H-025

# 時間情報を埋め込んだ新たな映像特徴についての理論的検討
# Theoretical Study of Data Compression of a New Video Feature with Temporal Embedding

塚谷俊介 † **Poullot Sebastien**‡ **Jégou Hervé**§ 佐藤 真一 ‡
Shunsuke Tsukatani Poullot Sebastien Jégou Hervé Shin'ichi Satoh

## 1. はじめに

インターネットの発達とスマートフォンの急速な普及によって, 映像データの数が爆発的に増加している. このような莫大な映像データをビッグデータとして取り込んで活用するためには, 目的とする内容を持つ映像を検索し, 取得する技術が必要不可欠である.

今回は特に時間的整合性の保持が必要とされるコピー検出 [1] や特定イベント検索 [2] に焦点を当てる. これらのタスクに対して, これまでの局所画像特徴量や時空間点を直接ストアしていく手法 [1] ではその巨大なデータセットに対してあまりにもコストがかかる.

そこで我々は, 時間不定長な映像を対象とし, オフセットを考慮しつつ時間的整合正を保持し, 映像内容のローカライズ化を行うコンパクトなフレームベクトル表現手法を開発した. しかし, この手法にはスケーラビリティの性能に問題があり, 改善の余地が残されている. そこで本研究では, 検索精度を維持しつつ省メモリな特徴の圧縮方法について検討する.

## 2. 関連研究

### Explicit feature map

Explicit feature map [3] は BoVW における非線形カーネルの代わりに, 高次元空間への写像を明示的に書き下すことによって直接線形識別器を使えるようにするというものである. 非線形カーネル $k(\cdot,\cdot) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ を $\varphi : \mathbb{R} \to \mathbb{R}^M$ を用いて次のように推定する.

$$k(x, y) \approx \langle \varphi(x) | \varphi(y) \rangle \tag{1}$$

今, このカーネルは時不変系のカーネルであると仮定する. つまり, カーネルは $k(x, y) = g(x - y)$ のように表すことが出来るとする. ここで $g$ をフーリエ級数を用いた近似としてみると,

$$g(z) = \sum_{i=0}^{m} a_i \cos\left(\frac{2\pi}{T} iz\right) \tag{2}$$

このように $g$ は周期 $T$ の関数として扱うことができる. ここで, $z = x - y$ として置き換える事を考える. 三角関数の加法定理を用い, さらにフーリエ級数の全ての係数がすべて正の値を取ることを仮定する. すなわち $\forall i\ a_i > 0$ に対して (2) は次のように書き直すことが出来る.

$$g(x-y) \approx \underbrace{\begin{bmatrix} \sqrt{a_0} \\ \sqrt{a_1}\cos\left(\frac{2\pi}{T}x\right) \\ \sqrt{a_1}\sin\left(\frac{2\pi}{T}x\right) \\ \vdots \\ \sqrt{a_m}\cos\left(\frac{2\pi}{T}mx\right) \\ \sqrt{a_m}\sin\left(\frac{2\pi}{T}mx\right) \end{bmatrix}^{\top}}_{\varphi(x)^{\top}} \underbrace{\begin{bmatrix} \sqrt{a_0} \\ \sqrt{a_1}\cos\left(\frac{2\pi}{T}y\right) \\ \sqrt{a_1}\sin\left(\frac{2\pi}{T}y\right) \\ \vdots \\ \sqrt{a_m}\cos\left(\frac{2\pi}{T}my\right) \\ \sqrt{a_m}\sin\left(\frac{2\pi}{T}my\right) \end{bmatrix}}_{\varphi(y)} \tag{3}$$

$\varphi(x)$ は $y$ に対して独立に計算することが出来るという特性が見て取れる. したがって, $\varphi()$ によるマッピングの構築が (1) であげた時不変のカーネルの推定に利用することが可能である.

## 3. 提案手法

### 3.1.Temporal match kernels

今, 画像フレーム $x$ についてタプル $(f_x, t_x)$ と表すことにする. ここでは $f_x \in \mathbb{R}^d$ は $d$ 次元のベクトルであり, $t_x$ はスカラー値のタイムスタンプである. 二つの画像フレームの間で定義されるカーネルは次のように表される

$$k\left((x, t_x), (y, t_y)\right) = k_{\mathrm{f}}(f_x, f_y)\, k_{\mathrm{t}}(t_x, t_y) \tag{4}$$

$$= \langle f_x | f_y \rangle\, k_{\mathrm{t}}(t_x, t_y) \tag{5}$$

$$\approx \langle f_x | f_y \rangle\, \varphi(t_x)^{\top} \varphi(t_y) \tag{6}$$

ここではフレーム記述子は正規化されており, コサイン類似度で比較するとする. つまり, $k_{\mathrm{f}}(f_x, f_y) = \langle f_x | f_y \rangle$ と表す. このカーネルをさらに次のように線形化する.

$$k\left((x, t_x), (y, t_y)\right) \approx \langle f_x \otimes \varphi(t_x) | f_y \otimes \varphi(t_y) \rangle \tag{7}$$

ここで $\otimes$ はクロネッカー積である. この場合, タプル $(f_x, t_x)$ は一つのベクトル $f_x \otimes \varphi(t_x)$ によって表され, このフレーム記述子を内積で比較する. このフレーム記述子の別の解釈として, この結合符号化はフレーム記述子 $f_x$ を連続するタイムスタンプ $t_x$ で変調していると見ることが出来る. よって, 同じタイムスタンプのフレーム記述子のみが重要な値を取る. つまり $|k((x, t_x), (y, t_y))|$ は $|k_{\mathrm{t}}(t_x, t_y)|$ にのみ制約される.

†東京大学大学院情報理工学系研究科
‡国立情報学研究所
§INRIA

### 3.2.Temporal match kernels with feature map

式 (7) は Bo ら [4] のカーネル記述子を元にして導き出された. ここでは efficient match kernels [5] の代わりに explict feature map [3] に基づいて考える. ここでは二つの連続しているタイムスタンプ ¶ がエンコードされた記述子 $\mathbf{x} = (x_0, \ldots, x_t, \ldots)$ と $\mathbf{y} = (y_0, \ldots, y_{t'}, \ldots)$ の比較を行う. いま, カーネルを下記の配列に近似することを考える

$$\mathcal{K}_0(\mathbf{x}, \mathbf{y}) = \alpha(\mathbf{x})\alpha(\mathbf{y}) \sum_t k_{\mathrm{f}}(x_t, y_t) \tag{8}$$

$$\propto \sum_{t=0}^{\infty} x_t^\top y_t \tag{9}$$

比例係数は $\mathcal{K}_0(\mathbf{x}, \mathbf{x}) = \mathcal{K}_0(\mathbf{y}, \mathbf{y}) = 1$ として計算する. シーケンスが $t$ よりも小さい場合は $x_t = \mathbf{0}$ として扱う.

このカーネルは次式と同等である

$$\mathcal{K}_0(\mathbf{x}, \mathbf{y}) \propto \sum_{t=0}^{\infty} \sum_{t'=0}^{\infty} x_t^\top y_{t'}\, \delta(t, t') \tag{10}$$

ここで $t = t'$ ならば $\delta(t, t') = 1$, その他は $\delta(t, t') = 0$ である. デルタ関数 $\delta(\cdot, \cdot)$ を任意の temporal kernel $k_{\mathrm{t}}(\cdot, \cdot)$ に置き換える. 式 (7) を用いることで, カーネルは下記のように因数分解できる,

$$\mathcal{K}_0(\mathbf{x}, \mathbf{y}) \propto \underbrace{\left(\sum_{t=0}^{\infty} x_t \otimes \varphi(t)\right)^\top}_{\varphi_0(x)^\top} \underbrace{\left(\sum_{t'=0}^{\infty} y_{t'} \otimes \varphi(t')\right)}_{\varphi_0(y)}, \tag{11}$$

$\varphi_0(x)$ はシーケンス $x$ のベクトルで表現される.

### 3.3.PCA for frame descriptors

式 (11) において $\varphi_0(x)$ の中身を展開してみる.

$$x_t = [x_t(1), x_t(2), \cdots, x_t(d)] \tag{12}$$

とおくと

$$\sum_{t=1}^{n} x_t \otimes \varphi(t) = \begin{bmatrix} x_1(1)\varphi(1) & +\cdots+ & x_n(1)\varphi(n) \\ \vdots & \vdots & \vdots \\ x_1(d)\varphi(1) & +\cdots+ & x_n(d)\varphi(n) \end{bmatrix} \tag{13}$$

各 $x_t$ に対して $A \in \mathbb{R}^{d \times k}$ による主成分分析を施すことで $\varphi_0(x)$ の計算量を減らすことが出来る. ところで、この主成分分析の結果に着目すると,

$$\underbrace{\left[\sqrt{a_0}x_t, \cdots, \sqrt{a_m}\sin\left(\tfrac{2\pi}{T}mx\right)x_t\right]}_{2m+1} \underbrace{\begin{bmatrix} A & & 0 \\ & A & \\ & & \ddots \\ 0 & & A \end{bmatrix}}_{B} \tag{14}$$

この上記式と同値の計算を施していることになる. この PCA の主成分の対角行列 $B$ について各対角成分が今後のスケーラビリティに関わるパラメータとなる.

### 3.4. 主成分分析の三重対角行列

ここで各周波数成分が他の周波数成分に与える影響について考える. 周波数成分に対して主成分分析を行う対角行列 $B$ について, 主対角線とその上下に隣接する対角線にだけ非零の成分を持つ行列である三重対角行列 $B'$ について考えると

$$\underbrace{\begin{bmatrix} A & c_1 & & & 0 \\ a_2 & A & c_2 & & \\ & a_3 & A & \ddots & \\ & & \ddots & \ddots & c_{2m} \\ 0 & & & a_{2m+1} & A \end{bmatrix}}_{B'} \tag{15}$$

この $B'$ は, 隣接する周波数成分の影響を加味した映像特徴ベクトルを生成する行列である. この $B'$ のパラメータ $a, b$ によって映像特徴ベクトルの周波数特性を探ることとなる.

## 4. 終わりに

本稿では, Temporal matching kernel を用いて映像を検索する手法について述べたあと, 各フレーム記述子を主成分分析にかけることが主成分からなる対角行列をかけることと同義であることを確かめ, 主成分分析による次元削減でスケーラビリティの確保と検索性能の両立を提案した. また, この新たな映像特蝶の周波数特性を調べるために三重対角行列を用いて各周波数成分の影響を測ることを提案した. 今後の課題として, 映像特徴量の周波数特性の調査を行いたい.

## 参考文献


[1] Law-To, Julien, et al. ”Video copy detection: a comparative study.” CIVR, 2007.

[2] Revaud, Jerome, et al. ”Event retrieval in large video collections with circulant temporal encoding.” CVPR, 2013.

[3] Vedaldi, Andrea, and Andrew Zisserman. ”Efficient additive kernels via explicit feature maps.” PAMI, 2012.

[4] Bo, Liefeng, Xiaofeng Ren, and Dieter Fox. ”Kernel descriptors for visual recognition.” NIPS, 2010.

[5] Bo, Liefeng, and Cristian Sminchisescu. ”Efficient match kernel between sets of features for visual recognition.” NIPS, 2009.


¶タイムスタンプは必ずしも整数値を取る必要はなく, いかなる実数値を取ることが可能である