l'intensité à l'intérieur des frames d'un même plan est donc semblable.

On représente cette distribution par l'**histogramme d'intensité** des frames.



Histogramme

Avantage de l'histogramme d'intensité :

- **Insensible au mouvement des objets :** Représentation qui ignore les arrangements spatiaux du contenu des frame.
- **Insensible aux petits mouvements de camera :** La représentation est insensible aux secousses de caméra.

Histogramme - Mesure de similarité

Comment calculer une mesure de similarité avec des histogrammes ?

On calcul une **distance** entre les histogrammes normalisés de chaque deux frames successifs de la vidéo et on pose un seuil pour l'une des distances suivantes :

- **Distance Euclidienne :** Une approche classique de différence au carré.

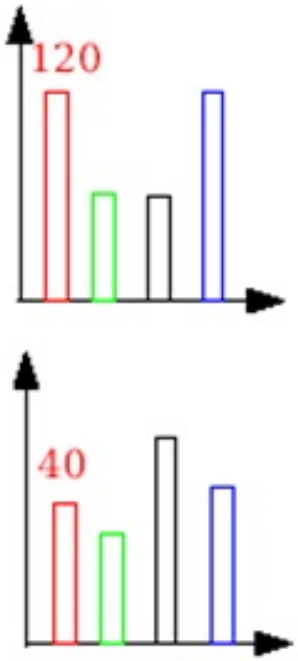
$$D_e(\mathbf{H}_t, \mathbf{H}_{t-1}) = \sum_{u=0}^{m} (H_t(u) - H_{t-1}(u))^2 \quad (3)$$

$m \rightarrow$ *Valeur d'intensité maximale*

Histogramme - Mesure de similarité

- **Intersection d'histogramme :** On calcul la proportion commune aux deux histogrammes.

$$D_i(\mathbf{H}_t, \mathbf{H}_{t-1}) = 1 - \frac{\sum_{u=0}^{m} \min\left(H_t(u), H_{t-1}(u)\right)}{\textit{Nb d'intensité différentes}} \tag{4}$$



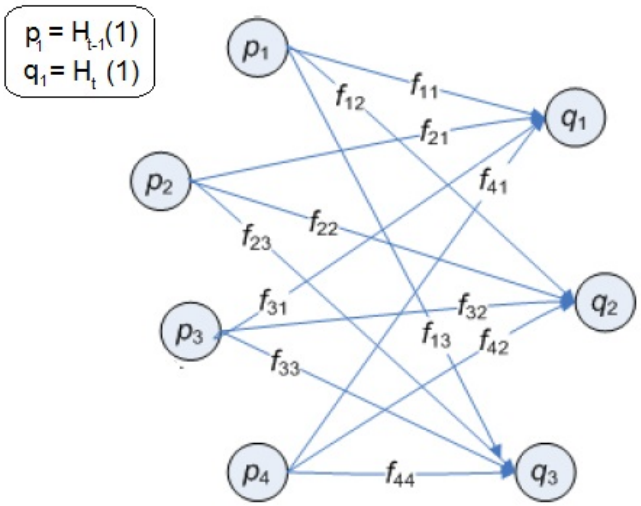
Au moins 40 pixels rouges dans chacune des deux images
\+ Au moins 35 pixels verts dans chacune des deux images
\+ Au moins 34 pixels noirs dans chacune des deux images
\+ Au moins 50 pixels noirs dans chacune des deux images

= 159 pixels ayant une couleur commune dans les deux images

Histogramme - Mesure de similarité

- **Earth mover's distance (EMD) :** Considérant les histogrammes comme l'ensemble de points dans l'espace métrique d'intensité, il faut calculer le cout minimal pour déplacer les données de façon à passer d'un ensemble à une autre ensemble.

Considérons $S_t(i)\{H_t(i), i\}$ une **signature**, où $H_t(i)$ correspond à un poids et i correspond à une position. On dénote $f(i, j)$ le flot d'une position i à j comme étant la valeur des quantités déplacées entre $H(i)$ et $H_{t-1}(j)$.



Histogramme - Mesure de similarité

On définit alors EMD comme étant :

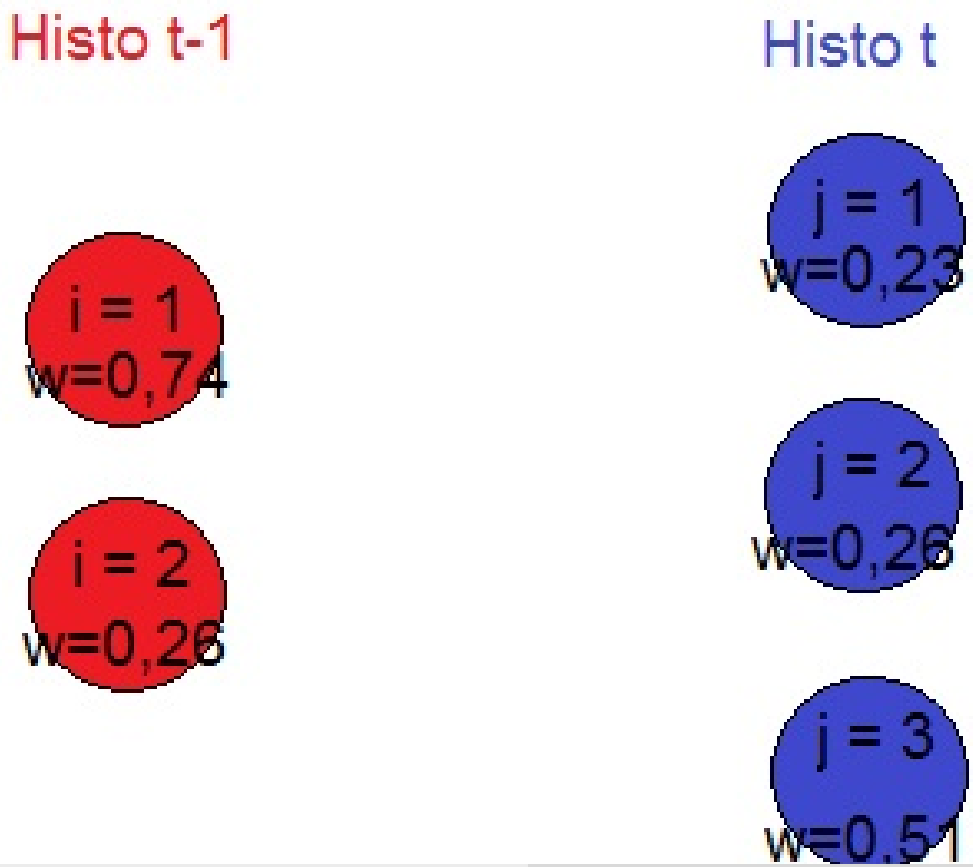
$$
\begin{aligned}
D_{EMD}(\mathbf{H}_t, \mathbf{H}_{t-1}) &= \min_{F=\{f(i,j):\ i\in H_{t-1}, j\in H_t\}} \sum_{i,j} \big(f(i,j)\cdot d(i,j)\big) \qquad (5)\\
d(i,j) &= |i-j|
\end{aligned}
$$

F définissant l'ensemble des flots (ou le flot global de H_{t-1} à H_t). On impose aux flots les contraintes suivantes :

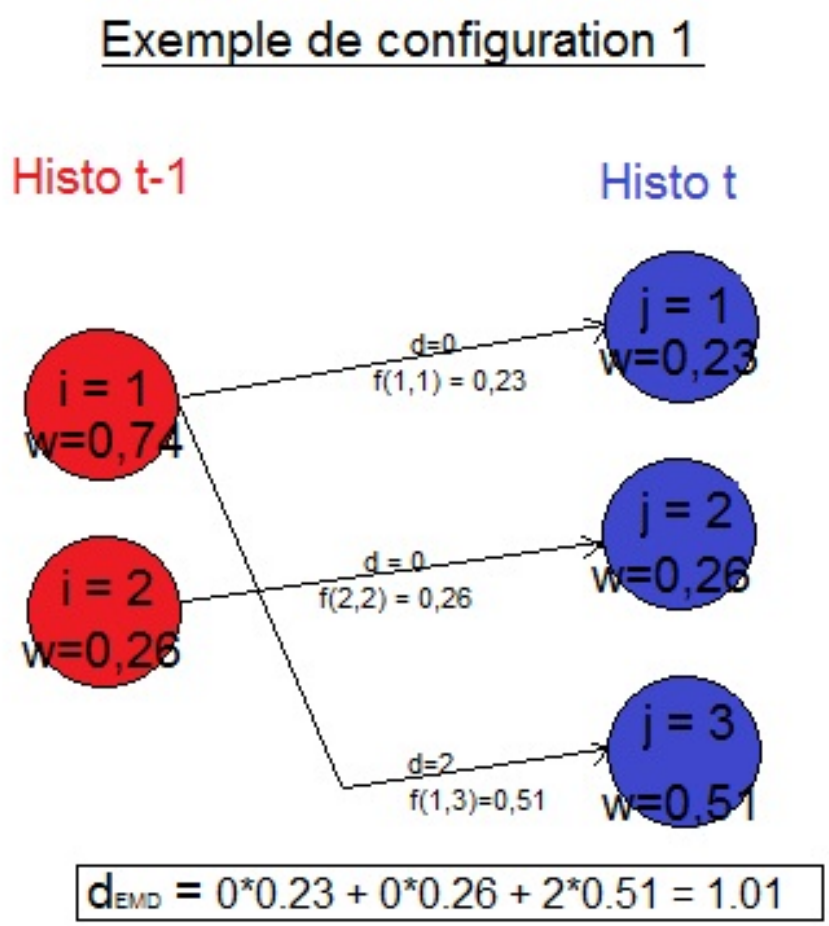
$$
\begin{aligned}
\sum_{j\in H_t} f(i,j) &= H_{t-1}(i), \quad \forall i \in H_{t-1}\\
\sum_{i\in H_{t-1}} f(i,j) &= H_t(j), \quad \forall j \in H_t\\
f(i,j) &\geq 0, \quad \forall i,j
\end{aligned}
$$

Histogramme - Mesure de similarité

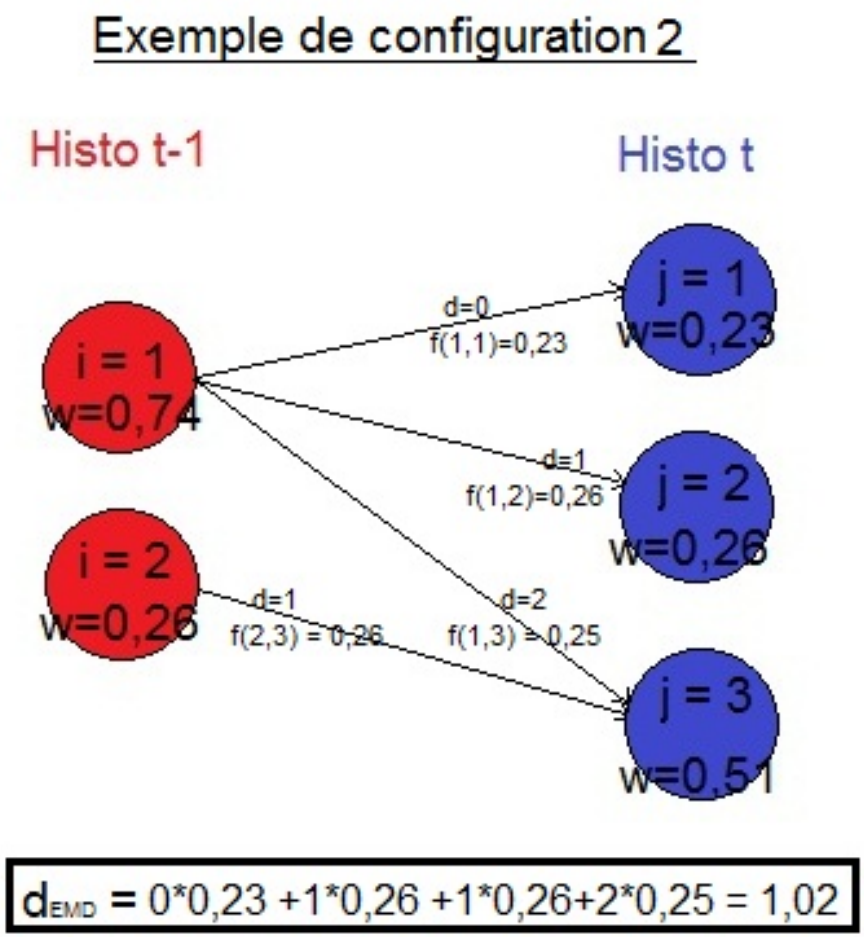
On peut résoudre l'équation par un algorithme ROP linéaire. Le problème est vu comme étant un **problème de transport**, où on doit trouver la meilleure configuration :



Histogramme - Mesure de similarité



Histogramme - Mesure de similarité



Mouvement (flux global)

Dans la méthode utilisant le flot optique comme primitive, on suppose qu'il existe un mouvement global dominant entre chaque deux frames successifs d'un plan.

L'absence d'un mouvement dominant entre deux frames successifs est alors considérée comme un indice de rupture du plan.

$$\begin{array}{rll} \text{Pixel conforme} & \textit{si} & |I(x,y,t) - I(T((x,y)|p),t-1)| < T \\ \text{Pixel non-conforme} & \textit{sinon} & \end{array} \quad (6)$$

Mouvement

On évalue ensuite les quantités suivantes, où S_t définit un ratio moyen, et M_t définit un ration maximal :

$$\text{Rapport moyen } S_t = \sum_{\tau=0}^{t} \left(|\frac{N_\tau}{N}\xi_\tau - m| - \delta \right), \quad (7)$$

$$\text{Rapport maximum } M_t = \max_{0<\tau<t} \{S_\tau\} \quad (8)$$

Où

$$\begin{aligned} N &= \text{Nombre de pixels dans le frame,} \\ N_t &= \text{Nombre de pixels conforme au mouvement estimé,} \\ \delta &= \text{Amplitude de la variation moyenne de la conformité,} \\ m &= \begin{cases} \frac{1}{t-1}\sum_{\tau=0}^{t-1} S_\tau & \text{si } t > 1 \\ \xi_0 & \text{sinon} \end{cases} \end{aligned}$$

Pour un temps t, on détecte une rupture de plan si $|M_t - S_t| > T$.

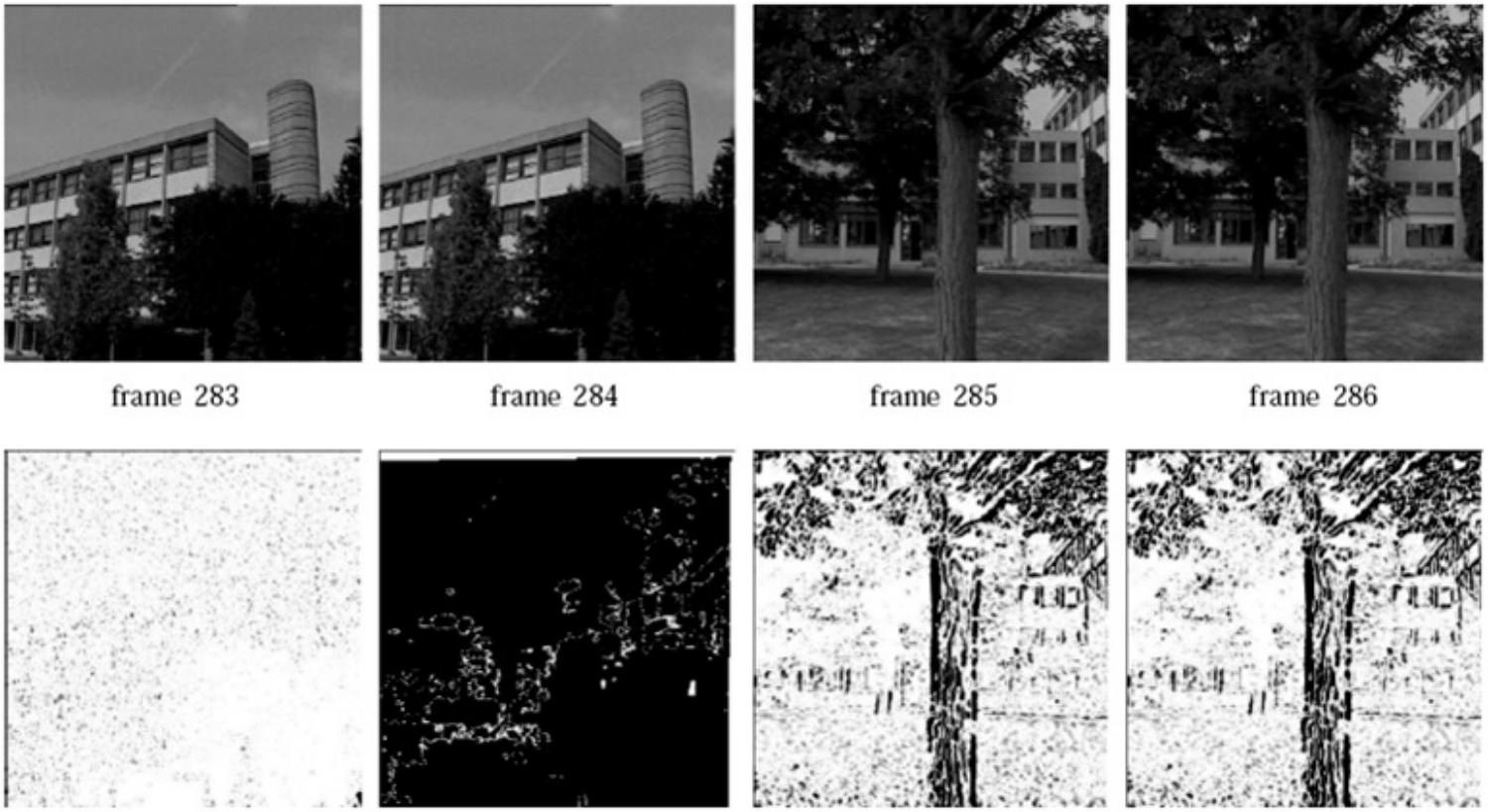
Mouvement

On peut traduire par :

Pour chaque plan, on calcule entre chaque deux frames la proportion de conformité des pixels au mouvement global estimé entre des derniers, et on garde le maximum et la moyenne de ces proportions à travers le plan.

Une rupture est détectée entre deux frames courants, si la proportion des pixels conformes est trop petite par rapport à la moyenne des autres proportions dans les frames précédents.

Mouvement



Pixels conformes au modèle du mouvement en blanc

Image de Bouthemy99

Mouvement

Pixels conformes au modèle du mouvement en blanc

Image de Bouthemy99

Plan de chapitre

Création de résumé

Pourquoi ?

Le contenu de la vidéo peut être accédé de différentes manières.

- **Recherche sur le contenu** $\downarrow\uparrow$: On connait avec précision ce que l'on cherche. On va chercher un plan, une scène, que l'on associera à une vidéo.
- **Consultation d'une vidéo** $\uparrow\downarrow$: On cherche une vidéo, une scène, un plan appartenant à un thème vague. On cherche alors parmi des résumés, telle une table des matières.

Dans les deux cas, la représentation ou la recherche se fait par **image-clé**, conçue de façon à optimiser le contenu sémantique de la vidéo.

Création de résumé

Pourquoi ?

On veut fournir des informations pertinentes et concises afin d'aider l'utilisateur à naviguer ou à organiser des fichiers vidéos plus efficacement.

Résumé statique (video summary) Sélectionner les **images les plus représentatives** de la vidéo. Ces images appelées **images clés** se présentent en général sous la forme d'un scénarimage (storyboard)

Résumé dynamique (video skimming) Version courte de la vidéo originale (Ex : construction automatique d'une bande-annonce du film).

Création de résumé statique

On peut regrouper les familles de résumés statiques en quatre catégories distinctes :

1. méthodes reposant sur l'échantillonnage ;
2. méthodes reposant sur les plans ;
3. méthodes reposant sur les scènes ;
4. méthodes autres (autre sémantique, méthode avancée, etc).

Résumé statique par échantillonnage

Dans le **résumé statique par échantillonnage**, il faut choisir les images clés en sous-échantillonnant uniformément ou aléatoirement la séquence originale.

Inconvénient :

- Certaines parties de la vidéo ne seront représentées
- Redondance de certaines images clés avec un contenu similaire.

Cette approche n'a pas d'attache au contenu de la vidéo.

Résumé statique par plan

Dans le **résumé statique par plan**, la détection des plans est réalisée pour mieux ajuster la sélection des images clés au contenu de la vidéo.

Pour ce faire, on va extraire des images-clés des différents plans extraits de la vidéo.

Cas 1 : Première image

On extrait **la première image du plan** comme image clé **ou les première et dernière images du plan**.

- Efficace pour décrire les plans stationnaires où le contenu varie peu.
- Pas de représentation satisfaisante pour les plans avec de forts mouvements de caméra.

Résumé statique par plan

Cas 2 : Utilisation de l'histogramme

Approche 1

La première image du plan est sélectionnée comme image clé. Puis, si la **distance entre l'histogramme** d'intensité de la **dernière image clé** sélectionnée et l'**image courante** est **supérieure à un seuil**, alors l'image courante est la nouvelle image clé.

- Efficace pour décrire les plans stationnaires où le contenu varie peu.
- Pas de représentation satisfaisante pour les plans avec de forts mouvements de caméra.
  $\rightarrow$ On risque de retrouver la dernière image comme image clé en cas de grands mouvements.

Résumé statique par plan

Cas 2 : Utilisation de l'histogramme

Approche 2

Comparaison des images d'un plan suivant leurs **histogrammes d'intensité**, puis réalise le **rassemblement des images en plusieurs groupes**.

Si une image au temps t dépasse un certain seuil de similarité, on crée un nouveau groupe. Seuls **les groupes de taille assez importante sont conservés** et les **images les plus proches du centre de gravité** de chaque groupe sont alors choisies comme images clés.

- Efficace pour décrire les plans dynamiques.
- Représentation satisfaisante pour les plans avec de forts mouvements de caméra.
  $\rightarrow$ On risque de retrouver beaucoup d'images clés superflues.

Résumé statique par scène

Une **scène** est un regroupement de **plans ayant un lien sémantique** entre eux. Nous supposerons ici que les scènes sont préalablement divisées, ou bien que la vidéo n'est qu'une seule scène.

On suppose que **les plans peuvent se ressembler entre eux, mais être disjoints**.

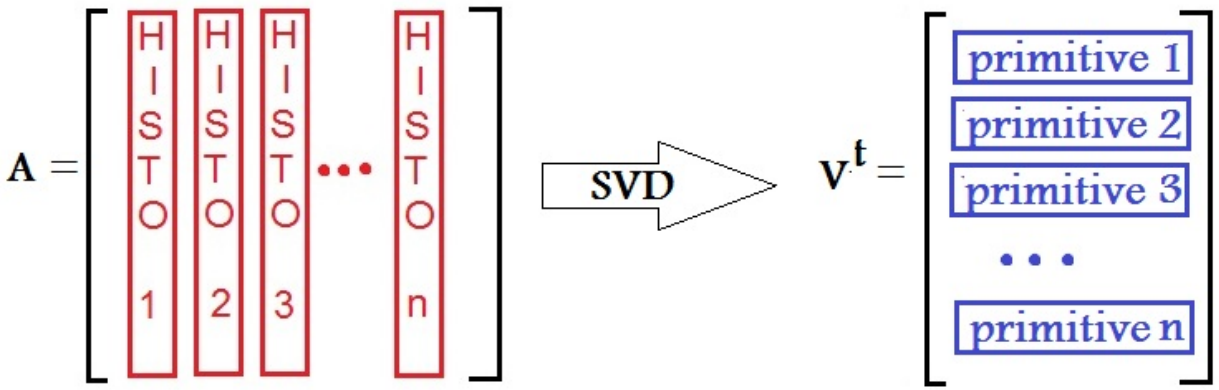
Par exemple, une conversation entre deux personnes donne lieu à plusieurs plans semblables qui, rassemblés, donneront seulement deux plans distincts.

Il faut noter que l'on **perd la logique temporelle** en effectuant un résumé statique par scène !

Résumé statique par scène

L'idée est d'assigner chaque image à un groupe, puis à réunir les groupes les plus similaires de manière itérative jusqu'à un critère d'arrêt.

Une matrice est créée où chaque colonne contient un vecteur caractéristique (l'histogramme) associé à un groupe d'images de la vidéo.



Une SVD est réalisée pour réduire l'espace des caractéristiques.

Étapes en résumé

1. On crée la matrice d'histogramme A avec tous les frames de la scène ;
2. On réduit A avec la SVD pour avoir l'espace de primitive réduit V^T ;
3. **Partitionnement** : On choisie une image-clé de départ, puis
   1. On **classe en ordre** les vecteur-primitives (VP) selon la distance avec le vecteur-primitive de l'image clé S_1 ;
   2. Pour chaque VP non classé, on trouve l'image ayant la distance minimum $d(S_{min})$.
   3. Si $d(S_{min}) < T$, on associe le VP à l'image S_{min}. Sinon, on crée une nouvelle image clée S_c.
   4. On arrête lorsque tous les VP sont associés à une image clé S_c.
4. On retourne les images clés S_c.

Résumé statique par scène

Remarques

- On travaille sur la totalité des frames de la scène.
- Le temps de calcul peut être long.

Solution :

$\rightarrow$ Approche hiérarchique (séquentielle)

Résumé statique par scène

Définition de résumé hiérarchique à deux niveaux

Premier niveau On applique l'approche précédente avec une séquence fixe d'images (20 et plus) ;

Deuxième niveau : On applique l'approche précédente, mais **en utilisant les images clés sorties du niveau précédent**.

Cette approche permet de paralléliser la première étape et de réduire le coût de la décomposition SVD.