

RCP217 – IA pour des données multimédia

Données vidéo

Responsable UE : Marin FERECATU
Conservatoire National des Arts et Métiers (CNAM)
Lab. CEDRIC, Équipe Vertigo (Données Complexes,
Apprentissage et Représentations)
http://cedric.cnam.fr/~ferecatu/

le cnam

Plan de la séance

- Structure d'un document vidéo
- Représentation numérique (codage/décodage)
- Segmentation temporelle
- Caractérisation du mouvement
- Caractérisation du contenu visuel
- Caractérisation sémantique
- Applications (reconnaissance des objets, suivi, etc.)

Structure d'un document vidéo

Document vidéo = combinaison de plusieurs flux d'informations

Deux sources principales combinées :

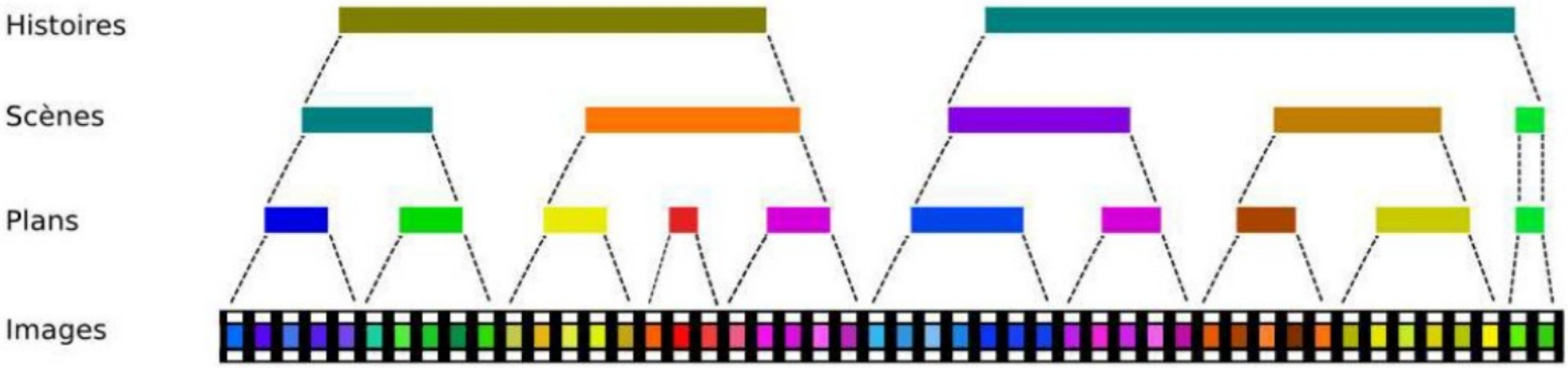
- L'image et le son
  - Synchronisés pour former une histoire
- D'autres sources aussi (télétexte, sous-titres, etc.)

Organisation hiérarchique :

- Séquences de granularité différentes.
- Plusieurs niveaux de structure liés à la donnée vidéo : cadre, plan, scène, histoire, etc.

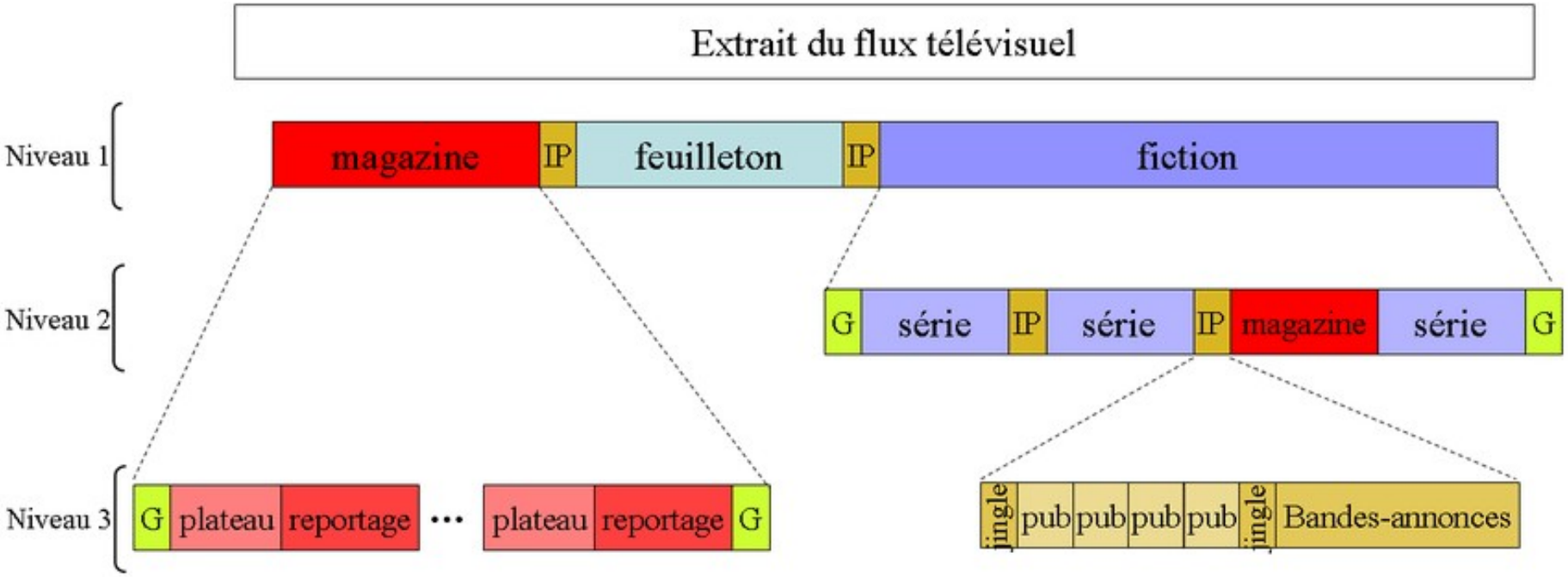
Structure d'un document vidéo

Structure hiérarchique d'un document vidéo :



Structure d'un document vidéo

Structure hiérarchique d'un flux télévisuel [7] :



G : générique

IP : interprogramme

Structure d'un document vidéo

Structure hiérarchique

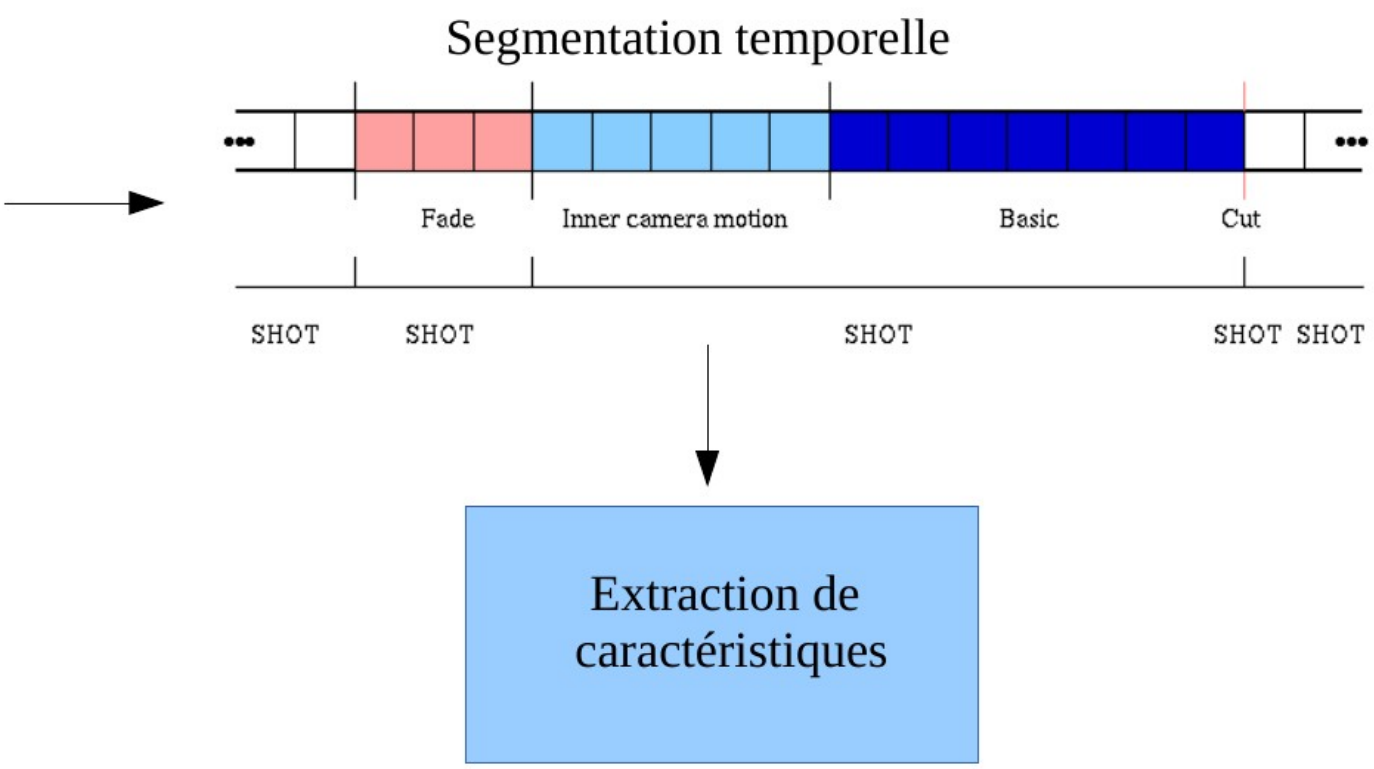
Un plan = une séquence d'images durant laquelle l'acquisition du signal est continue. L'acquisition de ces plans se fait dans une zone limitée de l'espace où se situe et se joue/déroule l'action.

Le montage consiste à regrouper les plans en scènes puis à assembler les scènes pour former un film.

L'analyse du contenu d'une vidéo passe alors par sa segmentation temporelle, tout d'abord en plans puis en scènes.

Structure d'un document vidéo

L'indexation se fait habituellement plan par plan (en anglais : shot).



Structure d'un document vidéo

Le flux visuel :

- Séquence d'images fixes animées sur l'axe temporel à une fréquence de 24 à 30 images/seconde.

Le flux sonore :

- Composé d'un ou plusieurs canaux (mono, stéréo)
- Signal sonore typiquement échantillonné entre 16000 et 48000 kHz.

Structure d'un document vidéo

Méta-description :

- Le contexte de la prise de vue, comme par exemple, la date ou l'auteur
- Données techniques issues de la camera vidéo
- Données de diffusion : langue, date et heure de diffusion, chaîne
- Données d'édition : liste de reportages pour JT, chapitres (DVD) etc.

Structure d'un document vidéo

Flux textuel :

- Provient soit d'un flux séparé, soit il est dérivé des sources audio et visuelle
- Sous-titrage
- Issu de la technologie télétexte,
- Texte mis à disposition pour les mal-entendants
- Les films sur DVD contiennent naturellement des sous-titres, généralement en plusieurs langues
- Analyse des flux audio et visuel (ASR, Automatic Speech Recognition et OCR, Optical Character Recognition)

Structure d'un document vidéo

Flux textuel :

- Aligné avec les séquences vidéos, le texte est une source d'informations riche pour l'indexation vidéos.
- En particulier : l'extraction des entités nommées (ou noms propres) permet de décrire les séquences vidéos par des éléments sémantiques précis tels que les personnes, organisations et lieux géographiques
- Les éléments sémantiques (concepts) et le flux textuel sont fortement corrélés.

Plan de la séance

- Structure d'un document vidéo
- Représentation numérique (codage/décodage)
- Segmentation temporelle
- Caractérisation du mouvement
- Caractérisation du contenu visuel
- Caractérisation sémantique
- Applications (reconnaissance des objets, détection de copies)

Codage du signal vidéo

MPEG-7 (2002) :

- Standard de représentation du contenu des documents
- Indépendant de la technique de codage ou de stockage

MPEG-4 (1998):

- Définit précisément la manière de décrire une scène
- Description scène 3D codée par MPEG-4

MPEG-2 (1994) :

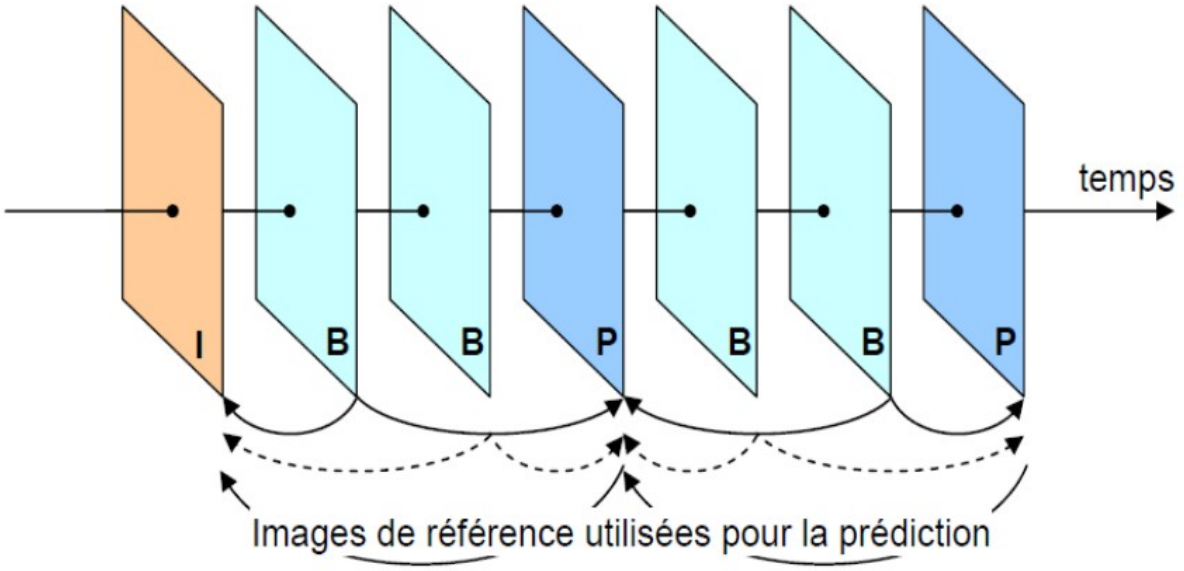
- MPEG-1 et MPEG-2 utilisées pour le codage prédictif (H263)
- DVD, télévision satellite, câble et réseau hertzien (TNT)

Codage du signal vidéo

Codage MPEG : système prédictif sur une séquence

Utilise trois types d'images :

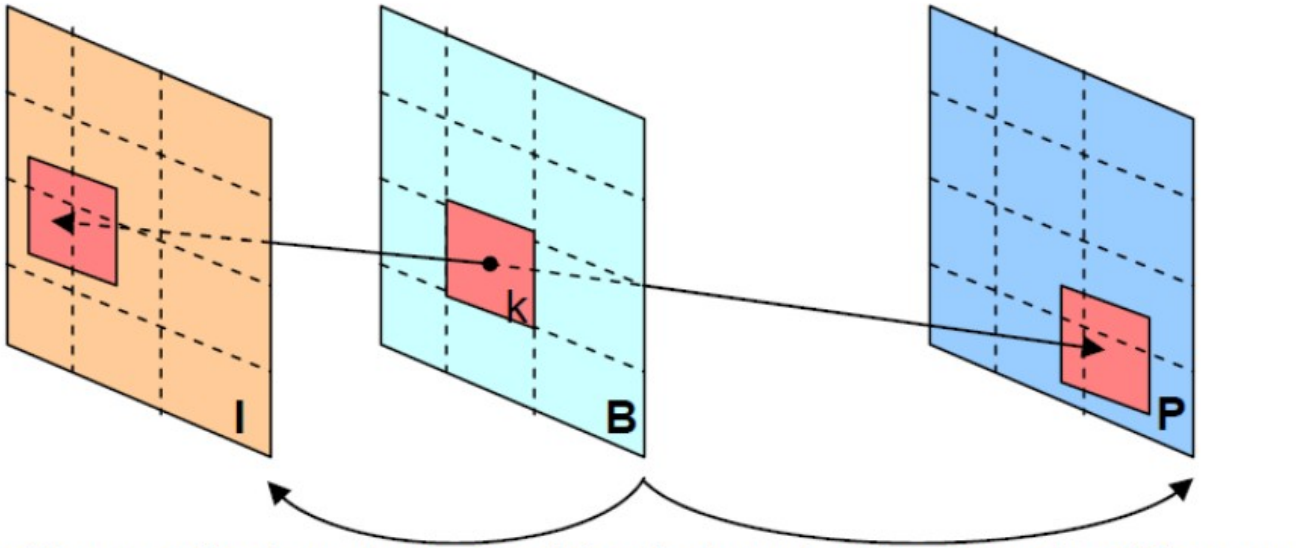
- Type I : Intra-codées
- Type P : Prédictives
- Type B : Bidirectionnelles



Codage du signal vidéo

Chaque image est découpé en *macro blocs* de taille 16 × 16 pixels.

Les macro blocs servent pour le codage par la transformée en cosinus discrète (DCT) des images I mais également pour les prédictions.



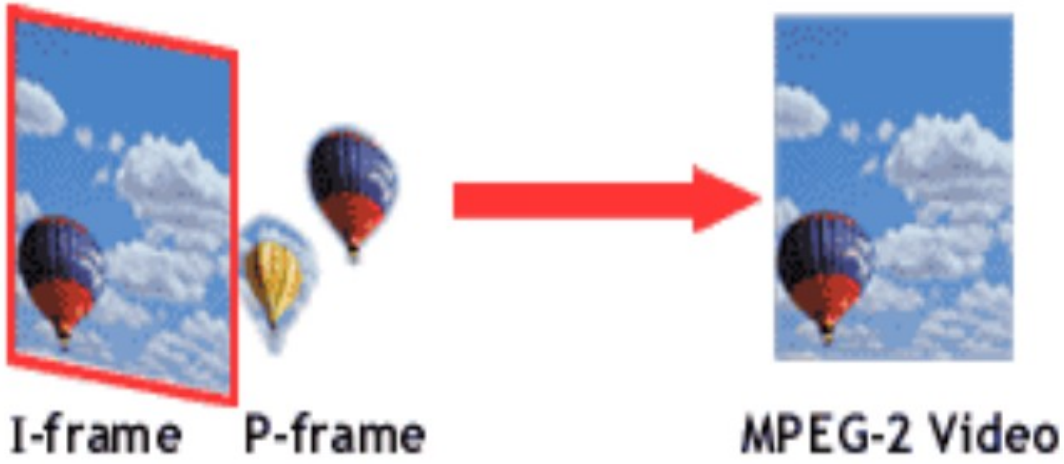
Positions estimées du macro bloc k dans les images de référence

Codage du signal vidéo

GOP (Group Of Pictures) : liste d'images entre deux images de type I

Longueur GOP est variable : valeur la plus courante entre 12 et 15

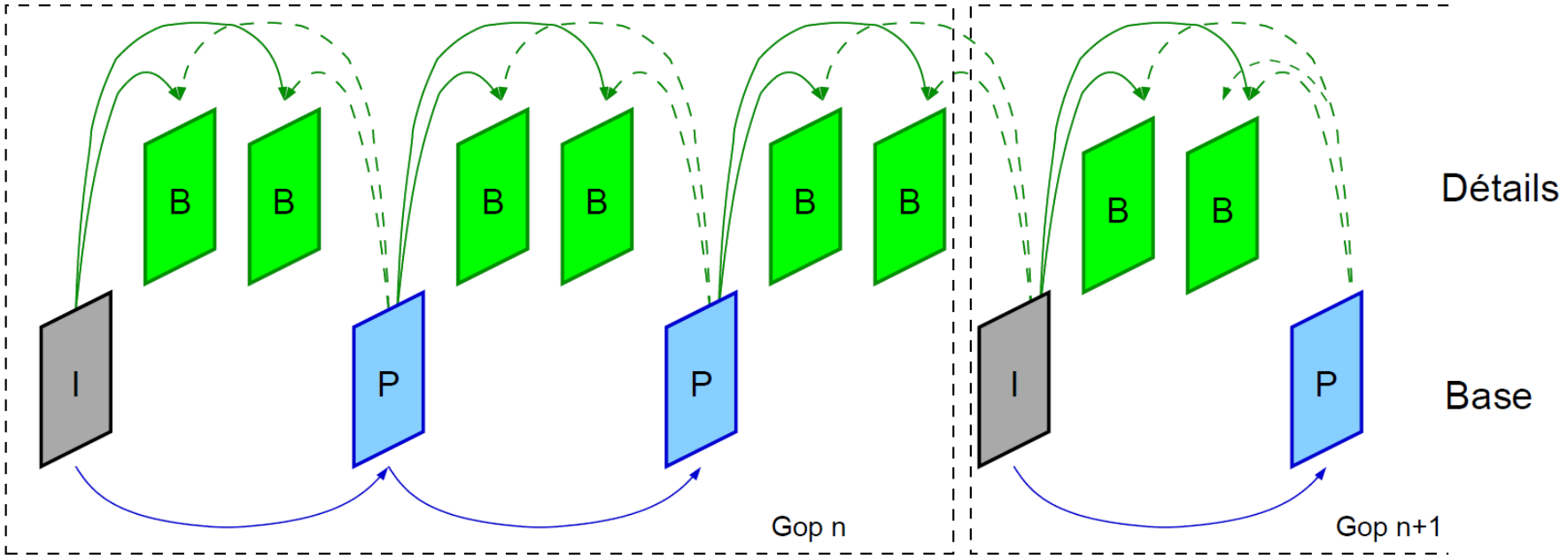
Décodage GOP par GOP : les images B sont prédites à partir des images antérieures ou postérieures et de type I ou P



Codage du signal vidéo

GOP (Group Of Pictures) : liste d'images entre deux images de type I

Décodage GOP par GOP : les images B sont prédites à partir des images antérieures ou postérieures et de type I ou P



Codage du signal vidéo

Étapes :

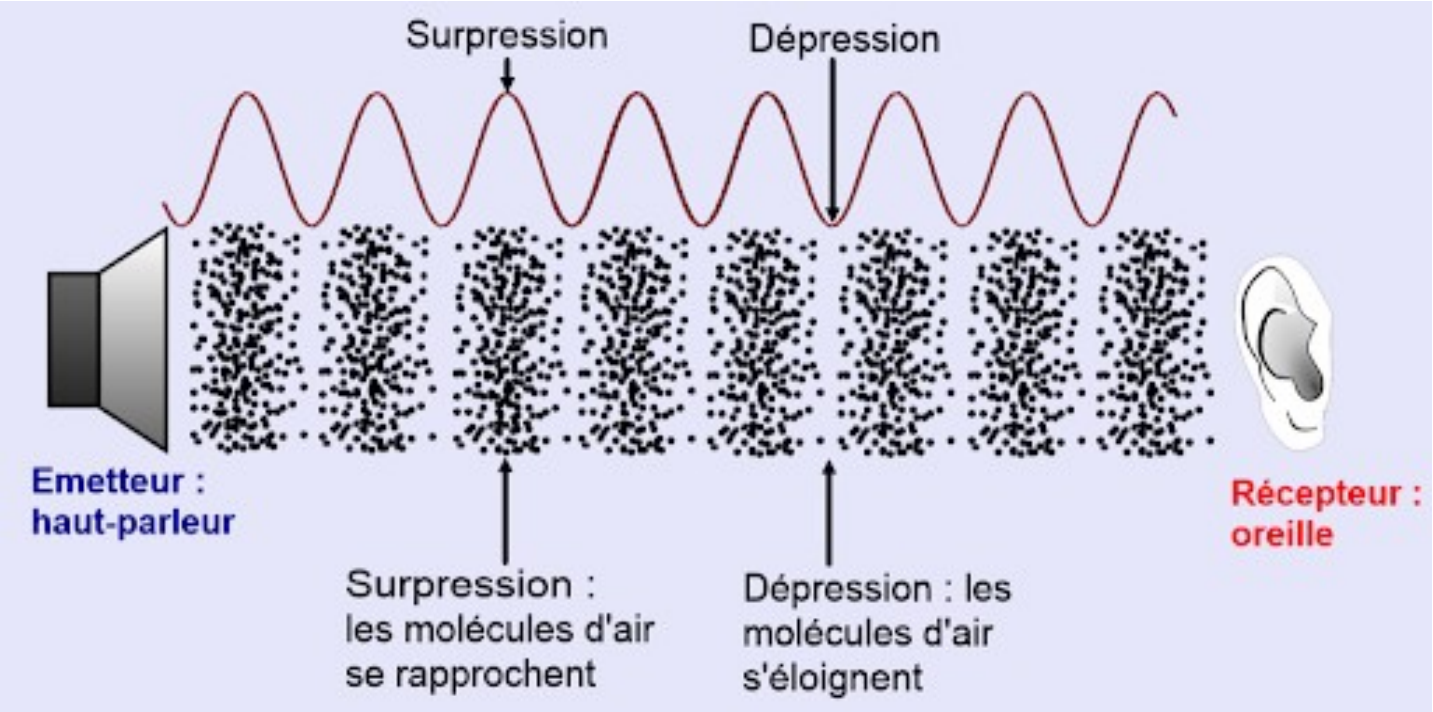
1. Les images I sont codées d'une manière proche des images JPEG.
2. Les macro blocs des images P et B sont estimés puis les erreurs sont compensées.
3. Le vecteur mouvement de chaque macro bloc d'une image P indique sa position estimée dans l'image I ou P précédente.
4. Pour les images B : deux vecteurs mouvements par macro bloc, chaque indiquant la position du macro bloc dans l'image I ou P précédente ou suivante.

L'extraction de données des images d'une séquence nécessite un supplément de mémoire dans le codeur et le décodeur, mais aussi génère du retard en fonction du nombre d'images bidirectionnelles.

Codage du son

Son : la forme d'onde (signal sonore)

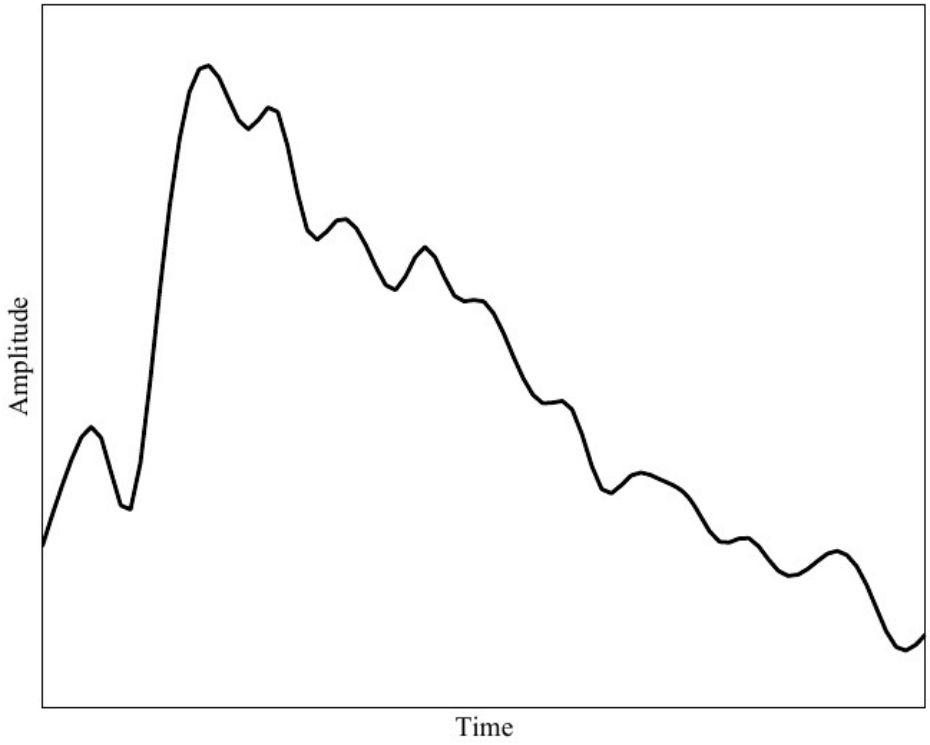
Evolution dans le temps de la pression de l'air au point de mesure (emplacement du microphone)



Codage du son

Son : la forme d'onde (signal sonore)

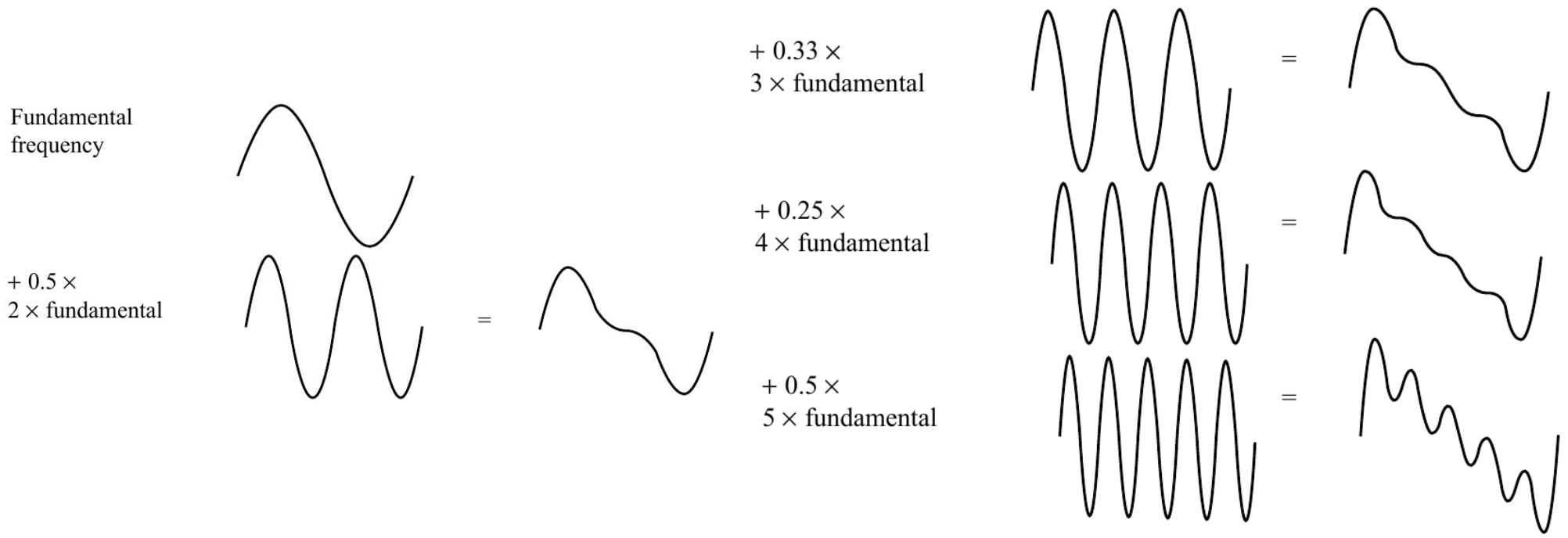
Evolution dans le temps de la pression de l'air au point de mesure (emplacement du microphone)



Codage du son

Le son peut vu comme une superposition de sinusoïdes

- composantes spectrales



Audition humaine : sons entre 20 Hz et 22 kHz (dans l'air)

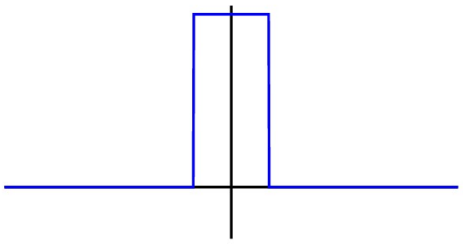
Codage du son

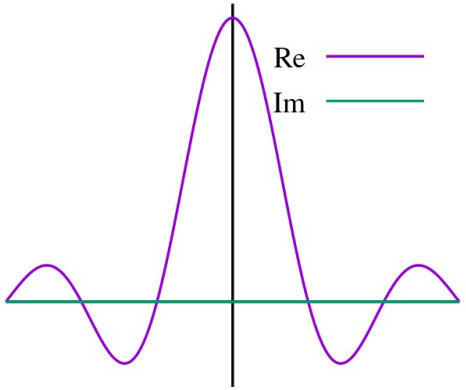
Le son peut vu comme une superposition de sinusoïdes

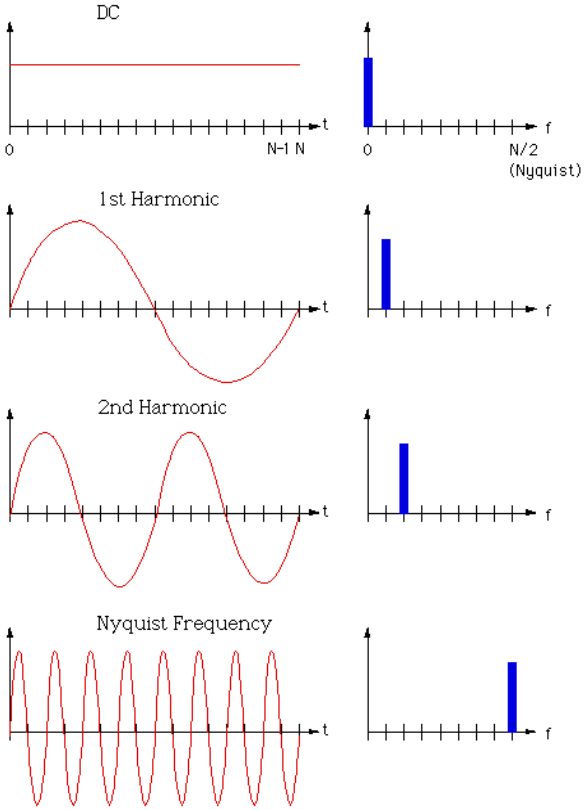
- composantes spectrales

Transformée de Fourier :

$$\mathcal{F}(f) : \nu \mapsto \hat{f}(\nu) = \int_{-\infty}^{+\infty} f(t)\, \mathrm{e}^{-\mathrm{i}2\pi\nu t}\, \mathrm{d}t$$





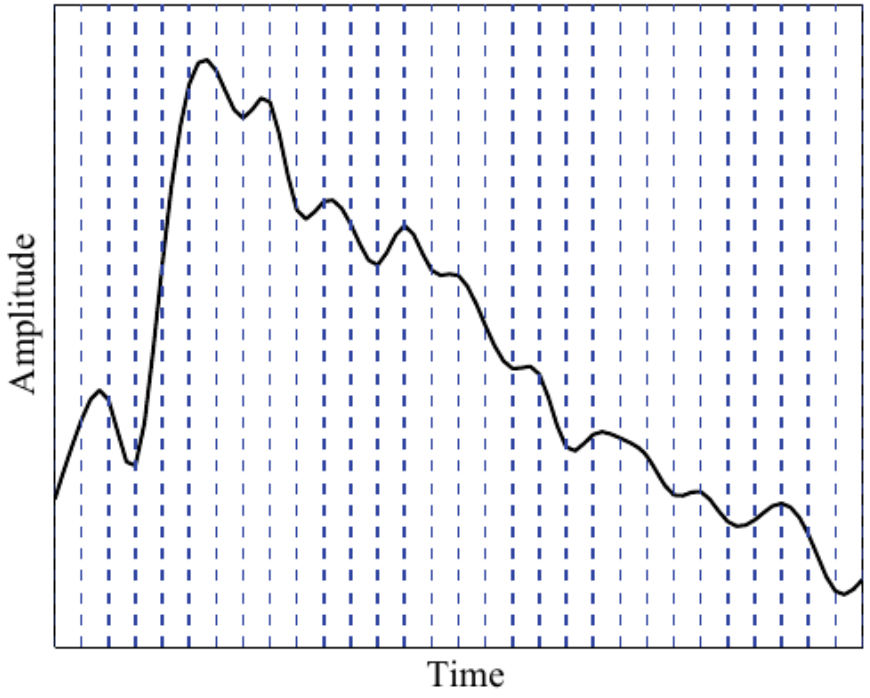


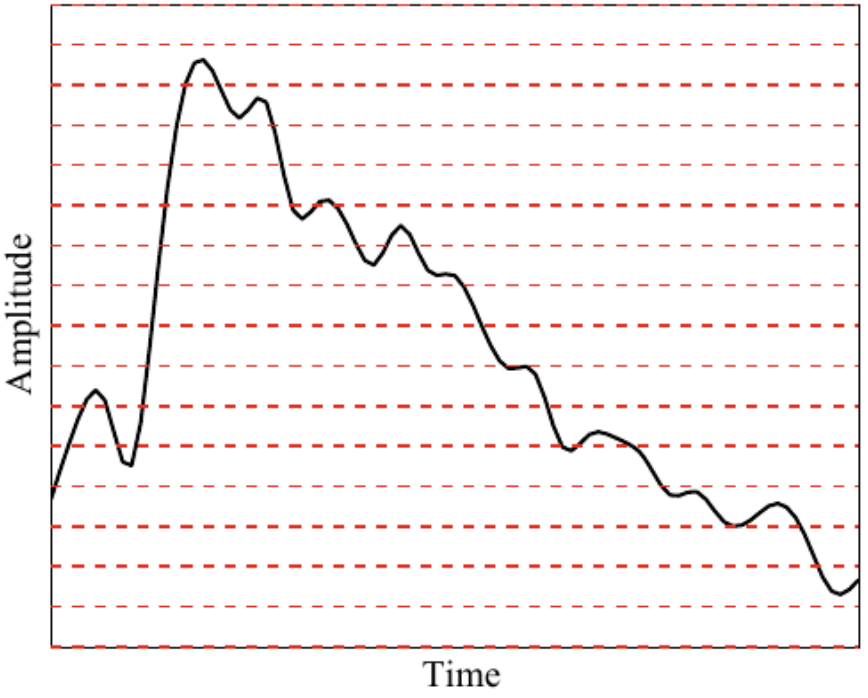
Audition humaine : sons entre 20 Hz et 22 kHz (dans l'air)

Codage du son

Échantillonnage et quantification

- échantillonnage : le temps
- quantification : l'amplitude





Codage du son

Échantillonnage : théorème de Nyquist

Un signal avec un spectre borné par la fréquence f_{max} : l'échantillonnage avec 2xf_{max} permet de récupérer le signal sans perte.

Fréquence max audible par l'homme : approx. 22kHz

Echantillonage à 44 Khz suffit pour représenter le son

Codage du son

Modulation par impulsions et codage (PCM)

PCM est une représentation numérique d'un signal électrique résultant d'un processus de numérisation.

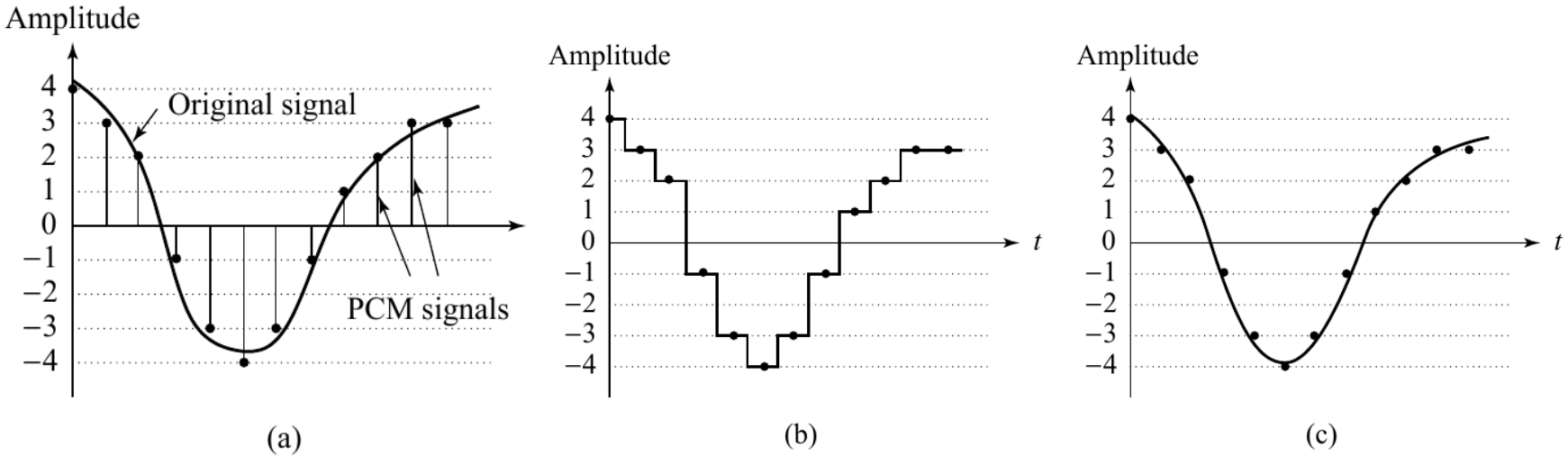
Le signal est d'abord échantillonné, puis chaque échantillon est quantifié indépendamment des autres échantillons, et chacune des valeurs quantifiées est convertie en un code numérique.

Modulation delta

La modulation différentielle (ou delta) d'impulsion-code (DPCM) code les valeurs PCM comme différences entre la valeur courante et la valeur précédente. Pour l'audio, ce type de codage réduit le nombre de bits exigé par rapport au PCM.

Codage du son

Modulation par impulsions et codage (PCM)



(a) signal d'origine

(b) signal décodé

(c) signal reconstruit avec filtrage passe bas

Stockage : formats de fichiers sans perte (e.g. FLAC) ou perte (e.g. MPG3)

Plan de la séance

- Structure d'un document vidéo
- Représentation numérique (codage/décodage)
- Segmentation temporelle
- Caractérisation du mouvement
- Caractérisation du contenu visuel
- Caractérisation sémantique
- Applications (reconnaissance des objets, détection de copies)

Segmentation temporelle (vidéo)

Décomposition structurée de la vidéo qui peut servir pour l'indexation et pour l'analyse sémantique.

Segments = unités stables pour étudier, comparer et caractériser les vidéos

Types de segmentation :

- Vidéo : plans, scènes, chapitres, émissions, ...
- Audio : parole/musique/bruit, locuteurs/chanteurs, ...

Segmentation temporelle (vidéo)

Plan : l'unité la plus courte après l'image et qui délimite deux prises de vue.

Méthodes :

- Codage : les images I correspondent parfois a un changement de plan
- Édition : les transitions (la coupure, le fondu et la dissolution)
- Analyse de contenu (dissimilarité forte entre cadres successifs)
- Domaine compressé (propriétés du flux MPEG : analyse de mouvement, des blocs et de débit, composante continue DCT)

Segmentation temporelle (vidéo)

Transition dans le flux : coupure franche

Segmentation temporelle (vidéo)

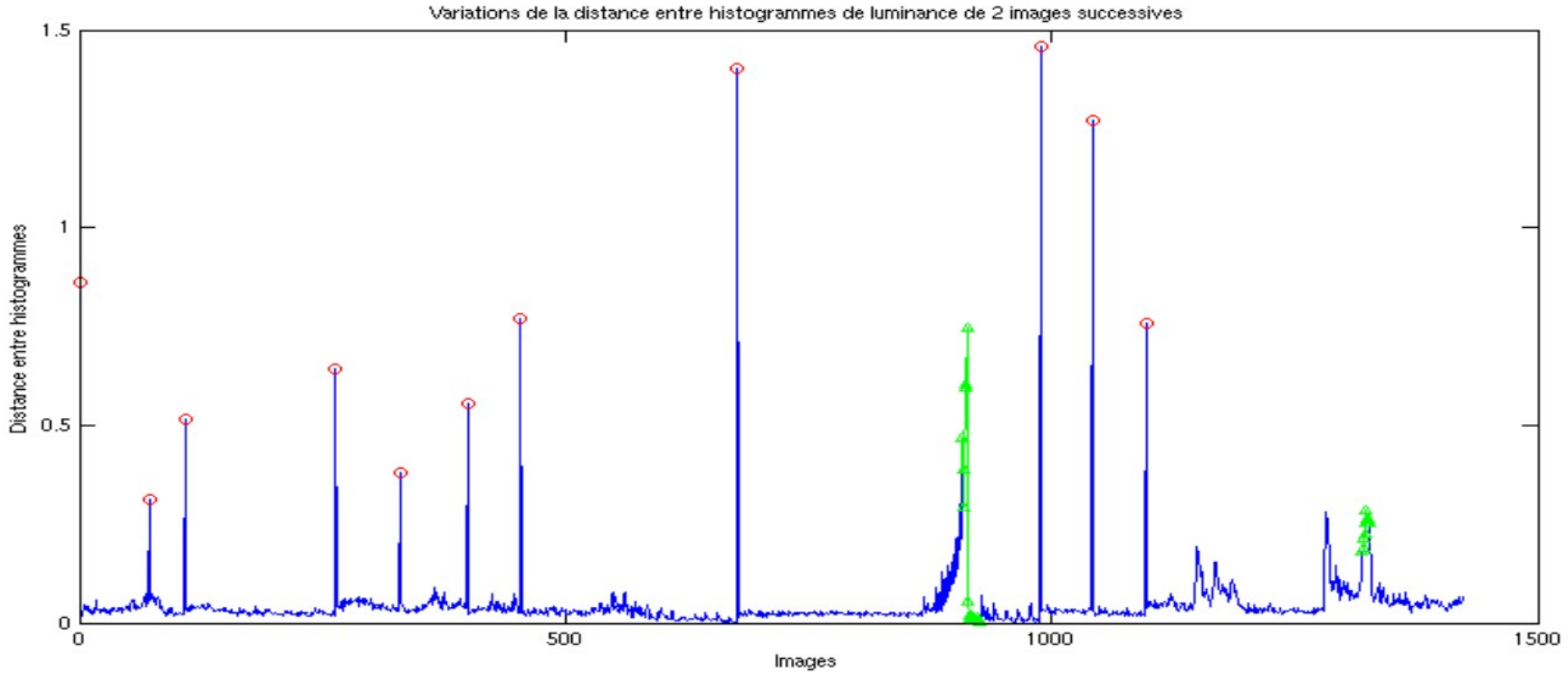
Transition dans le flux : fondus, volets

Segmentation temporelle (vidéo)

Détection de transition par analyse de contenu :

- Images successives dissimilaires au moment de la transition

Exemple : histogrammes en niveau de gris



Plan de la séance

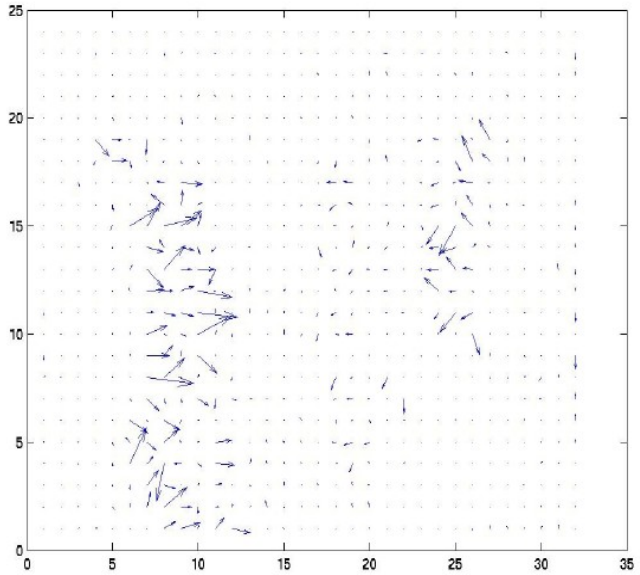
- Structure d'un document vidéo
- Représentation numérique (codage/décodage)
- Segmentation temporelle
- Caractérisation du mouvement
- Caractérisation du contenu visuel
- Caractérisation sémantique
- Applications (reconnaissance des objets, détection de copies)

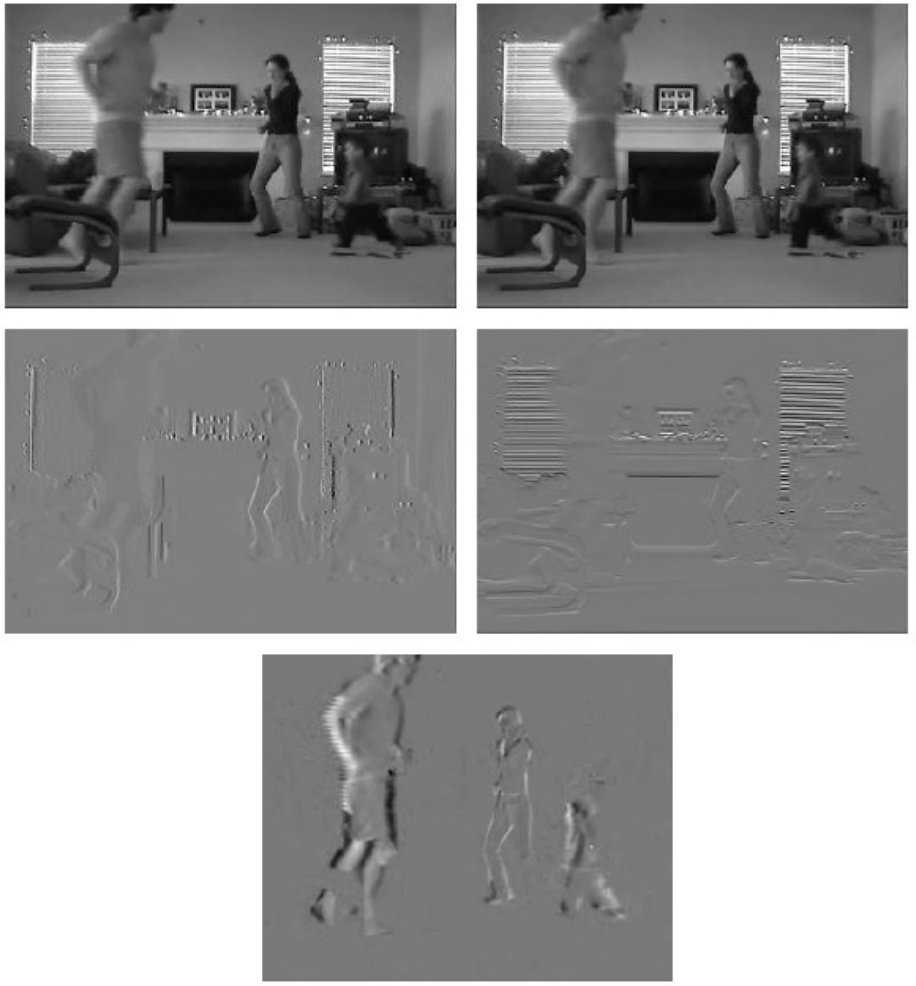
Caractérisation du mouvement

Le flot optique :

- au niveau des macro-blocs
- au niveau des pixels

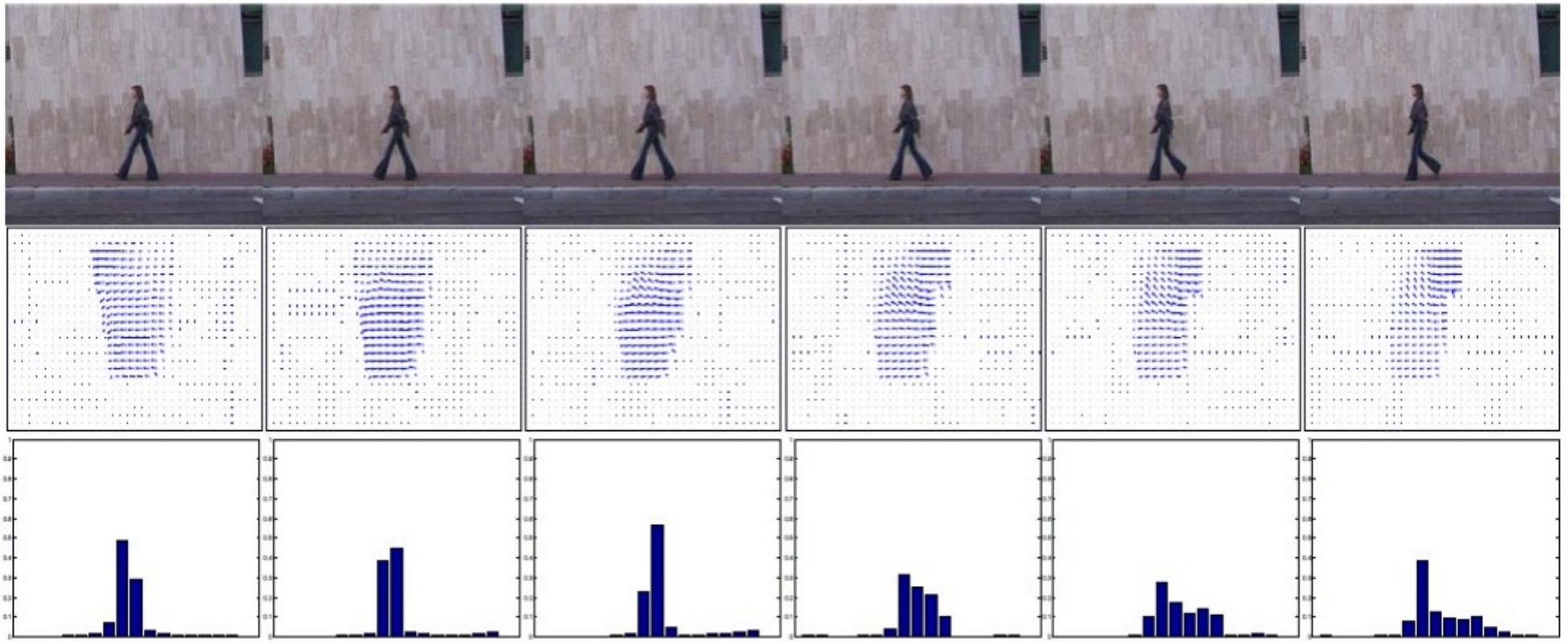
$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t)$$





Caractérisation du mouvement

Histogramme du flot optique



Plan de la séance

- Structure d'un document vidéo
- Représentation numérique (codage/décodage)
- Segmentation temporelle
- Caractérisation du mouvement
- Caractérisation du contenu visuel
- Caractérisation sémantique
- Applications (reconnaissance des objets, détection de copies)

Description du contenu visuel

- Rappel : document vidéo = séquence de plans, scènes
- Chaque plan : une image représentative
- Où : échantillonnage uniforme à plusieurs cadres par seconde (d'habitude entre 2 et 5)

Les méthodes de caractérisation des images sont généralement appliquées aux images représentatives de chaque plan (voir les séances concernant la description des images fixes).

Description du contenu visuel

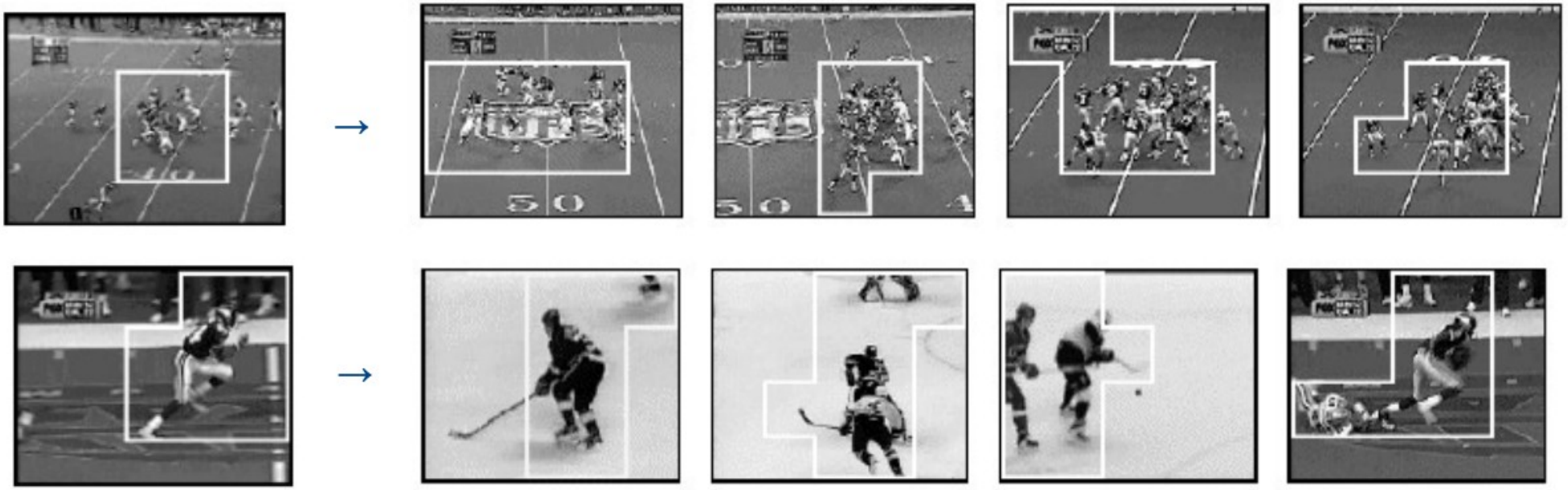
Descripteurs globaux : Couleur, texture, forme

requête

requête

Description du contenu visuel

Descripteurs locaux : points d'intérêt, activité de mouvement



Description du contenu visuel

Scènes, actions et suivi des objets

Points d'intérêts spatio-temporels

- Détecteur spatio-temporel [Laptev et Lindeberg 2003]
- Utilisation des dérivés gaussiennes avec une variance distincte sur l'axe temporel
- Détection des points qui ont des fortes variations de luminosité en x, y, et t
- Donne le même statut a l'axe temporelles est spatiales

Trajectoires des points d'intérêt :

- Clusters des trajectoires ↔ objets avec des mouvements similaires

Action recognition (classique)

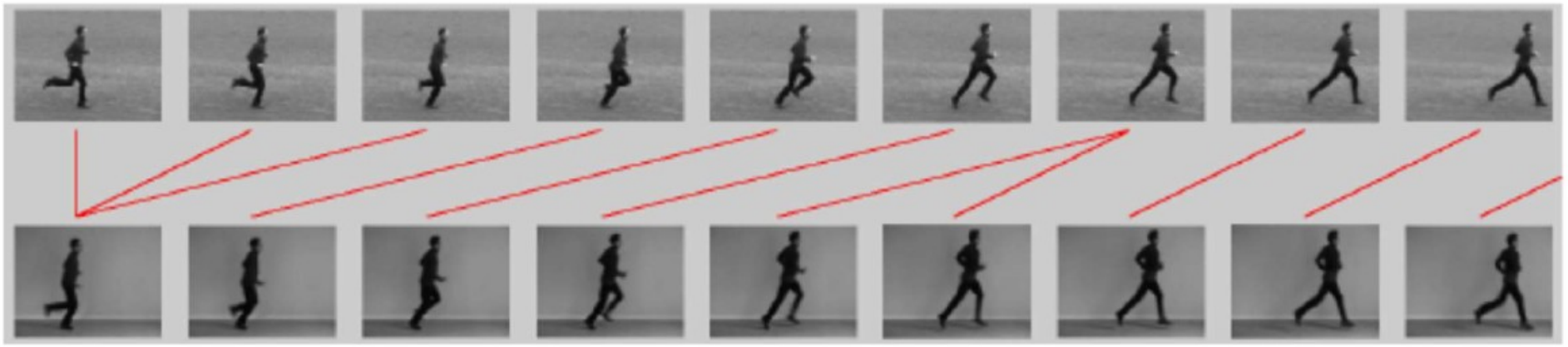
Dense trajectories: track grid points using optical flow

Histograms of: Gradients, optical flow, motion boundary

Action recognition

Dynamic time warp: Find the best warping path W^* (minimal cost path)

- Compute distances between all frames: $O(DL^2)$ complexity



Synchronisation par déformation temporelle
(Frame alignement)

Plan de la séance

- Structure d'un document vidéo
- Représentation numérique (codage/décodage)
- Segmentation temporelle
- Caractérisation du mouvement
- Caractérisation du contenu visuel
- Caractérisation sémantique
- Applications (reconnaissance des objets, détection de copies)

Caractérisation sémantique

Problème vaste, non résolu (beaucoup de travaux en cours)

Recherche type : « L'incroyable coup-de-tête de Zidane en finale de la coupe du monde 2006. »

Première approche : annotation manuelle

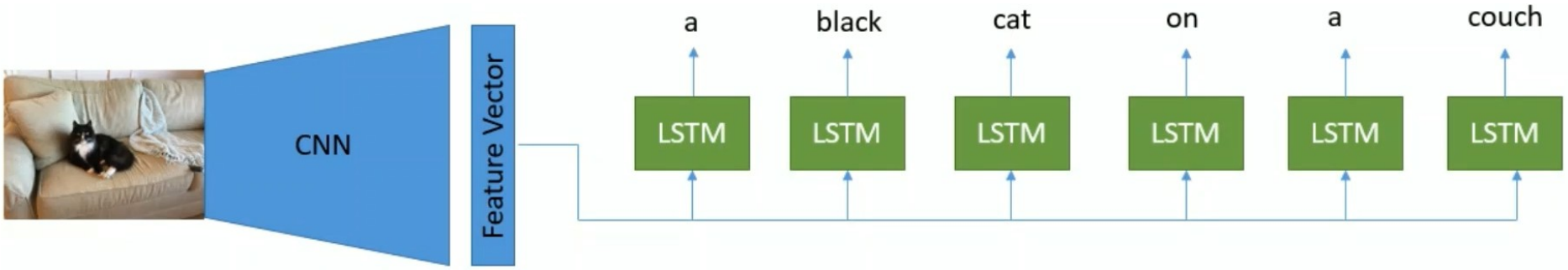
- Coûteux
- Ne passe pas à l'échelle

Approches récents :

- Extraction automatique des caractéristique sémantiques
- Algorithmes de classification automatique
- Apprentissage des classes (sujets sémantiques)

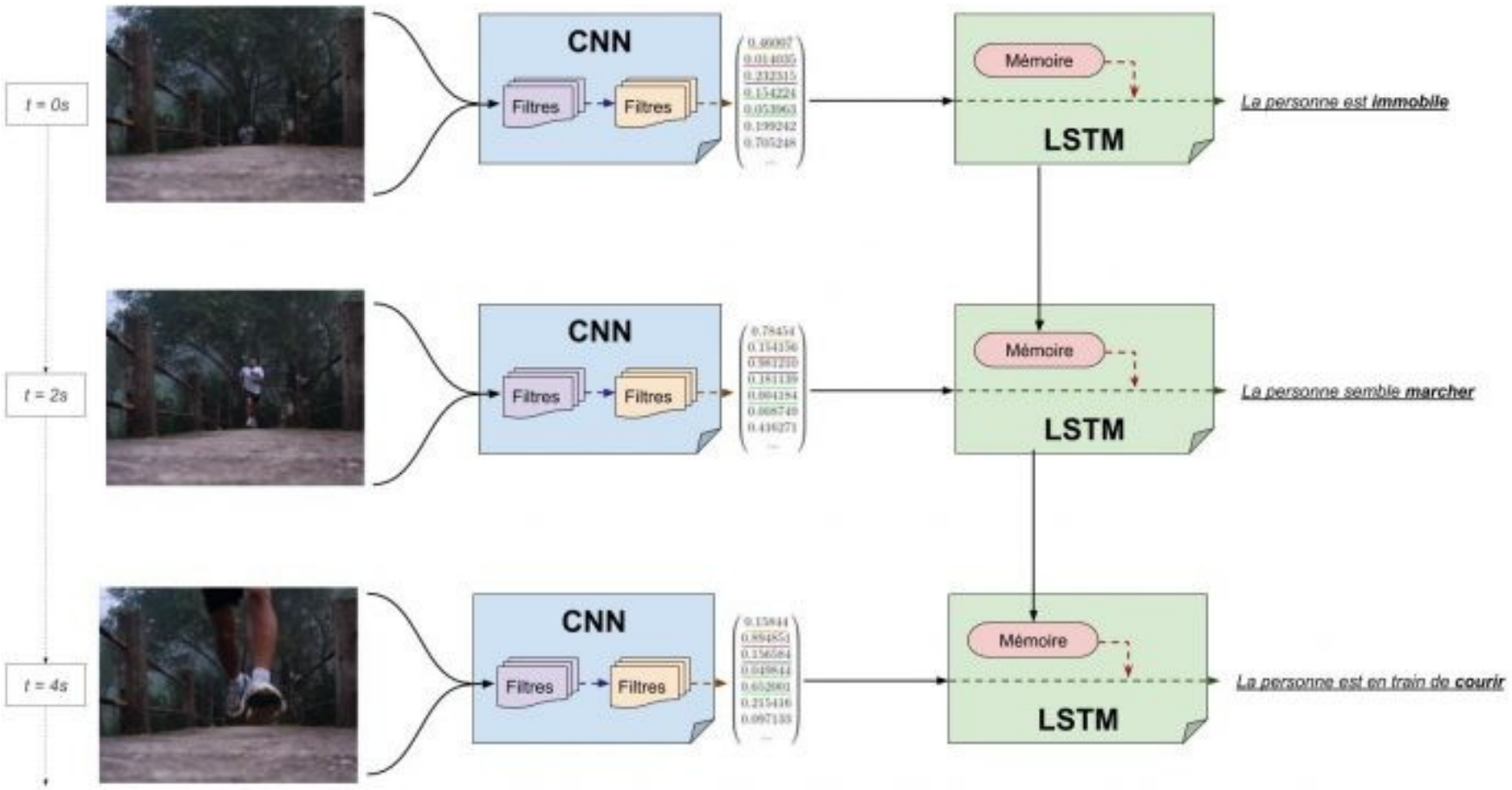
LSTM-CNN

Exemple description mixte : image captioning LSTM-CNN



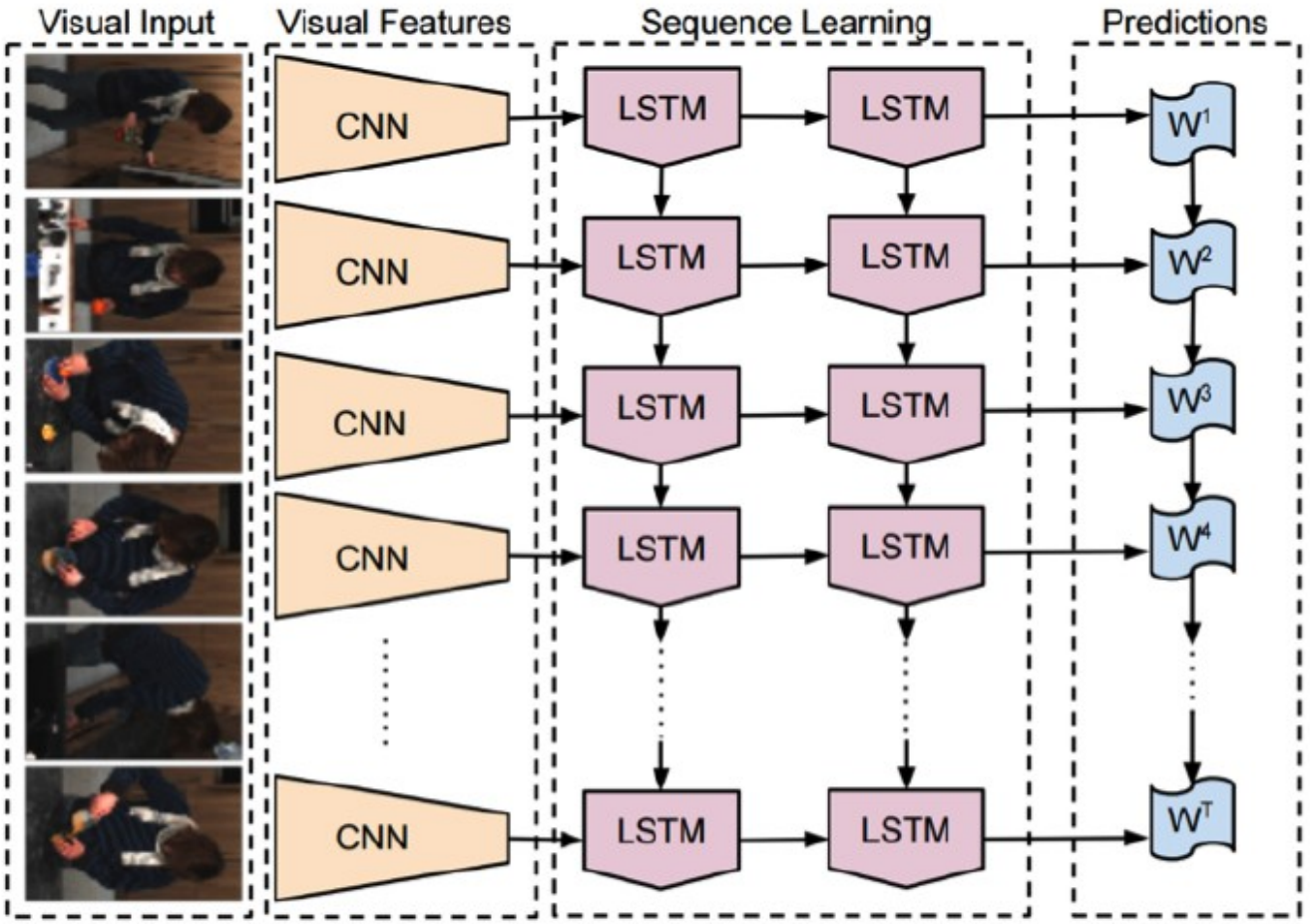
CNN : encode image to feature vectors
LSTM : decode feature vectors to natural language

LSTM-CNN



Classification de vidéos/actions

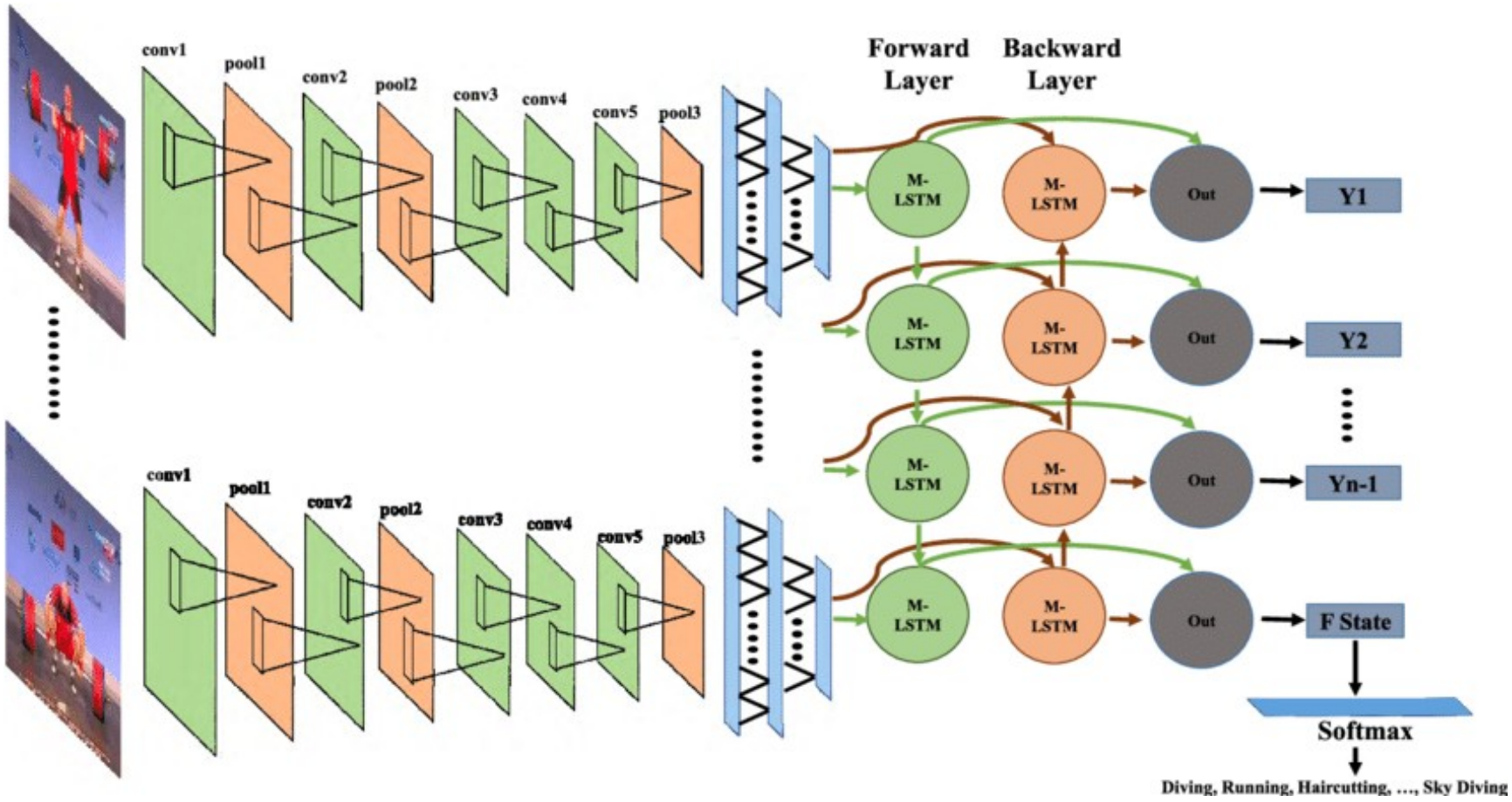
LSTM-CNN



Thung et al, 2015

LSTM à plusieurs couches

LSTM-CNN

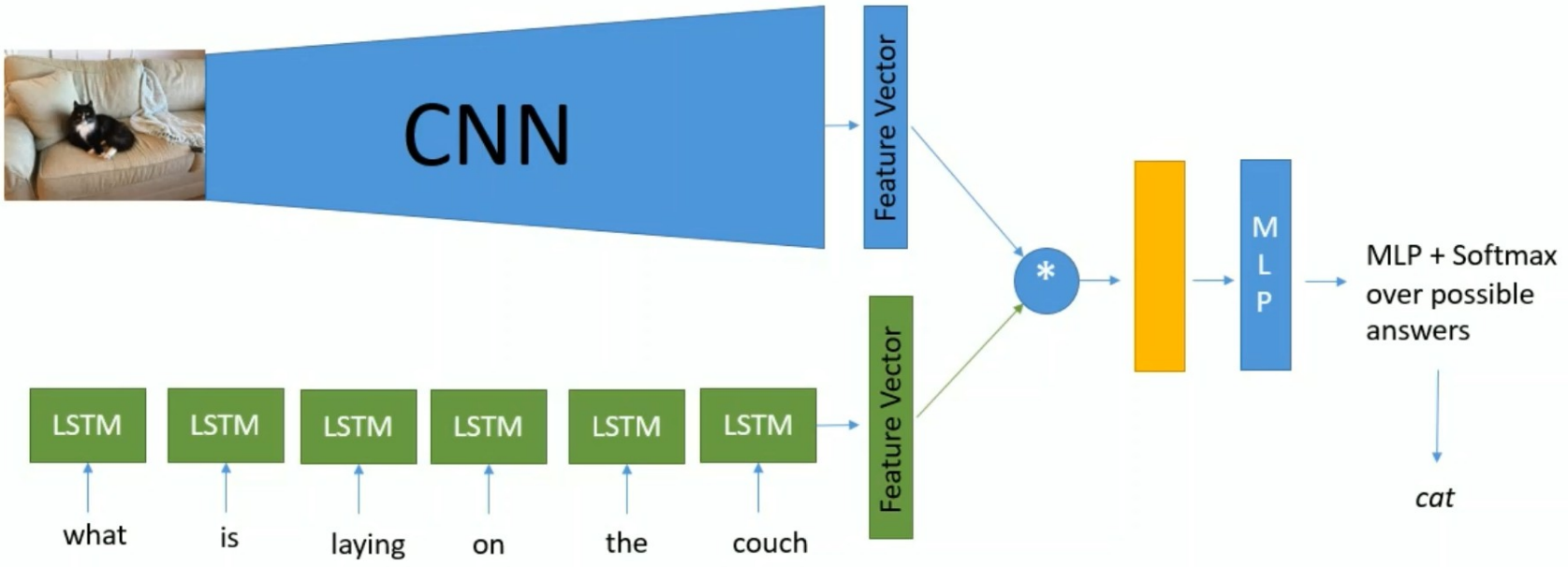


Ullah et al., 2018

LSTM bidirectionnel

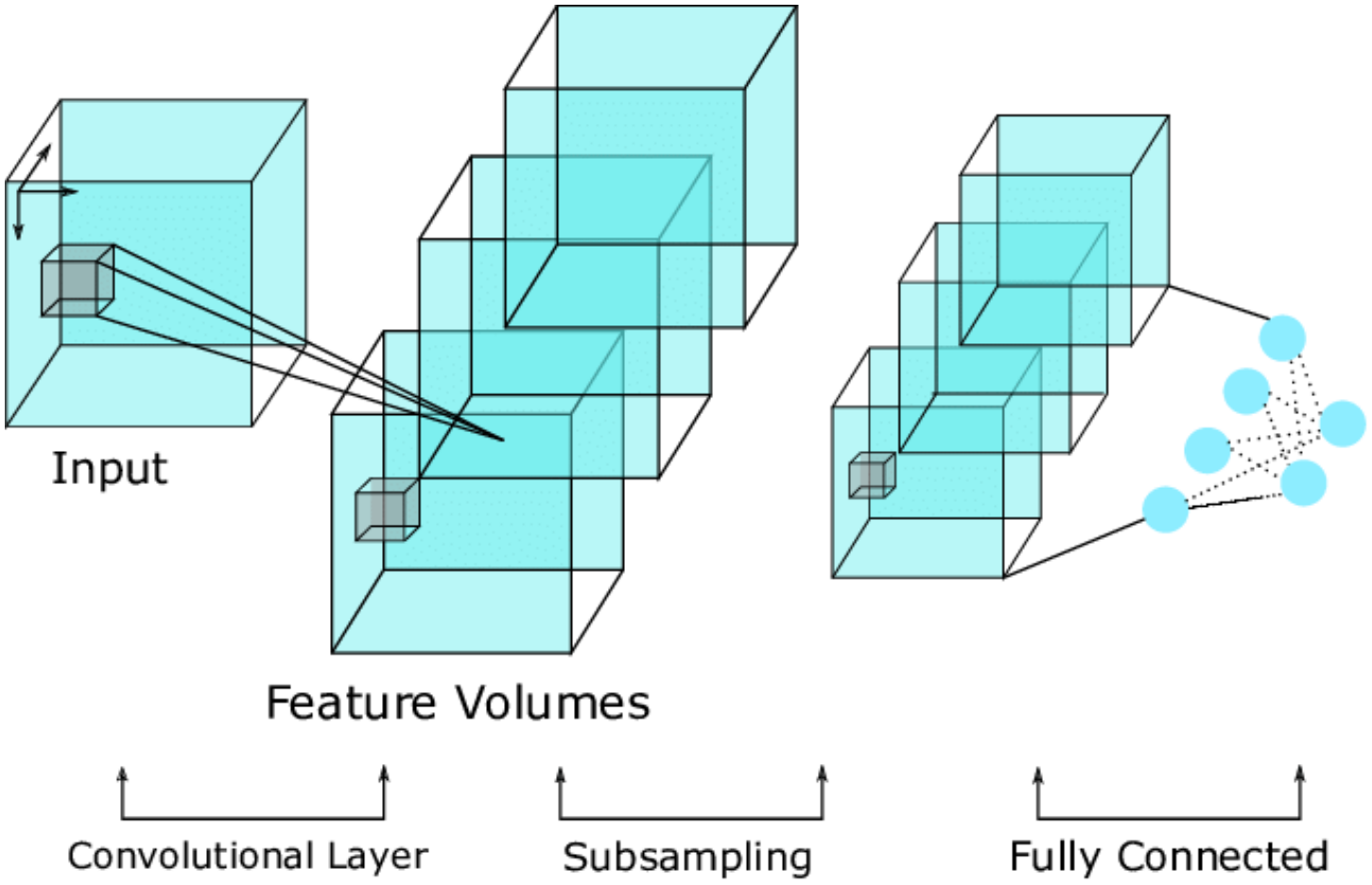
LSTM-CNN

Classification vidéo – approche par fusion CNN/LSTM

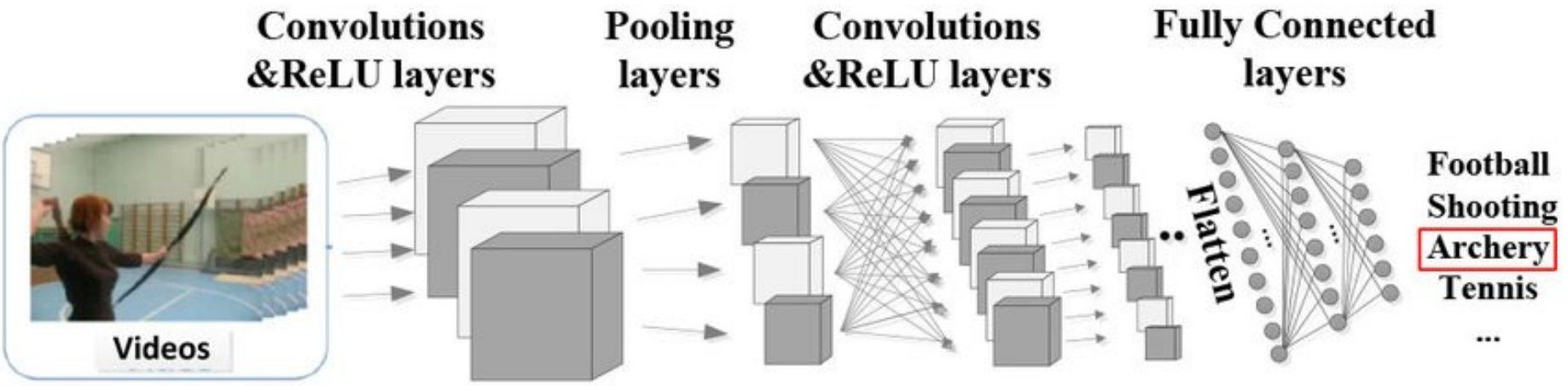


Architectures similaire au VQA

Convolution 3D pour la classification

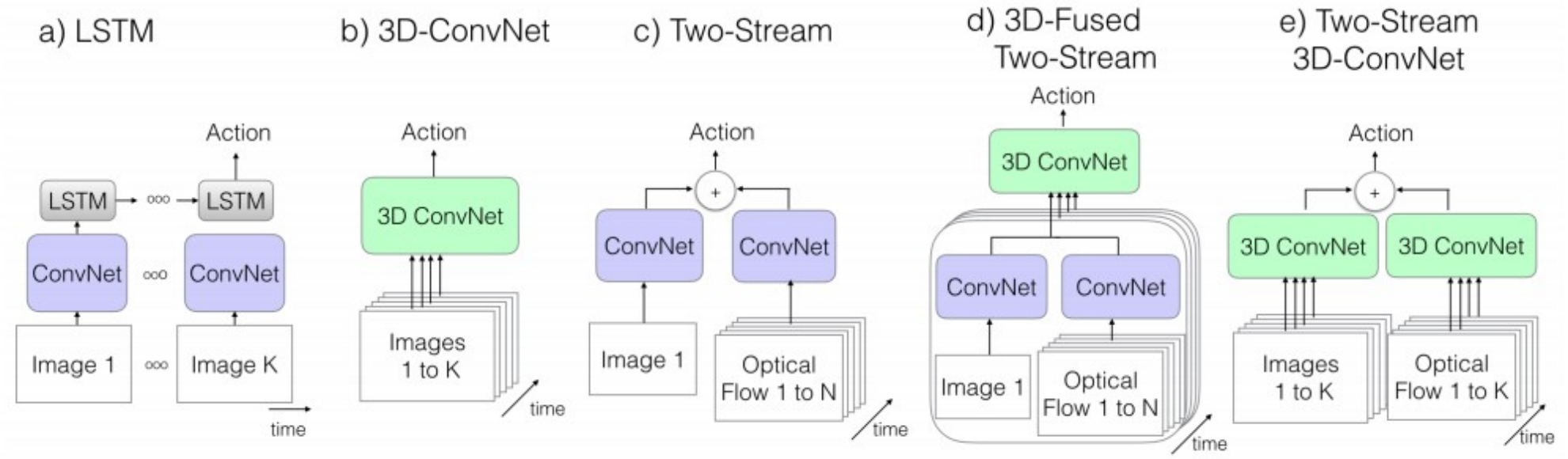


Convolution 3D pour la classification



Tran et al. 2015

(Deep-)Action recognition



Carreira et Zisserman, Quo Vadis, Action Recognition?

Description du contenu visuel

Scènes, actions et suivi des objets

- Extraction de caractéristiques : CNN, transfer learning
- Détection et localisation d'objets : YOLO, etc.
- Annotation automatique (image captioning)
- Suivi des objets (DeepSORT, Recurrent Yolo, etc.)

Référence littérature récente : Luo et al., *Multiple object tracking: A literature review*, Artificial Intelligence Journal, Volume 293, 2021

Further Topics

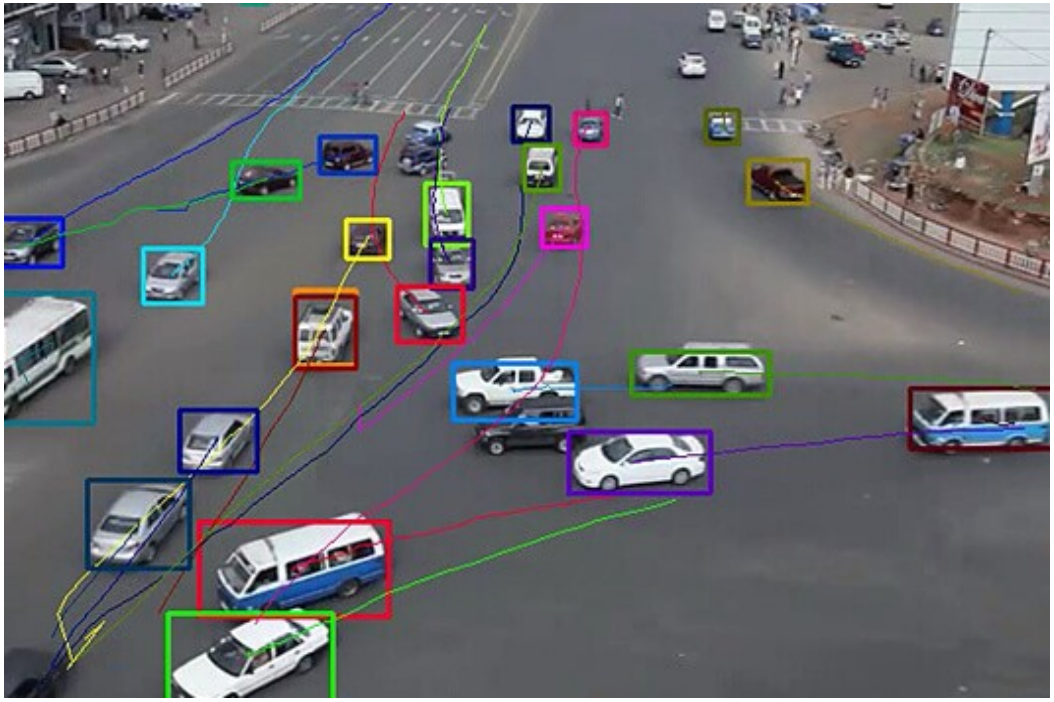
- Deep Video Action Recognition [16]
- Human Action Recognition [17]
- Predictive Understanding: Early Action Recognition and Future Action Prediction [18]
- Object Segmentation and Tracking [19]
- Video Summarization [20]

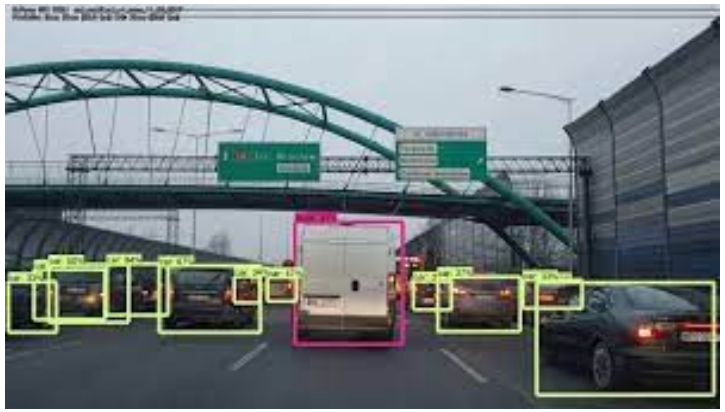
Plan de la séance

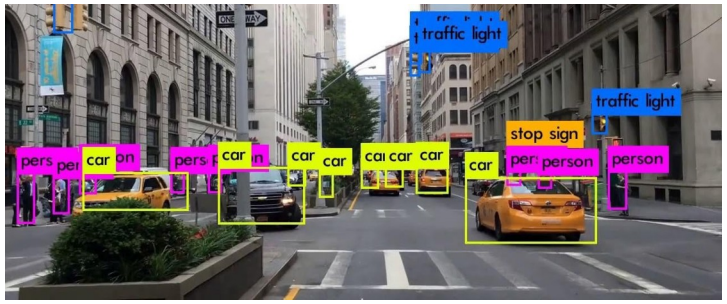
- Structure d'un document vidéo
- Représentation numérique (codage/décodage)
- Segmentation temporelle
- Caractérisation du mouvement
- Caractérisation du contenu visuel
- Caractérisation sémantique
- Applications

Applications (vidéo)

- Reconnaissance des objets
- Suivi des objets
- Reconnaissance des actions







Bibliographie

1. Fundamentals of Multimedia, Ze-Nian Li, Mark Drew, Jiangchuan Liu, Springer 2021.

2. Multimedia, Big Data, Computing for IoT Applications : Concepts, Paradigms and Solutions, Tenwar et al., Springer 2000

3. Image and Video Compression for Multimedia Engineering, Shi et al., CRC Press 2020

4. Intelligent video surveillance systems, Kolekar et al, CRC Press 2018

5. Multimodal analysis of user-generated multimedia content, R. Shah, R. Zimmermann, Springer 2017

6. Digital Image Processing, R. Gozalez, E. Woods, Pearson 2018, 4th edition

7. Structuration automatique de flux télévisuels, Jean-Philippe Poli, thèse doctorat, 2007

8. Computing 2D and 3D Optical Flow, J.L. Barron, N.A. Thacker, Tina Memo No. 2004-012, 2005.

Bibliographie

8. Cho et al., 2014. On the properties of neural machine translation: Encoder-decoder approaches

9. Chung et al., 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling

10. Hochreiter & Schmidhuber 1997. Long short-term memory

11. Fortun et al, Optical flow modeling and computation: A survey, Computer Vision and Image Understanding, 2015

12. Andrei Stoian, Scalable action detection in video collections, PhD thesis, 2016

13. Tran et al., Learning Spatiotemporal Features with 3D Convolutional Networks, ICCV 2015

14. Ullah et al., Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features, IEEE Access 2018

Bibliographie

16. Yi Zhu *et al.*, A Comprehensive Study of Deep Video Action Recognition, arXiv 2012.06567v1, 2020

17. Sun *et al.*, Human Action Recognition from Various Data Modalities: A Review, arXiv 2012.11866v4, 2021

18. Zhao He, Richard Wildes, Review of Video Predictive Understanding: Early Action Recognition and Future Action Prediction, 2107.05140v2, 2021

19. Yao *et al.*, Video Object Segmentation and Tracking: A Survey, arXiv 1904.09172v3, 2019

20. Apostolidis *et al.*, Video Summarization Using Deep Neural Networks: A Survey, arXiv 2101.06072v2, 2021