Bases de données multimédia
VIII – indexation vidéo & audio

ENSIMAG

2014-2015

Matthijs Douze & Karteek Alahari

Plan

- Spécificité du problème
- Filtrage temporel
- Détecteur spatio-temporels
- Aggrégation temporelle
- Zoom sur deux méthodes d'indexation vidéo
- Indexation audio
- Quelques perspectives

Spécificité du problème (début)

- Vidéo = suite d'images

- L'axe temporel *t* est singulier
  - processus physique différent des axes *x* et *y* : mouvement de la caméra et des objets qui la composent
  - la nature des invariants est différente sur cet axe (voir exemples)

- Segmentation temporelle en "plans" (au sens cinématographique)
  - c'est un problème en soi → assez bien maîtrisé aujourd'hui
  - indexation au niveau des plans plutôt que des images composant la vidéo
  - **Avantage?**

Spécificité du problème (fin)

- Indexation de vidéos = plusieurs problèmes possibles

1) Requête vidéo (souvent un plan)
   - problème qui est le plus souvent considéré
2) Requête image
   - asymétrie de la requête
   - plus difficile (car moins d'information dans la requête, et pas de description spatio-temporelle possible)
   - toutes les techniques ne s'appliquent pas

- Dans l'état de l'art, deux classes d'approches
  - indexation des images composant la vidéo + intégration cohérence temporelle
  - indexation spatio-temporelle (pas utilisable avec une requête image)

Plan

- Spécificité du problème
- Filtrage temporel
- Détecteur spatio-temporels
- Aggrégation temporelle
- Zoom sur deux méthodes d'indexation vidéo
- Indexation audio
- Quelques perspectives

Indexation vidéo = extension de l'indexation d'images

- Indexation 2D « usuelle »
- optionnellement, sous-échantillonnage (régulier ou par sélection d'images clés) pour réduire le nombre d'images
- génération de descripteurs pour chacune des images
- dans Video Google, [Sivic et Zisserman 2003]
  - appariemment des descripteurs de l'image *t* avec ceux des frames voisines
  - calcul d'un modèle dynamique simple (non détaillé, plusieurs variantes possibles) sur la position des descripteurs
  - suppression des descripteurs qui ne se reproduisent sur un voisinage temporel suffisant *ou* d'une manière non cohérente avec le modèle

Aggrégation temporelle (1)

- Contexte : appariement d'images de la vidéo
- Chaque frame de la requête est soumise à une base de frames qui représentent l'ensemble des vidéos à indexer

→ produit un ensemble de couple de frames et le score associé *(t_q, b, t_b, s)*

t_q	position temporelle dans la vidéo requête
b	numéro de vidéo dans la base
t_b	position temporelle dans la vidéo de la base
s	score obtenu par l'algorithme d'appariemment d'image

Aggrégation temporelle (2)

- Problème : estimer une fonction entre t_q et t_b

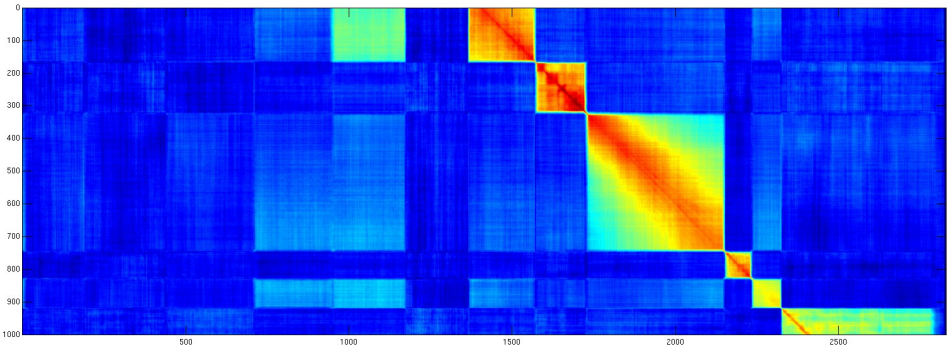
- Modèles possibles :
  - simple (décalage temporel uniquement): $t_q = t_b + \delta t$
  - incluant des variations globale de vitesse (slow-motion): $t_q = a^* t_b + \delta t$
  - complexe avec une table de décalage: $t_q = t_b + \text{shift}[t_q]$

- Méthodes d'estimation possibles
  - RANSAC
  - Transformée de Hough
  - Dynamic Time Warping (DTW)

Différents modèles

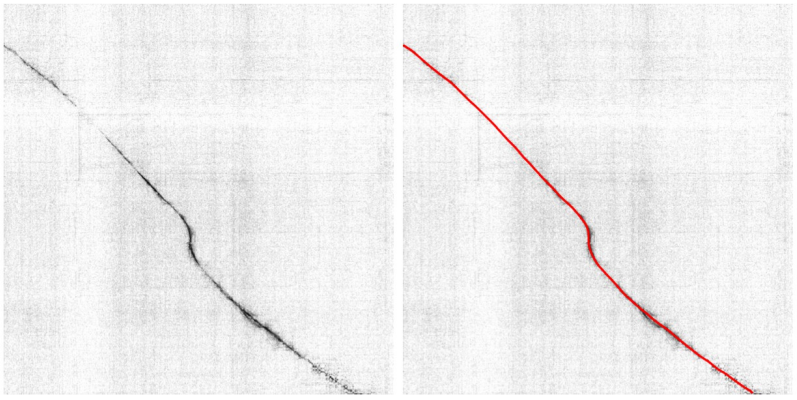
- Décalage temporel

"Event retrieval in large video collections with circulant temp encoding", Jérôme Revaud, Matthijs Douze, Cordelia Schi Hervé Jégou

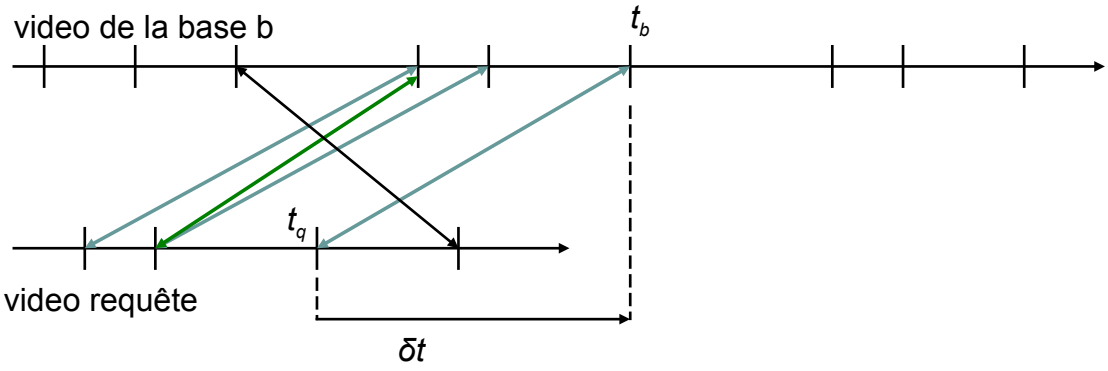


- Plus général

"VideoSnapping: Interactive Synchronization of Multiple Videos", Oliver Wang, Christopher Schroers, Henning Zimmer, Markus Gross, Alexander Sorkine-Hornung, siggraph 2014



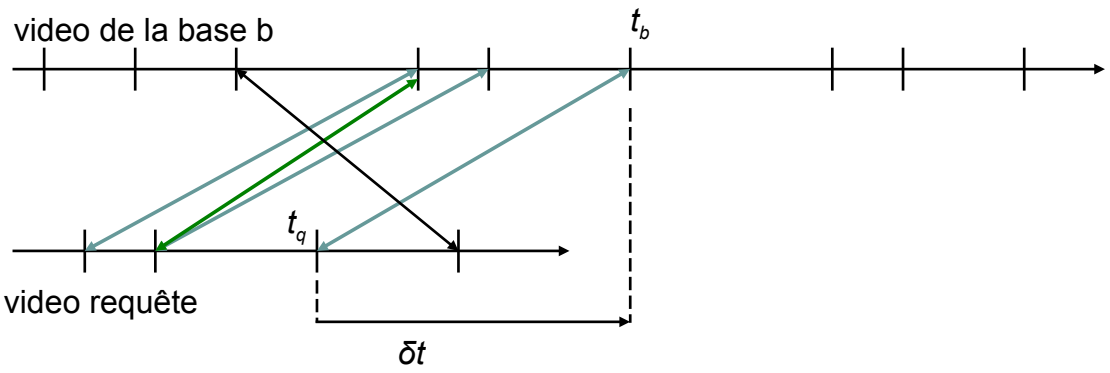
Aggrégation temporelle: Hough



- Exemple d'estimation
  - Utilisation du modèle simple de décalage temporel
  - Estimation en utilisant la transformée de Hough (temporelle)
  - Bins de la transformée ≈ taille du même ordre de grandeur que le pas d'échantillonage utilisé

→ Tolérance limitée aux changements modérés de vitesse (± 20 %)

Aggrégation temporelle: Hough (2)



- En sortie de l'estimation: liste d'hypothèses. Chaque hypothèse =
  - (b, δt) : video de la base et l'estimé du décalage δt
  - Le groupe de frames appariées (t_q, t_b).
  - Ce groupe doit être compact (<1 minute), ou bien il est découpé → "segment based matching"
  - Un score aggrégé calculé à partir des scores des frames appariées

Dynamic Time Warping (DTW)

- Algorithme de calcul de distance entre deux séquences
  - similaire à la distance d'édition
  - utilisée pour comparer les séquences sonores
- Algorithme au tableau
- Nécessité de normaliser les scores (typiquement, par la longueur de la séquence)
- Question : quelle est la complexité de cet algorithme ?
- Optimisation : techniques de "pruning"

Vérification spatio-temporelle (1)

- Objectif : estimation d'une transformée 2D entre la vidéo requête et la vidéo de la base de données

→ extension 2D → 3D des techniques utilisées en image

- Modèles 2D possibles :
  - Pour des caméras différentes: géométrie épipolaire
  - Pour le camcording : modèle homographique
  - Modèle affine 2D (affine complet, similitude, translation, etc)
- Problèmes
  - modèles peuvent changer dans le temps (d'une frame à l'autre)

→ Question : pourquoi ?

  - échantillonnage différent côté base et requête

Inria

Grenoble INP ensimag

Vérification spatio-temporelle (2)

- "Bon" choix: modèle affine 2D complet, avec paramètres *fixés*
- Le modèle est estimé sur un groupe de frames appariées

→ c'est-à-dire, prend en entrée une hypothèse générée en sortie de l'aggrégation

- Avantage d'un modèle de paramètres supposé fixe
  - tous points de tous les couples d'images appariées sont utilisés

  → beaucoup plus de points pour l'estimation que dans le cas image
- Modèle imparfait, mais pour lequel on peut obtenir une bonne estimation

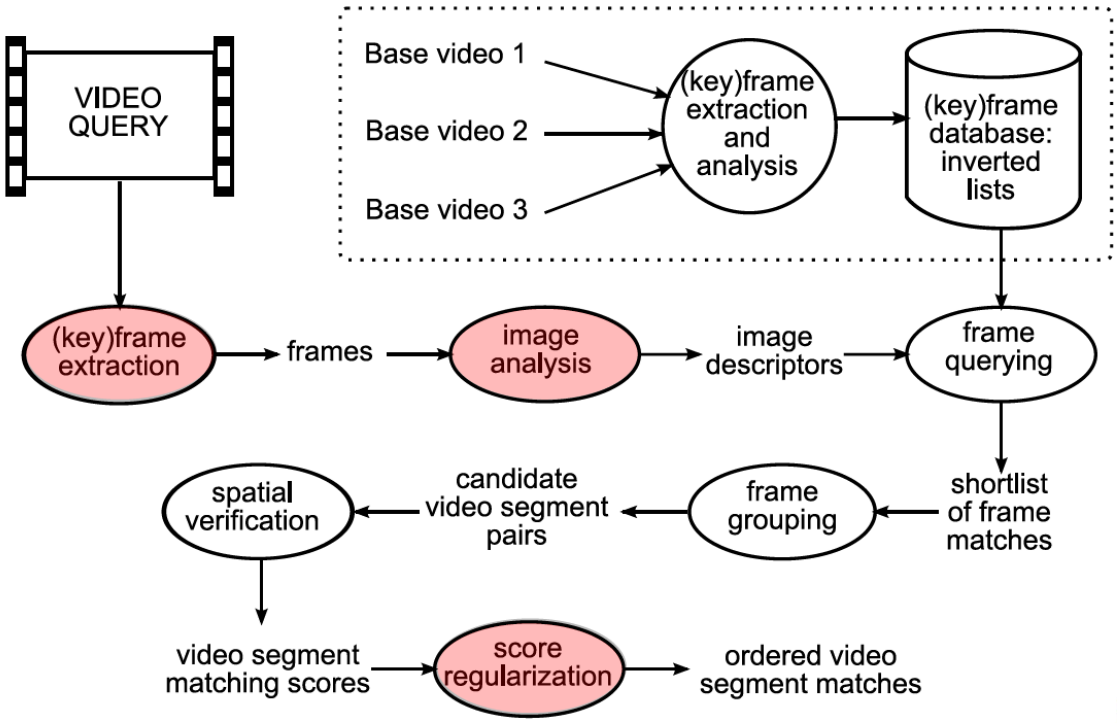
Inria

Grenoble INP ensimag

Vérification spatio-temporelle (3)

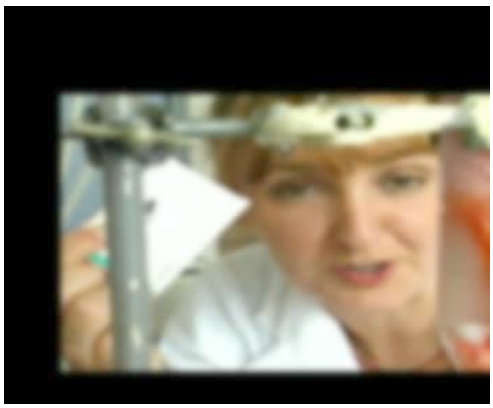
- Sortie pour chaque hypothèse (si la vérifiation spatio-temporelle réussit) :
  - un ensemble de paires de frames appariées et le score associé
  - la matrice de la transformée affine 2D

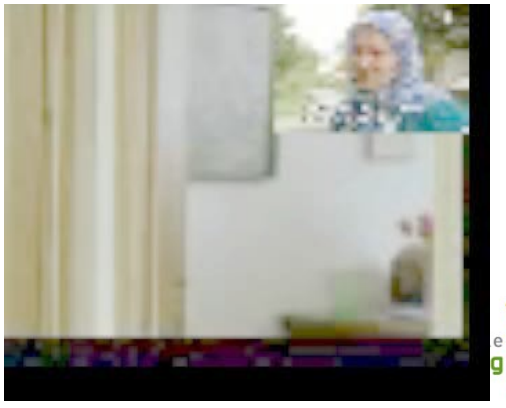
Méthode utilisée pour la compétition Trecvid : vision d'ensemble



Exemple de requêtes

Exemples de requêtes



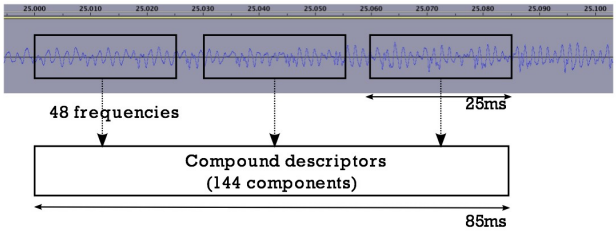


Plan

- Spécificité du problème
- Filtrage temporel
- Détecteur spatio-temporels
- Aggrégation temporelle
- Zoom sur deux méthodes d'indexation vidéo
- Indexation audio
- Quelques perspectives

Indexation audio

- Descripteurs "locaux" audio
  - dense (pas de détecteur)
  - fenêtres courtes (25 ms) qui se recouvrent (toutes les 10 ms)
  - description par 48 coefficients spectraux dans la bande de fréquence 300 – 3000 Hz (maximum de sensibilité de l'oreille humaine)
  - aggrégation de 3 fenêtres temporelles → descripteur 144 D



- Un autre exemple: MFCC
- Mise en correspondance = distance L2
- Indexation par quantificateur produit
- Vérification temporelle par Hough 1D

Indexation de gros volumes

- Hypothèses :
  - attaques faibles (pas/peu de déformation géométrique)
  - volume élevé (1000s d'heures)
- Méthode
  - sélection d'images clés
  - calcul d'une signature binaire de petite taille (DCT, 64 bits ou VLAD)
  - utilisation d'une structure de hachage pour la requête : accès en O(1)
- Xavier Naturel, Patrick Gros. *A Fast Shot Matching Strategy for detecting duplicate sequences in a television stream. CVDB'05*.
- *Compact video description with precise temporal alignment*, Matthijs Douze, Hervé Jégou, Cordelia Schmid, Patrick Pérez, ECCV 10
- Exemple d'application:
  - détection de séquences répétées sur des chaînes de télé en temps réel (génériques d'émission ou publicités)

Inria

Grenoble INP ensimag

Plan

- Spécificité du problème
- Filtrage temporel
- Détecteur spatio-temporels
- Aggrégation temporelle
- Zoom sur deux méthodes d'indexation vidéo
- Indexation audio
- Quelques perspectives

Inria

Grenoble INP ensimag

Problèmes d'intérêt et perspectives

- En conclusion
  - domaine de recherche très actif car récent

    → les capacités de calcul des ordinateurs permettent depuis très peu de temps de traiter des grands volumes de données

  - Compétitions pour comparer les méthodes : Trecvid, VideOlympics