An Introduction to Video Fingerprinting

Speaker: Wei-lun Chao  
Advisor: Prof. Jian-jiun Ding

DISP Lab  
Graduate Institute of Communication Engineering  
National Taiwan University, Taipei, Taiwan

Abstract

- A video copy detection technique
- A new way of thinking against watermarking
- Techniques of it are not too complicated until now
- Still lots of space to be improved
1. Introduction

What is video copy detection and its goal?
- For database and copyright management
- Copy vs Similarity

Two main direction:
- Content-based video copy detection (CBCD)
- Watermarking
**Watermarking**

- Watermarking relies on inserting a distinct pattern into the video stream.
- It can be simply classified into visible and invisible watermarking.
- A widely included research.

 CBCD

- Pertinent features are extracted as "fingerprints" or "signatures" of the video.
- Comparing the signatures to determine a copy or not.
- "The media itself is the watermark"
1. Comparison

<table>
<thead>
<tr>
<th>Difference/Technique</th>
<th>Watermarking</th>
<th>CBCD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embed something</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Main protecting process</td>
<td>Before propagation</td>
<td>After propagation</td>
</tr>
<tr>
<td>Additional memory</td>
<td>Yes (Key)</td>
<td>Yes (Feature database)</td>
</tr>
<tr>
<td>Matching step</td>
<td>Detection + Extraction + Comparison</td>
<td>Extraction + Comparison</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Watermarking</td>
<td>Multi-functional</td>
<td>Compare the embedded pattern, Afraid of attacking</td>
</tr>
<tr>
<td>CBCD</td>
<td>Fast search, no insertion, HVS preserved</td>
<td>Additional data memory, Database standard</td>
</tr>
</tbody>
</table>

2. Challenge

- **People feel the same, while computers don’t!**
- **Distortions**
  - Luminance
  - Geometrical modification
  - Compression & different formats
  - Post-processing
  - Malicious attack
- **CBCD V.S CBVR (content-based video retrieval)**

National Taiwan University, Taipei, Taiwan
DISP Lab @ MD531
3. Basic Structure

- There are three important properties a video fingerprinting technique should have!
  - Robustness: distortion tolerance
  - Pair-wise independent: identification
  - Database search efficiency: practical
System structure

- **Three parts of the system:**
  - Database operation (off-line)
  - Query operation (on-line)
  - Matching (on-line)

- **Functional blocks:**
  - Shot detection
  - Key-frame extraction
  - Feature extraction
  - Matching & decision

- **The most important one:**
  - Feature extraction + arrangement

Feature extraction

- **Difference between images and videos:** Time
- **3 dimensions to explore information (feature):**
  - Color
  - Spatial
  - Temporal

- **2 kinds of descriptor:**
  - Global
  - Local

- **Arrangement:** features -> a signature
Matching & Decision

- **Three steps:**
  - Searching
  - Voting algorithm: best match
  - Threshold: Is the best match a real match?

- Local description case is more complicated to deal with at this step:
  Temporal and spatial registration for each point

---

Is there any similarity?

- **Asymmetric relation!!!**
4. Existed Techniques

- **Four functional blocks:**
  - Shot detection
  - Key frame extraction
  - **Feature extraction**
  - Matching & decision

- **Feature extraction:**
  - global/local descriptors
  - color/spatial/temporal
  - image based/ sequence based
4.2 Global features

- Shot boundary
- Color
  Different color space histogram, moment, dominant color
- Changes in or between frames
  Motion, gradient
- Ordinal feature
  Spatial, temporal ordinal
- Transform-based

The signature for a shot: “A 125 x 8 matrix”
**Motion direction histogram**

![Diagram of motion direction histogram]

EX: 15x15 blocks per frame

\[ S_m(t) = q_0(t), q_1(t), q_2(t), q_3(t), q_4(t) \]

\[ q_i(t) \in \{0, \ldots, 225\} \quad i=0, \ldots, 4 \]

National Taiwan University, Taipei, Taiwan
DISP Lab @ MD531

**Centroids of gradient orientation**

![Diagram of centroid calculation]

\[
m(x, y) = \sqrt{G_x^2 + G_y^2}
\]

\[
\theta(x, y) = \tan^{-1}(G_y/G_x)
\]

\[
C = \frac{\sum_{x=2}^{X-1} \sum_{y=2}^{Y-1} \theta(x,y)m(x,y)}{\sum_{x=2}^{X-1} \sum_{y=2}^{Y-1} m(x,y)}
\]

The signature for each frame: “MN CGO scalars”

National Taiwan University, Taipei, Taiwan
DISP Lab @ MD531
Ordinal feature

- Spatial ordinal
  For each frame: \( S(t) = (r_1, r_2, \ldots, r_N) \)

- Temporal ordinal
  For each video clip:
  A 72-bin histogram

Compact Fourier-Mellin transform

Original formula:
\[
T_f(k, \nu) = \frac{1}{2\pi} \int_0^{2\pi} \int_0^\infty \! f(r, \theta) r^{-iv} e^{-ik\theta} \, d\theta \, \frac{dr}{r}
\]

AFMT:
\[
T_{f,\sigma}(k, \nu) = \frac{1}{2\pi} \int_0^{2\pi} \int_0^\infty \! f(r, \theta) r^{\sigma-iv} e^{-ik\theta} \, d\theta \, \frac{dr}{r}
\]

RST-invariant property:
\[
(y) = f(\alpha(xcos\beta+ysin\beta), \alpha(-xsin\beta+ycos\beta))
\]
\[
T_{g,\sigma}(k, \nu) = \alpha^{-\sigma} e^{ik\beta} T_{f,\sigma}(k, \nu)
\]
\[
| T_{g,\sigma}(k, \nu) | = | \alpha^{-\sigma} | T_{f,\sigma}(k, \nu) |
\]

Fast AFMT approximation:
\[
T_{f,\sigma}(k, \nu) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_0^{2\pi} e^{\sigma f(e^t, \theta)} e^{-i(k\theta + tv)} \, d\theta \, dt
\]
Compact Fourier-Mellin transform

PCA (principal component analysis)

Each frame: d-dim vector
Comparison of global descriptors

- Recall: What we want to solve, page 8

<table>
<thead>
<tr>
<th>Technique</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shot boundary</td>
<td>Compact, only frame index</td>
<td>Algorithm, only ok for whole movies</td>
</tr>
<tr>
<td>Color</td>
<td>Easy to implement</td>
<td>High noise sensitivity (inter video not intra video)</td>
</tr>
<tr>
<td>Spatial/temporal change</td>
<td>More information included</td>
<td>Computational complexity</td>
</tr>
<tr>
<td>Ordinal feature</td>
<td>Immune to global color change</td>
<td>Sensitive to post-processing</td>
</tr>
<tr>
<td>Transform-based</td>
<td>RST-invariant</td>
<td>Computational complexity, high feature dimension</td>
</tr>
</tbody>
</table>

- General problem: Post-product processing

4.3 Local features

- Additional work: Points-of-interest extraction
  Ex: Harris point, SIFT

- Method:
  - A. Joly method
  - Space time interest points
  - Video Copy tracking (ViCopT)
  - Scale invariant feature transform (SIFT)
A. Joly Method

- Harris point detector
- Key-frame based
- Signature for each point is a 20-dimensional vector in $[0,255]^{D=20}$:

$$\begin{align*}
\vec{S} &= \left( \frac{s_1}{\|s_1\|}, \frac{s_2}{\|s_2\|}, \frac{s_3}{\|s_3\|}, \frac{s_4}{\|s_4\|} \right) \\
\vec{s_i} &= \left( \frac{\partial \tilde{i}}{\partial x}, \frac{\partial \tilde{i}}{\partial x'}, \frac{\partial^2 \tilde{i}}{\partial x \partial y}, \frac{\partial^2 \tilde{i}}{\partial x' \partial y}, \frac{\partial^2 \tilde{i}}{\partial y^2} \right)
\end{align*}$$

at four nearby locations

ViCopT

- Try to find the trajectory of points of interest

Signal description $\vec{s}_{\text{mean}}$

Trajectory description $s_{\text{traj}}$

[ $t_{\text{Cin}}, t_{\text{Cout}}$ ] $[x_{\text{min}}, x_{\text{max}}]$ $[y_{\text{min}}, y_{\text{max}}]$
ViCopT

Labeling:
- label Background: motionless and persistent points along frames
- label Motion: moving and persistent points

#### Asymmetric Technique:
- Off-line: Trajectories
- On-line: Points of interest in key frames

Latter steps:
- Similarity searching
- Spatio-temporal registration
- Combination of labels
ViCopT framework

SIFT –
Scale invariant feature transform

- Using the SIFT-point detector (128-dim vector)
- Computational complexity due to pair-wise comparison
SIFT

- With location, scale, and orientation

16 regions, each with 8 orientations, totally 128-dim vector

National Taiwan University, Taipei, Taiwan
DISP Lab @ MD531
SIFT

- In order to simplify the matching step, the quantization (codebook) method is used for each point and the signature for each frame is a d-bin histogram.
- Each descriptor is quantized into a codeword based on the trained codebook.

4.4 Matching idea

- **Matching type:**
  - Sequential matching
  - Vector matching
  - Local point matching: intersection idea

- **Matching method:**
  - L1, L2 distance
  - Pattern recognition or machine learning idea
4.5 Complete comparison

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global descriptor:</td>
<td>Simple computation</td>
<td>Not robust for logo insertion, cropping, and post-product processing</td>
</tr>
<tr>
<td>(color, ordinal...)</td>
<td>Easy matching</td>
<td></td>
</tr>
<tr>
<td>Local descriptor:</td>
<td>Robust to post-product processing, and can be used for CBVR</td>
<td>Complex matching and decision (pair-wise, intersection...)</td>
</tr>
<tr>
<td>(ViCopT, SIFT...)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

More dimensions (color, spatial, temporal), more robust.

Local descriptor is stronger than global descriptor especially for logo and word insertion.

Ordinal recording is better than actual value recording.

It’s better to record the direction of an image by some algorithms in order to deal with image rotation.
5. Performance Evaluation

- **Precision-recall curve (P-R curve)**
  
  Recall = \( \frac{TruePositive\ (pa)}{NlITure} \)
  
  Precision = \( \frac{TruePositive\ (pa)}{NAllPosti(pa)} \)
  
- **F1 score**
  
  \[ F = \frac{2 \times P \times R}{P + R} \]

6. Our Works

- **Database:**
  
  2600 scenes, 100 for duplication. 63 duplicated videos are created, then totally we have \(6300 + 2600 = 8900\) videos (frame drop, blurring, AWGN, JPEG, Gamma correction)
  
- **Shot detection**
  
  Software
  
- **key-frame detection**
  
  Spatial temporal color distribution
  
- **Feature extraction**
  
  YChCr histogram, ordinal, CGO, CFMT, and SIFT
  
- **Matching & detection**
  
  L1, L2 distance
7. Conclusion & Future works

- **Is there any practical case of CBVD?**
  Youtube website
- **Another comparison**
<table>
<thead>
<tr>
<th>Techniques</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Watermarking</td>
<td>Malicious attack, for example, put fake patterns inside! Need some knowledge of cryptography and coding</td>
</tr>
<tr>
<td>Video fingerprinting</td>
<td>Could perform kinds of method on one video!!! Could use lots of image and video ideas</td>
</tr>
</tbody>
</table>
- **The importance of feature selection: V.S CBIR**
Future works

- New features
- Higher speed
- Put into practice
- Deal with more complicated cases
- While, there are practical cares!

8. Acknowledge

- I’d like to thank for my advisor, Prof. Pierre Moulin, and group members, Julien Dubois and Ryan Rogowski, in UIUC!!
9. Reference


