

# Energy Based Fast Event Retrieval in Video with Temporal Match Kernel

Junfu Pu<sup>1</sup> Yusuke Matsui<sup>2</sup> Fan Yang<sup>3,2</sup> Shin'ichi Satoh<sup>2,3</sup>

1. University of Science and Technology of China

2. National Institute of Informatics

3. The University of Tokyo





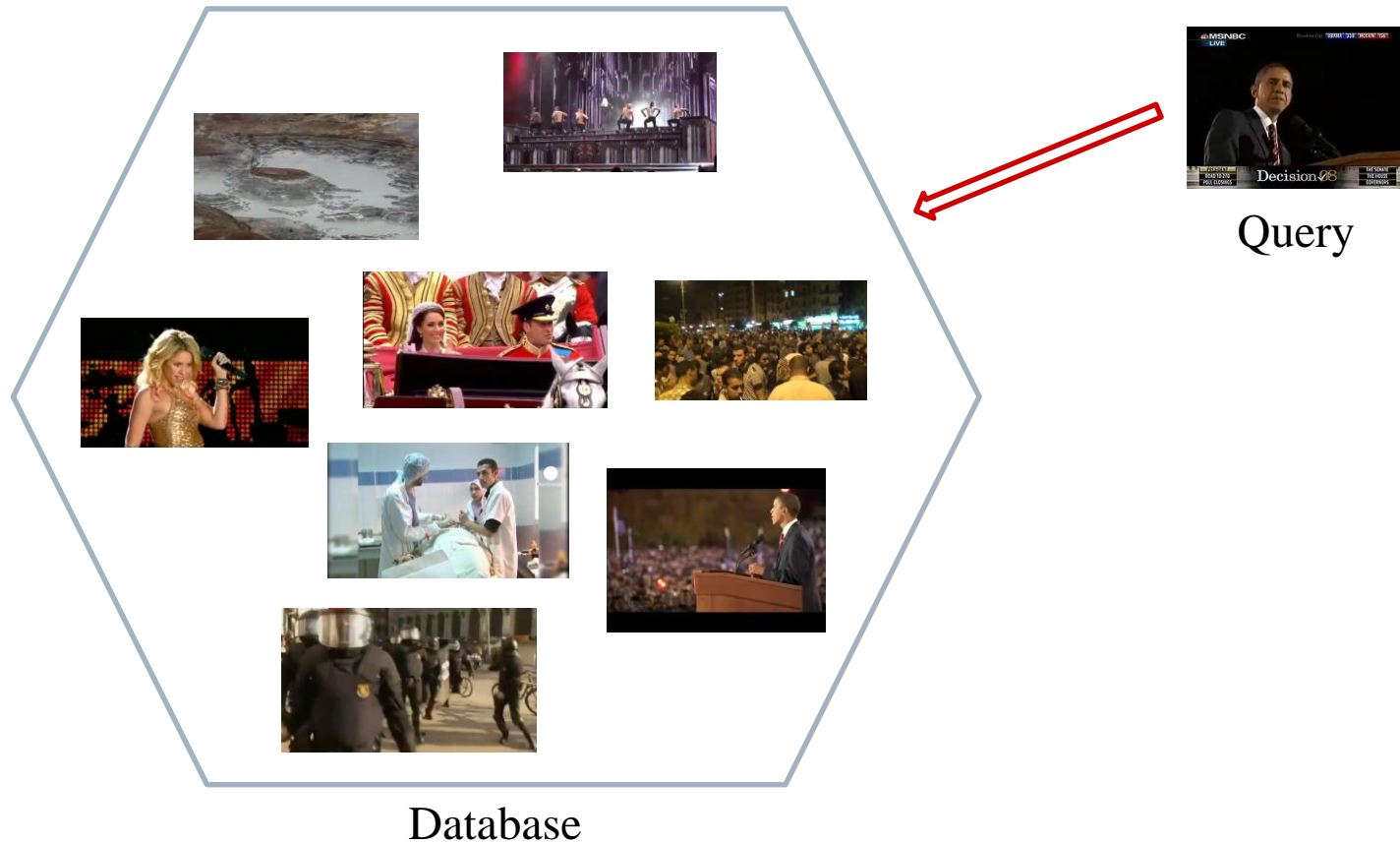
# Outline

---

- ☐ Introduction
- ☐ Background
- ☐ Matching with Energy
- ☐ Algorithm Speed up with PQ
- ☐ Experiments
- ☐ Conclusion

# Introduction

- Approach for fast content-based search in large video database



# Introduction

---

## □ Related work

- Jerome Revaud, et al., Event retrieval in large video collections with circulant temporal encoding, CVPR, 2013
- Matthijs Douze, et al., Stable hyper-pooling and query expansion for event detection, ICCV, 2013
- **Sebastien Poullot, et.al, Temporal matching kernel with explicit feature maps, ACM MM, 2015**

## □ Contribution

- Simplify the similarity metric by calculating the energy of the score function
- Derive the energy formulation by Parseval's theorem
- Accelerate the computation with product quantization

# Background

$\mathbf{x} = (\mathbf{x}_0, \dots, \mathbf{x}_t \dots)$      $\mathbf{y} = (\mathbf{y}_0, \dots, \mathbf{y}_t \dots)$     time offset:  $\Delta$

A kernel defined with  $\mathbf{x}$ ,  $\mathbf{y}$ , and  $\Delta$

$$\kappa_{\Delta}(\mathbf{x}, \mathbf{y}) \propto \sum_{t=0}^{\infty} \mathbf{x}_t^T \mathbf{y}_{t+\Delta} = \underbrace{\left( \sum_{t=0}^{\infty} \mathbf{x}_t \otimes \varphi(t) \right)^T}_{\psi_0(\mathbf{x})} \underbrace{\left( \sum_{t'=0}^{\infty} \mathbf{y}_{t'} \otimes \varphi(t' + \Delta) \right)^T}_{\psi_{\Delta}(\mathbf{y})}$$

$$\varphi(t) = \begin{bmatrix} \sqrt{a_0} \\ \sqrt{a_1} \cos(\frac{2\pi}{T} t) \\ \sqrt{a_1} \sin(\frac{2\pi}{T} t) \\ \vdots \\ \sqrt{a_m} \cos(\frac{2\pi}{T} m t) \\ \sqrt{a_m} \sin(\frac{2\pi}{T} m t) \end{bmatrix}$$

$$\psi_0(\mathbf{x}) = [\mathbf{v}_0^T, \mathbf{v}_{1,c}^T, \mathbf{v}_{1,s}^T, \dots, \mathbf{v}_{m,c}^T, \mathbf{v}_{m,s}^T]^T$$

$$\mathbf{v}_0 = a_0 \sum_{t=0}^{\infty} \mathbf{x}_t \in \mathbb{R}^D,$$

$$\mathbf{v}_{i,c} = a_i \sum_{t=0}^{\infty} \mathbf{x}_t \cos(\frac{2\pi}{T} i t) \in \mathbb{R}^D$$

$$\mathbf{v}_{i,s} = a_i \sum_{t=0}^{\infty} \mathbf{x}_t \sin(\frac{2\pi}{T} i t) \in \mathbb{R}^D$$

$a_i$ : the fourier coefficients

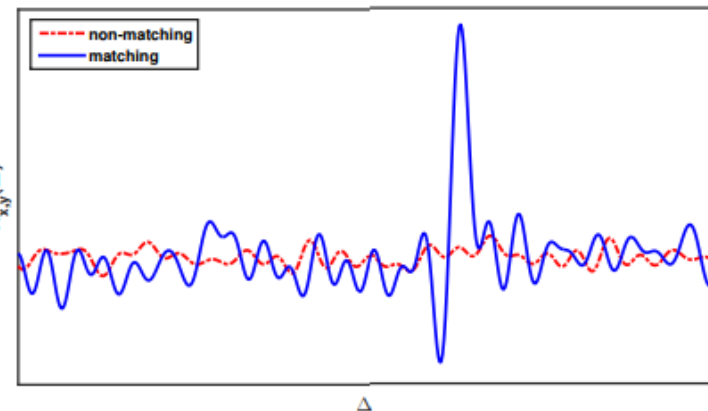
# Background

## Final Formulation

$$\begin{aligned} \kappa_{\mathbf{x},\mathbf{y}}(\Delta) = & \left\langle \mathbf{v}_0^{(\mathbf{x})}, \mathbf{v}_0^{(\mathbf{y})} \right\rangle \\ & + \sum_{n=1}^m \cos(n \Delta) \left( \left\langle \mathbf{v}_{n,c}^{(\mathbf{x})}, \mathbf{v}_{n,c}^{(\mathbf{y})} \right\rangle + \left\langle \mathbf{v}_{n,s}^{(\mathbf{x})}, \mathbf{v}_{n,s}^{(\mathbf{y})} \right\rangle \right) \\ & + \sum_{n=1}^m \sin(n \Delta) \left( -\left\langle \mathbf{v}_{n,c}^{(\mathbf{x})}, \mathbf{v}_{n,s}^{(\mathbf{y})} \right\rangle + \left\langle \mathbf{v}_{n,s}^{(\mathbf{x})}, \mathbf{v}_{n,c}^{(\mathbf{y})} \right\rangle \right) \end{aligned}$$

## Similarity Score

$$\begin{aligned} S(\mathbf{x}, \mathbf{y}) = \max_{\Delta} \kappa_{\mathbf{x},\mathbf{y}}(\Delta) \\ t_m = \arg \max_{\Delta} \kappa_{\mathbf{x},\mathbf{y}}(\Delta) \end{aligned}$$



# Our Method

## □ Matching with energy

$$E(\kappa_{\mathbf{x}, \mathbf{y}_1}) > E(\kappa_{\mathbf{x}, \mathbf{y}_2}) \quad \text{if } S(\mathbf{x}, \mathbf{y}_1) > S(\mathbf{x}, \mathbf{y}_2)$$



$$\tilde{S}(\mathbf{x}, \mathbf{y}) = E(\kappa_{\mathbf{x}, \mathbf{y}}(\Delta))$$

Denote the Fourier series of  $f(x)$  as

$$f(x) = \frac{1}{2}c_0 + \sum_{n=1}^m c_n \cos(nx) + \sum_{n=1}^m s_n \sin(nx)$$

The energy of  $f(x)$  is

$$E(f(x)) = \int_{-\infty}^{\infty} [f(x)]^2 dx$$

According to the Parseval's Theorem

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} [f(x)]^2 dx = \sum_{i=1}^n (c_i^2 + s_i^2) + c_0^2$$

# Our Method

## □ Matching with energy

The final form of the energy  $\tilde{S}(\mathbf{x}, \mathbf{y})$  for  $\kappa_{\mathbf{x}, \mathbf{y}}(\Delta)$  is

$$\begin{aligned}\tilde{S}(\mathbf{x}, \mathbf{y}) &= E\left(\kappa_{\mathbf{x}, \mathbf{y}}(\Delta)\right) \\ &= \sum_{n=1}^m \left[ \left( \left\langle \mathbf{v}_{n,c}^{(\mathbf{x})}, \mathbf{v}_{n,c}^{(\mathbf{y})} \right\rangle + \left\langle \mathbf{v}_{n,s}^{(\mathbf{x})}, \mathbf{v}_{n,s}^{(\mathbf{y})} \right\rangle \right)^2 \right]\end{aligned}$$

## □ Generalized formulation

$$\begin{aligned}S^{(p)}(\mathbf{x}, \mathbf{y}) &= \sqrt[p]{\sum_{i=1}^m (c_i^2 + s_i^2)^p} \\ S^{(\infty)}(\mathbf{x}, \mathbf{y}) &= \lim_{p \rightarrow \infty} \frac{1}{M} \sqrt[p]{\sum_{i=1}^m (c_i^2 + s_i^2)^p} = \max_n \{(c_n^2 + s_n^2)^p\}\end{aligned}$$



# Our Method

---

## □ Matching with energy

- Given a query video, go through the candidate in database
- Calculate the  $\tilde{S}(\mathbf{x}, \mathbf{y})$  between query and candidate
- Retrieval with  $\tilde{S}(\mathbf{x}, \mathbf{y})$

## □ Advantages

- More stable (maximum of  $S(\mathbf{x}, \mathbf{y})$  is sensitive to noise)
- Lower computational complexity
- Further accelerate the computation using approximate nearest neighbor method such as PQ

# Our Method

---

## □ Algorithm speedup with PQ

$j$ th codebook  $\mathbf{c}_{j*}$  generated from

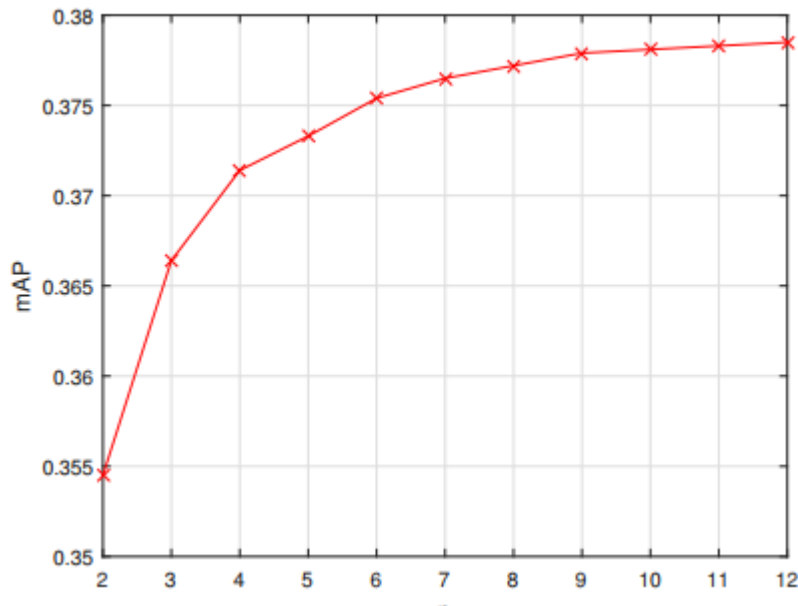
$$\left\{ \mathbf{v}_{j,c}^{(\mathbf{x}_i)} : i \in \{1, \dots, N\} \right\} \cup \left\{ \mathbf{v}_{j,s}^{(\mathbf{x}_i)} : i \in \{1, \dots, N\} \right\}$$

## □ Searching steps

- Quantize query  $q$  to its  $\omega$  nearest neighbors with  $\tilde{S}(\mathbf{x}, \mathbf{y})$
- Compute the squared distances and dot product for each subquantizer  $j$  and each of its centroid  $\mathbf{c}_{ji}$
- Using the subvector-to-centroid distance, calculate the similarity score  $\tilde{S}(\mathbf{x}, \mathbf{y})$
- Order the candidates by decreasing  $\tilde{S}(\mathbf{x}, \mathbf{y})$

# Experiments

- Event VidEo (EVVE) dataset [CVPR'13]
  - 620 queries, 2375 database videos, 13 events
  - 1024-D multi-VLAD frame descriptor
- Experimental results



$p \uparrow \rightarrow \text{mAP} \uparrow$

$$S^{(p)}(\mathbf{x}, \mathbf{y}) = \sqrt[p]{\sum_{i=1}^m (c_i^2 + s_i^2)^p}$$

The average mAP using  $S^{(p)}(\mathbf{x}, \mathbf{y})$  for different  $p$

# Experiment

## □ Results on EVVE and comparison

**Table 1:** Performance (mAP and time) on EVVE. The bold values show the best score.

Event No.	Baseline	Ours		
		$\tilde{S}$	$\tilde{S} + PQ$	$S^{(\infty)} + PQ$
#1	0.1521	0.2013	0.1985	<b>0.2483</b>
#2	0.2424	0.2503	<b>0.2621</b>	0.2133
#3	<b>0.1186</b>	0.1130	0.0651	0.0905
#4	0.1370	0.1390	0.1419	<b>0.1467</b>
#5	0.2486	0.2538	<b>0.2675</b>	0.2671
#6	0.2913	0.3189	0.3511	<b>0.3917</b>
#7	<b>0.1856</b>	0.1854	0.1177	0.1139
#8	0.2004	0.2216	0.2128	<b>0.2736</b>
#9	0.6119	0.6351	0.6276	<b>0.6728</b>
#10	0.3737	0.4519	0.4913	<b>0.5529</b>
#11	0.7979	0.7879	<b>0.8584</b>	0.8218
#12	0.2295	0.3084	0.3224	<b>0.4344</b>
#13	0.6187	0.6331	<b>0.6915</b>	0.6762
ave-mAP	0.3237	0.3461	0.3545	<b>0.3772</b>
time	9.31s	1.74s	$\approx 0.3s$	$\approx 0.3s$

**Table 2:** Comparison with state of the art.

methods	state of the art				Ours
	MMV	CTE	SHP	MMV+CTE	
ave-mAP	0.334	0.352	0.363	0.376	<b>0.377</b>

Baseline (temporal match kernel): MM'15

MMV (mean-multiVLAD): CVPR'13

CTE (circulant temporal encoding): CVPR'13

SHP (stable hyper-pooling): ICCV'13

# Conclusion

---

- Propose a fast event retrieval method in video database with temporal match kernel
- Use the energy of the score function as similarity metric
- Derive the simplified energy formulation by using Parseval's theorem
- With the energy formulation, we use PQ to accelerate the computation
- Achieve competitive performance with the-state-of-the-art

Thank you! ☺