Reconnaissance d'événements dans des bases vidéo

Royer Amélie

5 septembre 2013

1. Conditions expérimentales
   - EVVE
   - Protocole
   - Décrire les vidéos
   - Objectif
2. MMV
   - Comparaison
   - Bilan
3. CTE
   - Comparaison
   - Bilan
4. R1C
   - Comparaison
   - Description formelle
   - Bilan
5. Résultats

Introduction

Problème

- Une vidéo requête ($\Leftrightarrow$ un événement)
- Une base de données vidéo

Objectif : Retrouver toutes les vidéos de l'événement contenues dans la base de données

Introduction

Problème

- Une vidéo requête ($\Leftrightarrow$ un événement)
- Une base de données vidéo

Objectif : Retrouver toutes les vidéos de l'événement contenues dans la base de données

Événement

1. « Événement d'actualité »
2. Localisé précisément dans le temps et l'espace

Introduction

3 Approches

1. Base de données EVVE
2. - **MMV** : Approche visuelle
   - **CTE** : Approche temporelle
   - **R1C** : Approche plus compacte
3. Présentation des résultats

EVVE

EVent VidEo

- 2975 vidéos
- 13 événements
- Annotées positive ou négative
- Distinction de vidéos requêtes et vidéo data

Évaluer l'efficacité de la méthode

Précision de l'algorithme

1. La vidéo requête $\mathbf{q}$ interroge la base de données
2. $\Longrightarrow$ L_q : base de données triée par ordre de similarité à $\mathbf{q}$ décroissant

Représentation vectorielle

Descripteurs

- Une image = un vecteur réel normalisé : **Descripteur**
- Peu sensibles au changement de point de vue, d'échelle, de luminosité ;

Comparer deux images

images **a** et **b** similaires $\Leftrightarrow$ $\|\vec{a} - \vec{b}\|$ faible $\Leftrightarrow$ $\langle \vec{a}, \vec{b} \rangle$ élevé

Représentation vectorielle

Notation

v : vidéo de longueur **n**,

$$v = [v_1 \dots v_n] \in \mathbb{R}^{d \times n}$$

matrice de descripteurs ;
v_i, descripteur de l'image i ;
d, dimension des descripteurs ($d = 512$).

Objectif

Créer une fonction de score **s** : $s(v, w)$ = similarité entre **v** et **w**.

- efficace
- rapide à calculer

Mean Multi Vlad

Notation

Somme des descripteurs

$$\sum_{i=1}^{n} v_i = \text{descripteur } \textbf{MMV} \text{ de la vidéo } v.$$

Mean Multi Vlad

Notation

Somme des descripteurs

$$\sum_{i=1}^{n} v_i = \text{descripteur } \textbf{MMV} \text{ de la vidéo } v.$$

Score MMV

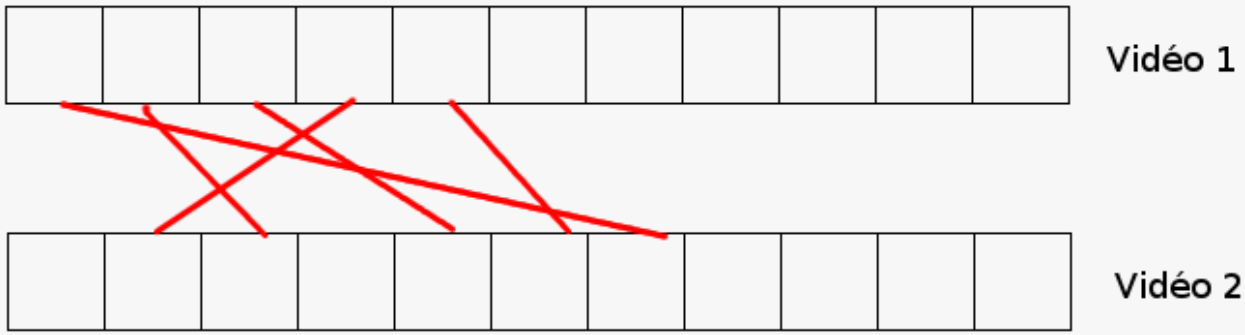
$$s(v, w) = \langle \sum_{i=1}^{n} v_i, \sum_{j=1}^{m} w_j \rangle$$

Mean Multi Vlad

Score MMV

$$s(v, w) = \langle \sum_{i=1}^{n} v_i, \sum_{j=1}^{m} w_j \rangle = \sum_{i,j} \langle v_i, w_j \rangle$$

i.e. Somme des similarités inter-images



Mean Multi Vlad

Avantages

- Représentation compacte des vidéos
- Score rapide à calculer

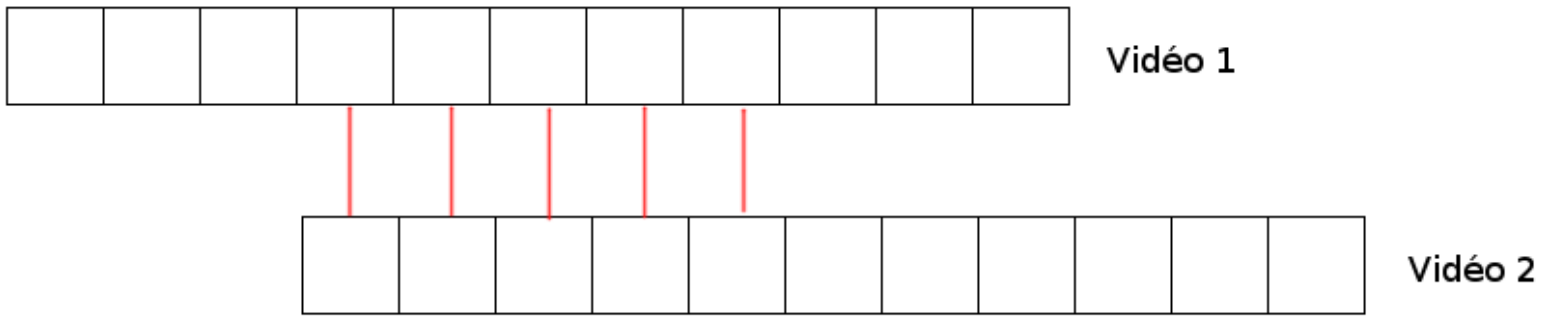
Inconvénient

- Pas d'information chronologique

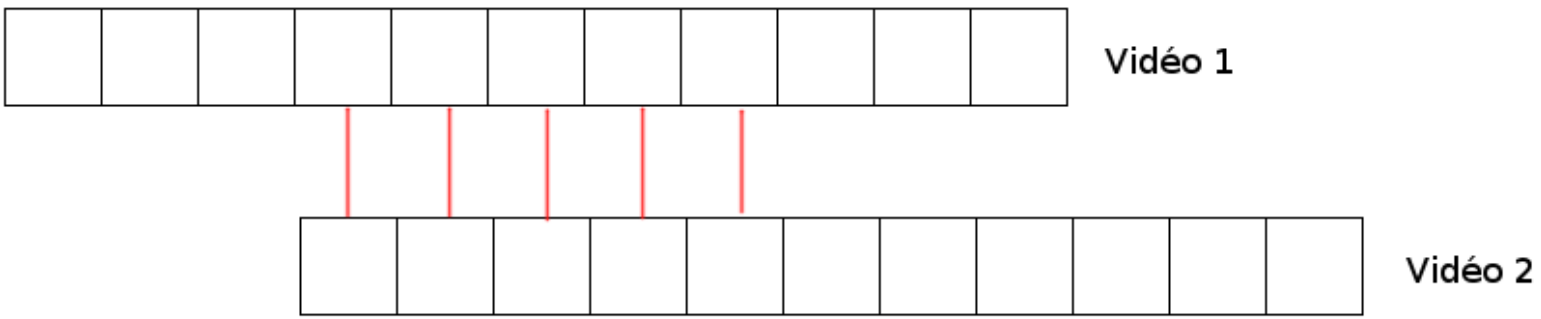
Circulant Temporal Encoding

Comparaison image par image

- $s_\delta(v, w)$: Ressemblance visuelle dans l'alignement δ
- Comparaison image à image
- Pour tous les décalages δ possibles



Circulant Temporal Encoding



$$s_\delta(v, w) = \sum_{t=-\infty}^{+\infty} \langle v_{t-\delta}, w_t \rangle \text{ , with } w_i = v_i = 0 \text{ when } i \notin [\![0; n]\!].$$

Circulant Temporal Encoding

Fonction de score finale

Ressemblance visuelle dans l'alignement optimal

$$\Longrightarrow \quad s(v, w) = \max_{\delta} s_{\delta}(v, w) \tag{1}$$

D'où+

- Calculs longs $\Longrightarrow$ perfectible
- Nécessité de garder un alignement optimal précis

Bilan de CTE

Inconvénients

- Plus lent
- Nécessite plus d'espace

Avantages

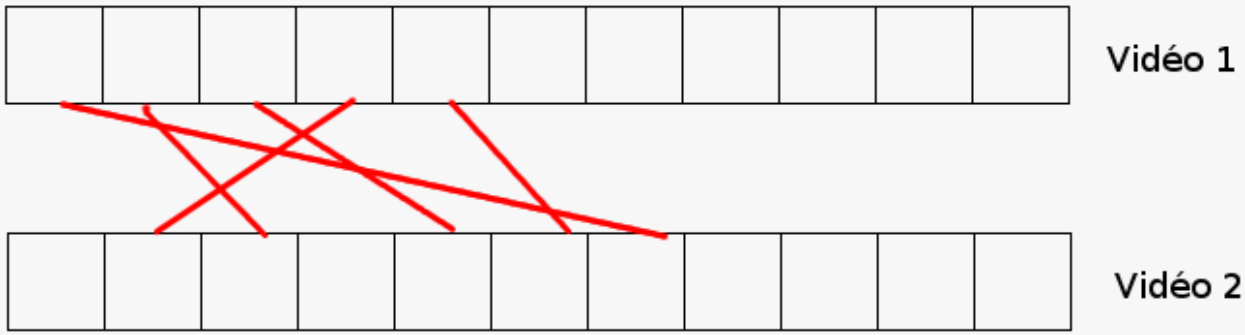
- Tient compte de l'aspect chronologique
- Plus efficace
- Améliorations possibles dans le domaine de Fourier
- Peut se combiner à MMV pour de meilleurs résultats

Random 1-Cycle

Idée générale

- Rappel de MMV (contribution de toutes les similarités)
-

Rappel MMV :

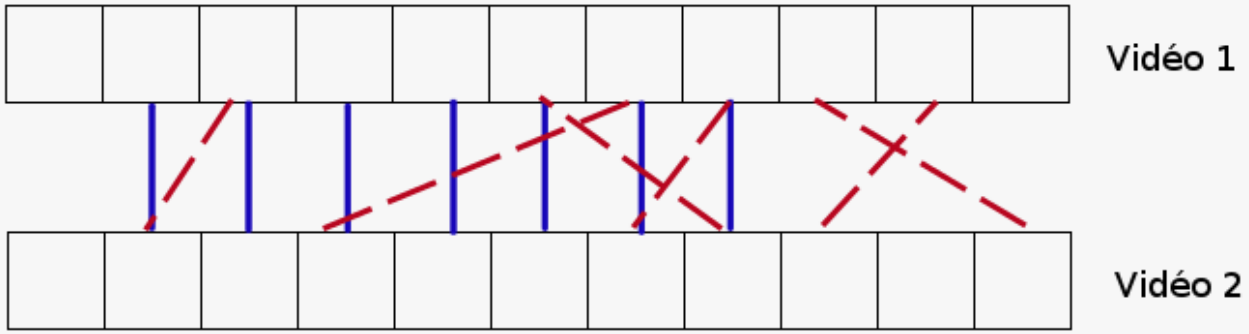


Random 1-Cycle

Idée générale

- Rappel de MMV (contribution de toutes les similarités)
- Privilégier les similarités des images alignées (—)
- Diminuer les autres similarités (images corrélées) (—)

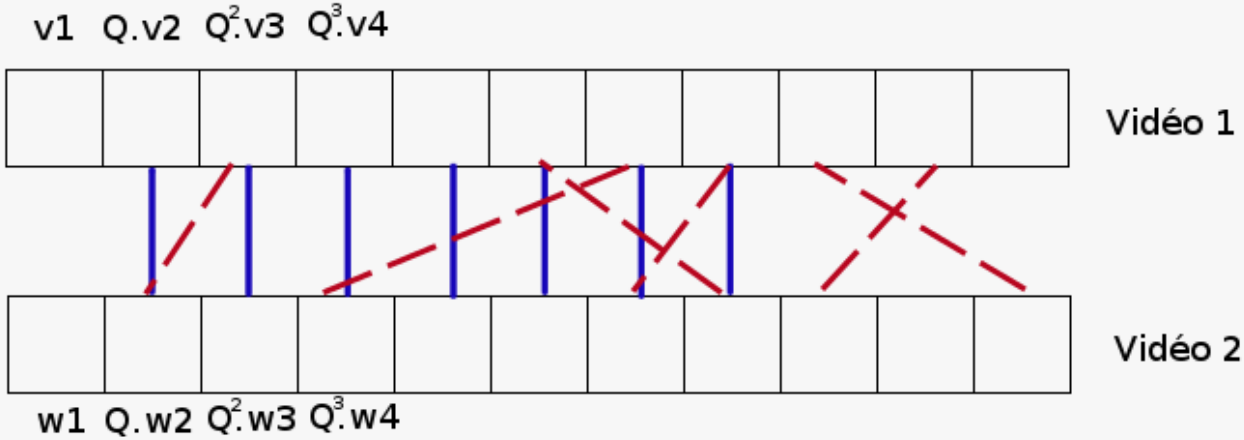
Pour retrouver l'information temporelle :



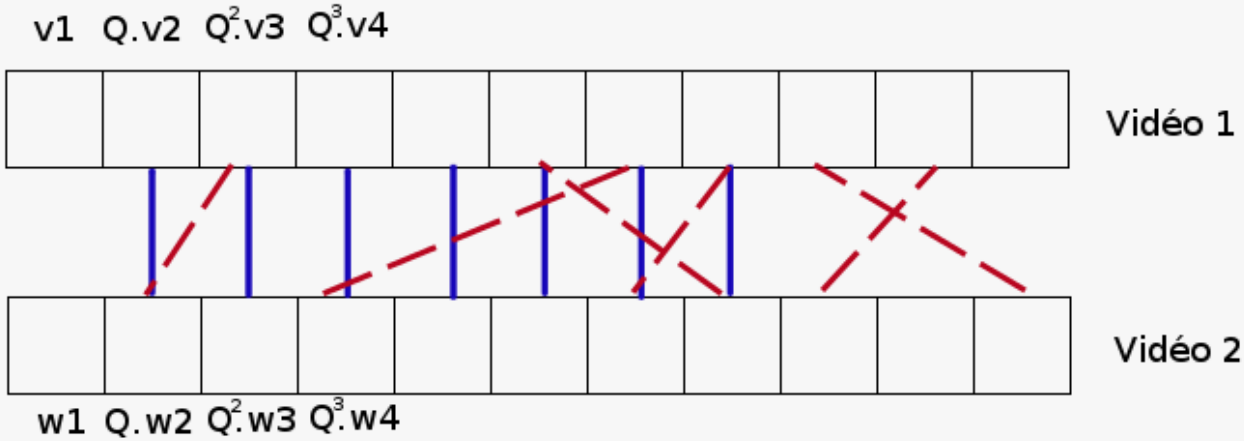
Random 1-Cycle

Influencer les similarités

- Augmenter la dimension (L) des vecteurs (ajout de 0)
- Permuter chaque descripteur différement
- Utiliser les puissances d'une permutation Q



Random 1-Cycle



- Permutations différentes $\Longrightarrow$ produit scalaire bas
- Permutations identiques $\Longrightarrow$ similarité visuelle usuelle
- Changer d'alignement $\Longrightarrow$ multiplier par Q

Formellement

- Une vidéo v : $\sum_{i=1}^{n} Q^i v_i$
- Score de l'alignement : $S_\delta(v, w) =< Q^\delta v, w >$

$$S_\delta(\mathrm{v}, \mathrm{w}) = \sum_{\substack{j=i+\delta[ord] \\ i,j\in[\![1;n]\!]}} \langle \mathbf{v_i}, \mathbf{w_j} \rangle + \sum_{\substack{j\neq i+\delta[ord] \\ i,j\in[\![1;n]\!]}} \langle \mathbf{v_i}, Q^{j-i-\delta}\mathbf{w_j} \rangle$$

$$S_\delta(\mathrm{v}, \mathrm{w}) = \mathit{Alignement} + \mathit{bruit}$$

Choix de la permutation

Permutation

1. σ (random) pour occuper la dimension
2. Q décale les coordonnées d'un cran (1-cycle)

Coût temporel et en mémoire

Comparaison à CTE

- Pré Calcul plus coûteux
- Moins d'espace (quitte à bien choisir L)
- Temps de comparaison plus court (encore une fois lié à L)

Bilan

Avantages

- Représentation potentiellement plus compacte
- Améliorations avec Fourier possibles (cf CTE)
- Combinaison avec l'aspect visuel (MMV) possible

Inconvénients

- Ne donne pas l'alignement temporel exact
- Plus de paramètres à gérer dans les améliorations

Valeurs des mAPs

	MMV	CTE norm ∞	R1C norm 5	D1C norm 3
1	0.5316	0.7447	0.7032	0.7218
2	0.3386	0.3790	0.3195	0.3322
3	0.0872	0.1265	0.1389	0.1518
4	0.4554	0.3792	0.4216	0.4249
5	0.2346	0.2969	0.1761	0.2203
6	0.2542	0.2529	0.1584	0.2127
7	0.1997	0.1773	0.2496	0.2451
8	0.1299	0.1199	0.1076	0.1421
9	0.1246	0.2521	0.1559	0.2180
10	0.3669	0.2400	0.2196	0.3053
11	0.2392	0.1557	0.1218	0.1680
12	0.7732	0.7558	0.7834	0.7513
13	0.6043	0.6488	0.6474	0.6863
avg-mAP	0.3338	0.3484	0.3233	**0.3523**

Interprétation - Temporel vs Visuel

	MMV	CTE norm ∞	R1C norm 5	D1C norm 3
1	0.5316	**0.7447**	0.7032	0.7218
2	0.3386	0.3790	0.3195	0.3322
3	0.0872	0.1265	0.1389	**0.1518**
4	0.4554	0.3792	0.4216	0.4249
5	0.2346	0.2969	0.1761	0.2203
6	0.2542	0.2529	0.1584	0.2127
7	0.1997	0.1773	0.2496	0.2451
8	0.1299	0.1199	0.1076	0.1421
9	0.1246	0.2521	0.1559	0.2180
10	0.3669	0.2400	0.2196	0.3053
11	0.2392	0.1557	0.1218	0.1680
12	0.7732	0.7558	0.7834	0.7513
13	0.6043	0.6488	0.6474	**0.6863**
avg-mAP	0.3338	0.3484	0.3233	**0.3523**

Interprétation - Contribution de R1C

	MMV	CTE norm ∞	R1C norm 5	D1C norm 3
1	0.5316	0.7447	0.7032	0.7218
2	0.3386	0.3790	0.3195	0.3322
3	0.0872	0.1265	0.1389	0.1518
4	0.4554	0.3792	0.4216	0.4249
5	0.2346	0.2969	0.1761	0.2203
emph6	0.2542	0.2529	0.1584	0.2127
7	0.1997	0.1773	0.2496	0.2451
8	0.1299	0.1199	0.1076	0.1421
9	0.1246	0.2521	0.1559	0.2180
10	0.3669	0.2400	0.2196	0.3053
11	0.2392	0.1557	0.1218	0.1680
12	0.7732	0.7558	0.7834	0.7513
13	0.6043	0.6488	0.6474	0.6863
avg-mAP	0.3338	0.3484	0.3233	**0.3523**

Interprétation - Combiner visuel et temporel

	MMV	CTE norm ∞	R1C norm 5	D1C norm 3	MMV+CTE norm ∞	MMV+D1C norm ∞
1	0.5316	0.7447	0.7032	0.7218	0.6607	0.6307
2	0.3386	0.3790	0.3195	0.3322	0.3798	0.3559
3	0.0872	0.1265	0.1389	0.1518	0.1136	0.1402
4	0.4554	0.3792	0.4216	0.4249	0.4467	0.4716
5	0.2346	0.2969	0.1761	0.2203	0.2736	0.2452
6	0.2542	0.2529	0.1584	0.2127	0.2732	0.2451
7	0.1997	0.1773	0.2496	0.2451	0.2038	0.2453
8	0.1299	0.1199	0.1076	0.1421	0.1355	0.1457
9	0.1246	0.2521	0.1559	0.2180	0.2124	0.2512
10	0.3669	0.2400	0.2196	0.3053	0.3424	0.3535
11	0.2392	0.1557	0.1218	0.1680	0.2435	0.2266
12	0.7732	0.7558	0.7834	0.7513	0.7797	0.8005
13	0.6043	0.6488	0.6474	0.6863	0.6885	0.7123
avg-mAP	0.3338	0.3484	0.3233	**0.3523**	0.3657	**0.3711**

Conclusion

Bilan des 3 méthodes

- 3 méthodes, résultats similaires
- Aspect visuel (MMV) $\Longrightarrow$ bon résultats mais néglige l'aspect "vidéo"
- Aspect temporel (CTE) $\Longrightarrow$ plus coûteux
- R1C : gain en temps, mais moins précis temporellement
- Combiner temporel et visuel donne de meilleurs résultats

Calcul de mAP

Précision

- $D = D^+ \cap D^-$ base de données annotée
- q, vidéo requête
- L_q, base de données triée dans un ordre de similarité décroissant

$\forall v \in D^+$, $Pre_q(v)$ = Nombre de vidéos positives sur nombre total de vidéos retrouvées (jusqu'à v)

$$AP(q) = \frac{\sum_{v \in D^+} \mathrm{Pre}_q(v)}{\mathrm{card}(D^+)}$$

Corrélation croisée

Définition

$$S = [\dots s_\delta(v, w) \dots]. \implies s(v, w) = \max(S)$$

En développant

$$S = \sum_{i=1}^{d} \tilde{v}_i \circledast \tilde{w}_i.$$

($\circledast$: corrélation croisée, $\tilde{}$: vecteur ligne)

En développant

$$S = \sum_{i=1}^{d} \tilde{v}_i \circledast \tilde{w}_i .$$

Puis passage dans le domaine de Fourier

$$s(v, w) = \max(S) = \max\left(\mathcal{F}^{-1}\left(\sum_{i=1}^{d} \mathcal{F}(\tilde{v}_i)^* \odot \mathcal{F}(\tilde{w}_i)\right)\right) \tag{2}$$

Corrélation croisée

Définition

$$S = [\ldots \langle Q^{\delta} v, w \rangle \ldots]. \implies s(v, w) = \max(S)$$

En développant

$$\langle Q^{\delta} v, w \rangle = \sum_i v_i \cdot w_{delta+i} = v \circledast_{circ} \tilde{w}.$$

($\circledast$: corrélation croisée, $\tilde{}$: vecteur ligne)

Cout temporel et en mémoire

Comparaison à CTE

- Pré Calcul plus coûteux
- Moins d'espace (quitte à bien choisir L)
- Temps de comparaison plus court (encore une fois lié à L)

	Database size	Time(PreComputing)	Time(Comparison)
R1C	974M	34s/event	5.9s/query
CTE	3.282G	17s/event	6.8s/query

Produit scalaire en grande dimension

Proposition (Distribution gaussienne sur la sphère unitaire)

x et y étant deux vecteurs unitaires uniformément distribués de dimension D, pour $D \to \infty$, $\langle x, y \rangle \sim \mathcal{N}(0, \frac{1}{D})$.

Lemma (Théorème limite central et projection)

$x = [x_1 \dots x_D] \in \mathbb{R}^D$ vecteur uniformément distribué de norme $\|x\| = \sqrt{D}$, Alors $x_1 \xrightarrow[D \to +\infty]{\mathcal{L}} \mathcal{N}(0, 1)$.

Produit scalaire en grande dimension

Preuve - Lemme

Etude de la densité de probabilité de x_1

$$\begin{cases} F(t) & = \mathbb{P}(-\infty \leq x_1 \leq t) = 1 - \mathbb{P}(x_1 \geq t) \\ f(t) & = -\frac{d}{dt}\left(\mathbb{P}(x_1 \geq t)\right) \end{cases}$$

Produit scalaire en grande dimension

Preuve - Lemme

Etude de la densité de probabilité de x_1

$$\begin{cases} F(t) & = \mathbb{P}(-\infty \leq x_1 \leq t) = 1 - \mathbb{P}(x_1 \geq t) \\ f(t) & = -\frac{d}{dt}\left(\mathbb{P}(x_1 \geq t)\right) \end{cases}$$

Hypothèses

x uniformément distribué.

$\mathbb{P}(x_1 \geq t)$ est le rapport de l'aire de la partie de l'hypersphère au dessus de l'hyperplan $x_1 = t$), sur l'aire totale.

Produit scalaire en grande dimension

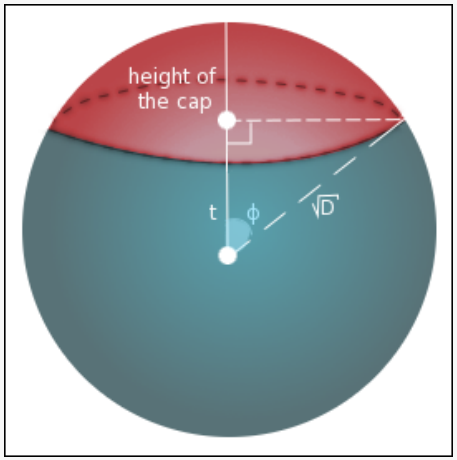
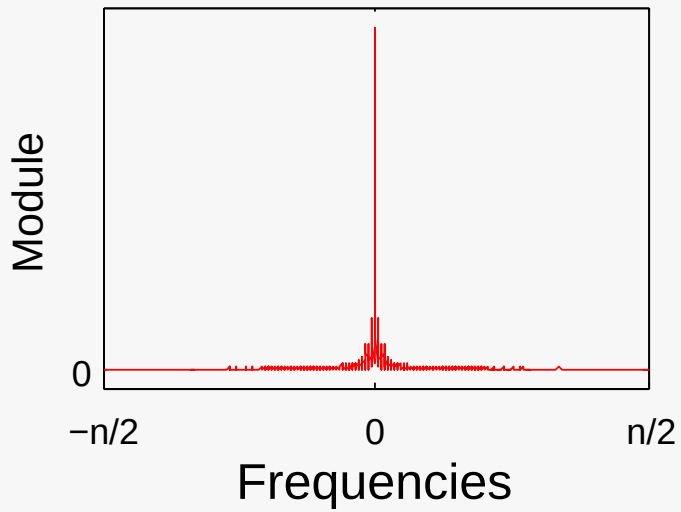


FIGURE: Schema for $D = 3$

$$f(t) \underset{D \to +\infty}{\longrightarrow} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}$$

Approximation



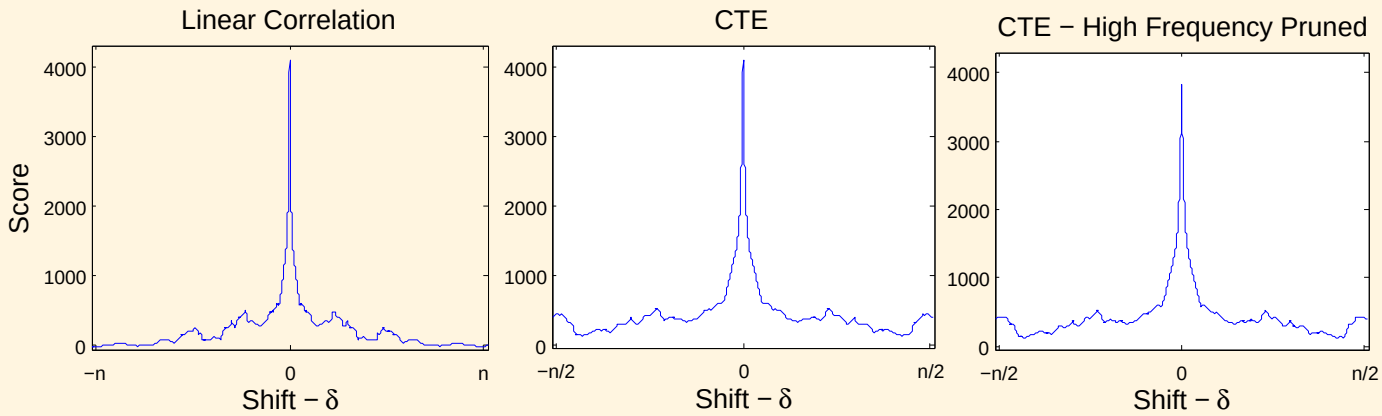
Fourier

Approximation à 0 des coefficients de haute fréquence

Préciser l'alignement

Tracé de s(v,v)

Similarité d'une vidéo à elle même s(v,v)



Optimalement $\Longrightarrow$ Dirac

Mex File

Objectifs

- Augmenter la dimension des descripteurs
- Appliquer une permutation σ des lignes sur tous les descripteurs pour répartir les coordonnées
- Permuter chaque descripteur indépendamment (Q^i)
- Sommer les descripteurs

Mex File

Remarques

- Seuls les indices des lignes des coordonnées changent
- Beaucoup de 0 dans les descripteurs

$\Longrightarrow$ Garder uniquement les positions des lignes des coordonnées non nulles