Bag-of-Words Computer Vision Methods for Forensic Applications

February Fourier Talks
The Norbert Wiener Center for
Harmonic Analysis and Applications
University of Maryland
February 18, 2011

Jeff Woodard
MITRE CORP, McLean, VA

Approved for Public Release: 10-4252
Distribution Unlimited
Problem addressed

- Forensic & biometrics: writer identification

  - Computer studies suggest a scientific basis for handwriting comparison, at least in the absence of intentional obfuscation or forgery
  - The scientific basis for handwriting comparisons needs to be strengthened
Summary of Past Automated Studies

- Strongly influenced by human document examiners
- Limited publically available reports and data bases
- Often manual supervision or preprocessing needed
- Features are linguistic and/or geometrically based
Initial Processing

Paper Document

Photo & Digitization

Scanners
Digital cameras, etc.

Digital Image

1 255 0 11 128 9 …
10 55 0 31 228 9 …
1 255 6 11 108 0 …
21 155 7 1 128 3 …
111 5 50 111 8 9 …
Bag-of-Words Approaches

- Popular in computer vision applications:
  - Object recognition & detection
  - Video tracking
  - Other biometrics, e.g. face, iris, fingerprint

- Bag-of-words methods:
  - Do not measure classical document features
  - Instead, extract “local features” to model local shape
  - Are completely automated and unsupervised
Bag of Words Algorithmic Components

- Local Features: Scale Invariant Feature Transform (SIFT)
- Vector Quantization (VQ)
- Either:
  - probabilistic Latent Semantic Analysis (pLSA)
  - Spatial Pyramid Matching
First learning...

Images: Class 1

Local Features

image representations

Image or Class Models

Class 1  ...  Class N

...Then recognition

Local Features

Clustering Codebook

Class Unknown

Adapted from L. Fei-Fei (UIUC, 2007)
Local Features
Local Features

- Popularized by D. Lowe, U. British Col.
- Represent distinctive shapes of small regions without segmentation

- Moderately invariant or robust to: Rotation, Scale, Viewpoint, Noise, Occlusions

Figure credit: adapted from Torralba, MIT
Local Features: Invariance

Local feature 1 detected

Transform

Descriptor 1

\{99 \ 52 \ 16 \ 24 \ldots\}

\cong

Descriptor 2

\{98 \ 52 \ 16 \ 23 \ldots\}

Local feature 2 detected
Models of Image Change

Geometry
- Rotation
- Similarity (rotation + uniform scale)
- Affine

Photometry
- Affine intensity change ($l \rightarrow a \ l + b$)

Lowe’s Scale Space
- Laplacian of Gaussian provides scale invariance
- Approximate with Difference of Gaussians (DoG)

```
\sigma(s, o) = \sigma_0 \cdot 2^{(s/S + o)}, \quad \text{DoG}(o,s) = I_{o(o,s+1)} - I_{o(o,s)}
\sigma_0 = 1.6, s = 0, \ldots, S - 1, o = o_{\text{min}}, \ldots, o_{\text{min}} + O - 1
s : \# \text{ of scales per octave}, \quad O : \# \text{ of octaves}
```

Adapted from A. Torralba, MIT, 2010
Lowe’s Scale Space

<table>
<thead>
<tr>
<th>Scale 1</th>
<th>Scale 2</th>
<th>Scale 3</th>
<th>Scale 4</th>
<th>Scale 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
</tr>
</tbody>
</table>

Octave 1

Octave 2

Octave 3

Gaussian Representation

Find Extrema

Difference of Gaussians Representation

© 2011 The MITRE Corporation. All rights reserved
Local Features

Original Image

Circle center: feature x-y position
Circle radius: feature scale
Circle arrow: dominant orientation(s)

DoG-SIFT features

© 2011 The MITRE Corporation. All rights reserved
Vector Quantization
Vector Quantization (VQ)

- i.e., K-means, Generalized Lloyd clustering
- Hierarchical or tree-structured for speed
- Many different local descriptors quantized to small codebook -- "visual words"
- Encode each visual word with unique index
- Images characterized by histogram of indexes
VQ “VISUAL WORDS”

100 random image regions from Arabic, DoG detector, SIFT descriptor, 512 codewords

Region sizes vary: all displayed identically

Regions for VQ codeword 1

Regions for VQ codeword 2

Regions for VQ codeword 3

“Visual words” represent fundamental properties associated with early to middle vision
Bag of Words Representation

Image

Local Features

Histogram

VQ encoding

Counts

VQ index

Counts

VQ index
Classification by pLSA


probabilistic Latent Semantic Analysis (pLSA):
Text Documents

Documents modeled as combinations of latent topics

Generative View
- Select a document $d_i$ with prob $P(d_i)$
- Pick latent class $z_k$ with prob $P(z_k|d_i)$
- Generate keyword $w_j$ with prob $P(w_j|z_k)$
- Boxes replicate

Bag of Words

Count Keywords

<table>
<thead>
<tr>
<th>Index</th>
<th>Word</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Filter</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>Particle</td>
<td>2</td>
</tr>
<tr>
<td>j</td>
<td>Method</td>
<td>12</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>Result</td>
<td>1</td>
</tr>
</tbody>
</table>
pLSA for Images

Images modeled as combinations of latent objects

Generative View
- Select an image \( d_i \) with prob \( P(d_i) \)
- Pick latent object \( z_k \) with prob \( P(z_k|d_i) \)
- Generate VQ index \( w_j \) with prob \( P(w_j|z_k) \)

Bag of Words

"Writing Style"
Extract Features
Quantize
Count Visual Words

<table>
<thead>
<tr>
<th>Word Index</th>
<th>Word</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>![Visual Word 1]</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>![Visual Word 2]</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>![Visual Word 3]</td>
<td>12</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>![Visual Word L]</td>
<td>1</td>
</tr>
</tbody>
</table>

© 2011 The MITRE Corporation. All rights reserved
probabilistic Latent Semantic Analysis (pLSA)

\[ w = VQ \text{ index} \]
\[ d = \text{image} \]
\[ z = \text{object} \]

Hierarchical Bayesian relation

\[ P(w_i | d_j) = \sum_{k=1}^{K} P(z_k | d_j) P(w_i | z_k) \]

Object distributions

Observed VQ index distributions per image

VQ index distributions per object

Object distributions per image

EM likelihood function

\[ L = \sum_d \sum_w n(d,w) \log P(d,w), \quad P(w | d) \approx n(d,w) \]

\[ n(d,w) = \text{raw histogram} \]

Adapted from W. Freeman, MIT (2005)
Querying

object overlap: probability that chosen objects in first and second images are similar: “cosine similarity”

\[ \text{sim}(d_i, d_m) \approx \sum_{k=1}^{K} P(z_k | d_i) P(z | d_m) \]

\[ + \alpha \sum_{j=1}^{L} \left( \sum_{k=1}^{K} P(z_k | d_i, w_j) P(z_k | d_m, w_j) \right) \frac{n_{ij} n_{mj}}{\sum_{j} n_{ij} \sum_{j} n_{mj}} \]

VQ index sense overlap: do both VQ indexes refer to the same object?

VQ index overlap: do both images contain common indexes?

\[ \text{retrieved class } d^* \equiv \arg \max_{j} \{ \text{sim}(d_i, d_j) \} \]
pLSA Object "Discovery"

Ranked by \( \max P(d|z_k) = \arg \max P(\text{Image} | \text{Object}_k) \)

Object 1  Style = "tight"

Object 2  Style = "loose"

Writer indexes shown in green
Classification by Spatial Pyramid Matching

Local feature correspondences

Optimal *partial* match: match all of smaller set to some of larger set

*Depicted are Feature Locations*

**X** = \{\vec{x}_1, \ldots, \vec{x}_m\} \quad **Y** = \{\vec{y}_1, \ldots, \vec{y}_n\}

\[ m < n \]

\[ \min_{\pi:X \rightarrow Y} \sum_{x_i \in X} ||x_i - \pi(x_i)|| \]

Slide credit: B. Leibe & K. Grauman, U. Texas, Ausin
Pyramid match Method

- Optimal matching too expensive
- Approximate optimal partial match

Optimal match: $O(r^3)$
Greedy match: $O(r^2 \log r)$

*Pyramid match: $O(r)$*

$r = \text{number of images}$
Pyramid match kernel: Histogram intersection

Feature histograms:

Level 3

Level 2

Level 1

Level 0

Total weight (value of pyramid match kernel): \( \mathcal{I}_3 + \frac{1}{2}(\mathcal{I}_2 - \mathcal{I}_3) + \frac{1}{4}(\mathcal{I}_1 - \mathcal{I}_2) + \frac{1}{8}(\mathcal{I}_0 - \mathcal{I}_1) \)

\[
\mathcal{I}(H(X), H(Y)) = \sum_j \min(H(X)_j, H(Y)_j) \quad \text{Take minimum of each bin}
\]

Slide adapted from Lazebnik, et al., U. North Carolina
Spatial Pyramid Matching: Processing Flow

Training Image Database

Training phase

SIFT

Train

VQ

VQ Codebook

Testing Image Database

Testing phase

SIFT

VQ

VQ indexes

Spatial Pyramid

Pyramid Histograms
## Experimental Results

### DUTCH

<table>
<thead>
<tr>
<th>Approach</th>
<th>Validation Paradigm</th>
<th>Features</th>
<th>Codebook Size</th>
<th>Rank-1 %Error</th>
<th>Time (hrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>pLSA</td>
<td>10-fold cross</td>
<td>Hessian-Laplace</td>
<td>512</td>
<td>11.8</td>
<td>15.4</td>
</tr>
<tr>
<td>Pyramid</td>
<td>SLOO</td>
<td>Hessian-Affine</td>
<td>512</td>
<td>10.4</td>
<td>4.01</td>
</tr>
</tbody>
</table>

### Arabic

<table>
<thead>
<tr>
<th>Approach</th>
<th>Validation Paradigm</th>
<th>Features</th>
<th>Codebook size</th>
<th>Rank-1 %Error</th>
<th>Time (hrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>pLSA</td>
<td>SLOO</td>
<td>Hessian-Affine</td>
<td>512</td>
<td>1.30</td>
<td>5.1</td>
</tr>
<tr>
<td>Pyramid</td>
<td>SLOO</td>
<td>Harris-Hessian-Laplace</td>
<td>2048</td>
<td>1.96</td>
<td>1.3</td>
</tr>
<tr>
<td>Pyramid</td>
<td>SLOO</td>
<td>Harris-Hessian-Laplace</td>
<td>2048</td>
<td>1.96</td>
<td>.16</td>
</tr>
<tr>
<td>Read Codebook</td>
<td>SLOO</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Strict Leave One Out (SLOO)**

---

250 Subjects
2 docs per subject
500 total documents
Style: printed, uppercase
Gray Scale
Text Independent

51 Subjects
3 docs per subject
153 total documents
Style: cursive
Gray Scale
Text Dependent

© 2011 The MITRE Corporation. All rights reserved
Summary

- First known application of generative local computer vision models to forensics or biometrics

- Beats best published results on Dutch

- Learning is totally unsupervised

- Works across different languages and text

Publications:


Backups
References


Local features

- Distinctive
- Robust to changes

Aggregate local statistics:

- Direct match
- Indirect match

Indirect Match

Direct Match

First Image

Second Image

Local image patches corresponding To feature locations
Lowe detector: Scale-Space

Low-Pass, Gaussian filtered versions of image  Bandpass filtered versions of image

Borrowed from Dave Lowe, U. of British Columbia, 2004
Scale & Orientation Assignment

- Use Gaussian & DoG representations

DoG image

Detected keypoint

Borrowed from Yung-Yu Chuang, National Taiwan U., 2006
Scale & Orientation assignment

Gradient Magnitude:
\[ m(x, y) = (L(x + 1, y) - L(x - 1, y))^2 + (L(x, y + 1) - L(x, y - 1))^2 \]

Gradient Orientation:
\[ \theta(x, y) = \tan^{-1}((L(x, y + 1) - L(x, y - 1))/(L(x + 1, y) - L(x - 1, y))) \]

Arrows
Direction: represents orientation
Length: represents magnitude

Gaussian filtered image, \( L \), at closest scale: This is the scale

Gradients superimposed over Gaussian filtered image

Adapted from Yung-Yu Chuang, National Taiwan U., 2006
Orientation assignment

1. Form bins containing sum of weighted gradient magnitudes corresponding to orientations at that range of angles.

2. Find peak

3. Determine angle corresponding to bin
   This is the orientation

4. Rotate gradient magnitudes to dominant orientation after orientation is found.

Adapted from Yung-Yu Chuang, National Taiwan U., 2006.
SIFT Descriptor

- **4 x 4 Gradient window.** Compute for 16 windows
- **Histogram each window in 8 directions**
- **16 x 8 = 128 dimensional feature vector**
- Represents stable, local image structure:
  - Strength of Gaussian filtered image gradient
  - Quantized to 8 directions, or 45 degree slices

---

**Parameters determined experimentally**

**16 gradient magnitudes quantized to 8 directions**

---

Adapted from slide by Jonas Hurrelmann
### Some SIFT-Family detectors

<table>
<thead>
<tr>
<th>SIFT Detector</th>
<th>Relative Strengths</th>
<th>Relative Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hessian-DOG (Lowe)</td>
<td>Invariant to rotation &amp; scale. Good for blobs.</td>
<td>Corners, not fully affine invariant. Occlusions &amp; clutter may be an issue.</td>
</tr>
<tr>
<td>Hessian-Laplace</td>
<td>Very similar to Hessian-DOG but slightly higher location accuracy</td>
<td>Corners, not fully affine invariant. Occlusions &amp; clutter may be an issue.</td>
</tr>
<tr>
<td>Maximally Stable</td>
<td>Affine invariant. Good for viewpoint change. Good with homogeneous regions with distinct boundaries. Good all-around.</td>
<td>Blur. Heterogeneous or complex-shaped regions.</td>
</tr>
<tr>
<td>Harris-Hessian-Laplace</td>
<td>Corners. Good for occlusions &amp; clutter.</td>
<td></td>
</tr>
</tbody>
</table>