Ubiquitous Fine-Grained Computer Vision

Shu Kong

Department of Computer Science, UC Irvine
Outline

1. Problem definition
2. Instantiation
3. Challenge and philosophy
4. Fine-grained classification with holistic representation
5. Fine-grained identification by matching local patches
6. Future work and conclusion
1. Problem definition
2. Instantiation
3. Challenge and philosophy
4. Fine-grained classification with holistic representation
5. Fine-grained identification by matching local patches
6. Future work and conclusion
Fine-grained

- marginally different or subtle
Fine-grained

• marginally different or *subtle*

• involving great attention to *detail* (Oxford dictionary)
Problem Definition

Fine-grained

- marginally different or **subtle**
- involving great attention to **detail** (Oxford dictionary)
Problem Definition

Fine-grained

• marginally different or subtle

• involving great attention to detail (Oxford dictionary)

• The devil is in the details!

• ...and everywhere!
Problem definition

Fine-grained computer vision

- distinguish subordinate categories within an entry-level category
Problem definition

Fine-grained computer vision

- distinguish subordinate categories within an entry-level category

- tasks are like classification, segmentation, specific case studies, etc.
previously, generic classification -- car vs. bird
now, fine-grained car model classification
now, fine-grained bird species classification
previously, in phytology, identifying by eye
now, automatically, accurately identifying species-level pollen and matching fossilized pollen grains with modern reference

modern pollen grain from glauca

fossil pollen grain from glauca
previously, in biology, semantic segmentation
e.g. binary label for biological data of *C. elegans*
now, instance segmentation
enabling study of worm population

now, instance segmentation enabling study of worm population
previously, modeling image aesthetics study as binary classification, low- vs. high- aesthetic
previously, modeling image aesthetics study as binary classification, low- vs. high- aesthetic
now, fine-grained ranking for personal photo album management
now, fine-grained ranking for personal photo album management
1. Problem definition
2. Instantiation
3. Challenge and philosophy
4. Fine-grained classification with holistic representation
5. Fine-grained identification by matching local patches
6. Future work and conclusion
Challenge and philosophy

• lack of training data
  – costly data collection and annotation
Challenge and philosophy

• lack of training data
  – costly data collection and annotation

• large numbers of categories
Challenge and philosophy

• lack of training data
  – costly data collection and annotation

• **large numbers of categories**
  – >14,000 birds
  – >278,000 butterfly&moth
  – >941,000 insects
Challenge and philosophy

- lack of training data
  - costly data collection and annotation
- large numbers of categories
- high intra-class vs. low inter-class variance
Challenge and philosophy

- lack of training data
  - costly data collection and annotation
- large numbers of categories
- high intra-class vs. low inter-class variance

Caspian Tern  Caspian Tern  Elegant Tern
Challenge and philosophy

- lack of training data
  - costly data collection and annotation
- large numbers of categories
- high intra-class vs. low inter-class variance

philosophy

- finding discriminative parts, and matching them effectively
Holistic representation based method

1. Problem definition
2. Instantiation
3. Challenge and philosophy
4. Fine-grained classification with holistic representation
5. Fine-grained identification by matching local patches
6. Future work
7. Conclusion
Holistic representation based method

recognizing bird species by seeing the photo

Red_Winged_Blackbird

Brandt_Cormorant

Acadian_Flycatcher

Yellow_Headed_Blackbird

Pelagic_Cormorant

Yellow_Billed_Cuckoo
recognizing bird species by seeing the photo

In literature, detecting keypoint/parts and stacking them as holistic representation

[Images of bird species]

- Red_Winged_Blackbird
- Brandt_Cormorant
- Acadian_Flycatcher
- Yellow_Headed_Blackbird
- Pelagic_Cormorant
- Yellow_Billed_Cuckoo

picture from Wah et al, 2011
But, this requires strong-supervised annotation, which is expensive to obtain.

picture from Wah et al, 2011
But, this requires strong-supervised annotation, which is expensive to obtain.

Preferably in weakly supervised manner --

- solely based on category labels
- without any part annotation/masks.
Holistic representation based method

One method for this is called bilinear pooling

Lin et al., Bilinear CNN models for fine-grained visual recognition, ICCV, 2015
Holistic representation based method

One method for this is called bilinear pooling. To compute second-order statistics of local features, and average them as a single holistic representation.

Lin et al., Bilinear CNN models for fine-grained visual recognition, ICCV, 2015
Holistic representation based method

One method for this is called bilinear pooling compute second-order statistics of local features, and average them as a single holistic representation.

The local features can be activations at hidden layers of a convolutional neural network (CNN).

Lin et al., Bilinear CNN models for fine-grained visual recognition, ICCV, 2015
Holistic representation based method

Bilinear Pooling

\[ x \in \mathbb{R}^{h \times w \times c} \]
Holistic representation based method

Bilinear Pooling

\[ \mathbf{X} \in \mathbb{R}^{h \times w \times c} \]

\[ \mathbf{x}_i \in \mathbb{R}^{c} \quad i \in [1, hw] \]

Lin et al., Bilinear CNN models for fine-grained visual recognition, ICCV, 2015
Holistic representation based method

Bilinear Pooling

\[ x \in \mathbb{R}^{h \times w \times c} \]

\[ x_i \in \mathbb{R}^c \quad i \in [1, hw] \]

\[ X \in \mathbb{R}^{c \times hw} \]

Lin et al., Bilinear CNN models for fine-grained visual recognition, ICCV, 2015
Holistic representation based method

Bilinear Pooling

\[ \mathbf{X} \in \mathbb{R}^{h \times w \times c} \]

\[ \mathbf{x}_i \in \mathbb{R}^c \quad i \in [1, hw] \]

\[ \mathbf{X} \in \mathbb{R}^{c \times hw} \]

\[ \mathbf{XX}^T = \sum_{i=1}^{hw} \mathbf{x}_i \mathbf{x}_i^T \]

Lin et al., Bilinear CNN models for fine-grained visual recognition, ICCV, 2015
Holistic representation based method

Bilinear Pooling

\[ \mathbf{X} \in \mathbb{R}^{h \times w \times c} \]

\[ \mathbf{x}_i \in \mathbb{R}^c \quad i \in [1, hw] \]

\[ \mathbf{X} \in \mathbb{R}^{c \times hw} \]

\[ \mathbf{XX}^T = \sum_{i=1}^{hw} \mathbf{x}_i \mathbf{x}_i^T \]

\[ \mathbf{z} = \text{vec}(\mathbf{XX}^T) \in \mathbb{R}^{c^2} \]

Lin et al., Bilinear CNN models for fine-grained visual recognition, ICCV, 2015
Holistic representation based method

Bilinear Pooling CNN -- training in an end-to-end manner

Lin et al., Bilinear CNN models for fine-grained visual recognition, ICCV, 2015
Holistic representation based method

Low-rank Bilinear Pooling

\[ z = vec(XX^T) \in \mathbb{R}^{c^2} \]
Holistic representation based method

Low-rank Bilinear Pooling

\[ z = \text{vec}(XX^T) \in \mathbb{R}^{c^2} \]

linear SVM

\[ \max(0, 1 - y_i w^T z_i + b) \]
Holistic representation based method

Low-rank Bilinear Pooling

\[ z = \text{vec}(XX^T) \in \mathbb{R}^{c^2} \]

linear SVM

\[ \max(0, 1 - y_i w^T z_i + b) \]

\[ w^T \text{vec}(XX^T) \iff \text{tr}(W^T XX^T) \]
Holistic representation based method

Low-rank Bilinear Pooling

\[ z = \text{vec}(XX^T) \in \mathbb{R}^{c^2} \]

linear SVM

\[ \max(0, 1 - y_i w^T z_i + b) \]

\[ w^T \text{vec}(XX^T) \iff \text{tr}(W^T XX^T) \iff \text{tr}(UU^T XX^T) \]
Holistic representation based method

Low-rank Bilinear Pooling

\[ z = \text{vec}(XX^T) \in \mathbb{R}^{c^2} \]

linear SVM

\[
\max(0, 1 - y_i w^T z_i + b)
\]

\[ w^T \text{vec}(XX^T) \iff tr(W^T XX^T) \iff tr(UU^T XX^T) \]

linear SVM in matrix

\[
\max(0, 1 - y_i \text{tr}(W^T X X^T) + b)
\]
Holistic representation based method

Low-rank Bilinear Pooling

\[ z = \text{vec}(XX^T) \in \mathbb{R}^{c^2} \]

linear SVM

\[ \max(0, 1 - y_i w^T z_i + b) \]

\[ w^T \text{vec}(XX^T) \iff \text{tr}(W^T XX^T) \iff \text{tr}(UU^T XX^T) \]

linear SVM in matrix

\[ \max(0, 1 - y_i \text{tr}(W^T X X^T) + b) \]

rank-r SVM

\[ \max(0, 1 - y_i \text{tr}(W_{r}^T X X^T) + b) \]
Holistic representation based method

Low-rank SVM

![Graph showing eigen values and index of sorted eigen values. The graph displays two different patterns; one with a smooth decline and another with a step-like increase.](image-url)
Holistic representation based method

Low-rank SVM
Holistic representation based method

When bilinear SVM meets bilinear feature

1. linear SVM
   \[ \max(0, 1 - y_i w^T z_i + b) \]

2. linear SVM in matrix
   \[ \max(0, 1 - y_i \text{tr}(W^T X_i X_i^T) + b) \]
Holistic representation based method

When bilinear SVM meets bilinear feature

1. linear SVM
   \[ \max(0, 1 - y_i w^T z_i + b) \]

2. linear SVM in matrix
   \[ \max(0, 1 - y_i \text{tr}(W^T X_i X_i^T) + b) \]

**Theorem 1** Let \( w^* \in \mathbb{R}^{c^2} \) be the optimal solution of the linear SVM in Equation 1 over bilinear features, then \( W^* = \text{mat}(w^*) \in \mathbb{R}^{c \times c} \) is the optimal solution in Equation 2. Moreover, \( W^* = W^{*T} \).
Holistic representation based method

When bilinear SVM meets bilinear feature

1. linear SVM
   \[ \max(0, 1 - y_i w^T z_i + b) \]

2. linear SVM in matrix
   \[ \max(0, 1 - y_i \text{tr}(W^T X_i X_i^T) + b) \]

Theorem 1 Let \( w^* \in \mathbb{R}^{c^2} \) be the optimal solution of the linear SVM in Equation 1 over bilinear features, then \( W^* = \text{mat}(w^*) \in \mathbb{R}^{c \times c} \) is the optimal solution in Equation 2. Moreover, \( W^* = W^*^T \).

\[
\begin{align*}
  w^* &= \sum_{y_i = 1} \alpha_i z_i - \sum_{y_i = -1} \alpha_i z_i \\
  W^* &= \sum_{y_i = 1} \alpha_i X_i X_i^T - \sum_{y_i = -1} \alpha_i X_i X_i^T \\
\end{align*}
\]

where \( \alpha_i \geq 0, \forall i = 1, \ldots, N \)
Holistic representation based method

When bilinear SVM meets bilinear feature

1. linear SVM

\[ \max(0, 1 - y_i w^T z_i + b) \]

2. linear SVM in matrix

\[ \max(0, 1 - y_i \text{tr}(W^T X_i X_i^T) + b) \]

\[
W^* = \Psi \Sigma \Psi^T = \Psi_+ \Sigma_+ \Psi_+^T + \Psi_- \Sigma_- \Psi_-^T \\
= \Psi_+ \Sigma_+ \Psi_+^T - \Psi_- |\Sigma_-| \Psi_-^T \\
= U_+ U_+^T - U_- U_-^T
\]

\[
w^* = \sum_{y_i=1} \alpha_i z_i - \sum_{y_i=-1} \alpha_i z_i
\]

\[
W^* = \sum_{y_i=1} \alpha_i X_i X_i^T - \sum_{y_i=-1} \alpha_i X_i X_i^T
\]

where \( \alpha_i \geq 0, \forall i = 1, \ldots, N \)
When bilinear SVM meets bilinear feature

1. linear SVM
\[ \max(0, 1 - y_i w^T z_i + b) \]

2. linear SVM in matrix
\[ \max(0, 1 - y_i \text{tr}(W^T X_i X_i^T) + b) \]

\[ w^T \text{vec}(X X^T) \leftrightarrow \text{tr}(W^T X X^T) \leftrightarrow \text{tr}(U U^T X X^T) \]

\[ \|U^T X\|_F^2 \leftrightarrow \text{tr}(U^T X X^T U) \]

\[ w^* = \sum_{y_i=1} \alpha_i z_i - \sum_{y_i=-1} \alpha_i z_i \]

\[ W^* = \sum_{y_i=1} \alpha_i X_i X_i^T - \sum_{y_i=-1} \alpha_i X_i X_i^T \]

where \( \alpha_i \geq 0, \forall i = 1, \ldots, N \)
Holistic representation based method

When bilinear SVM meets bilinear feature

1. linear SVM
   \[
   \max(0, 1 - y_i \mathbf{w}^T \mathbf{z}_i + b)
   \]

2. linear SVM in matrix
   \[
   \max(0, 1 - y_i \text{tr}(\mathbf{W}^T \mathbf{X}_i \mathbf{X}_i^T) + b)
   \]

\[
\mathbf{w}^T \text{vec}(\mathbf{X} \mathbf{X}^T) \iff \text{tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T) \iff \text{tr}(\mathbf{U} \mathbf{U}^T \mathbf{X} \mathbf{X}^T)
\]

\[
\|\mathbf{U}^T \mathbf{X}\|_F^2 \iff \text{tr}(\mathbf{U}^T \mathbf{X} \mathbf{X}^T \mathbf{U})
\]

\[
\max(0, 1 - y_i \left\{ \|\mathbf{U}_+^T \mathbf{X}_i\|_F^2 - \|\mathbf{U}_-^T \mathbf{X}_i\|_F^2 \right\} + b)
\]

\[
\max(0, 1 - y_i \left\{ \text{tr}(\mathbf{U}_+^T \mathbf{X}_i \mathbf{X}_i^T) - \text{tr}(\mathbf{U}_-^T \mathbf{X}_i \mathbf{X}_i^T) \right\} + b)
\]
Holistic representation based method

When bilinear SVM meets bilinear feature
maximum Frobenius norm

(d) our model (LRBP-I)

max(0, 1 − yi \{||U_+X_i||_F^2 − ||U_-X_i||_F^2\} + b)
max(0, 1 − yi \{\text{tr}(U_+U_+^TX_iX_i^T) − \text{tr}(U_-U_-^TX_iX_i^T)\} + b)
When bilinear SVM meets bilinear feature

maximum Frobenius norm

no need to compute bilinear features when testing

\[
\max(0, 1 - y_i \left\{ \| U^+_T X_i \|^2_F - \| U^-_T X_i \|^2_F \right\} + b)
\]

\[
\max(0, 1 - y_i \left\{ \text{tr}(U^+_U^T X_i X_i^T) - \text{tr}(U^- U^-_T X_i X_i^T) \right\} + b)
\]
Holistic representation based method

When bilinear SVM meets bilinear feature

maximum Frobenius norm

no need to compute bilinear features when testing

200 classes, then param size is reduced from $200 \times 512 \times 512$ to $200 \times 512 \times 8$

$$\max(0, 1 - y_i \{ \| U_+^T X_i \|_F^2 - \| U_-^T X_i \|_F^2 \} + b)$$

$$\max(0, 1 - y_i \{ \text{tr}(U_+ U_+^T X_i X_i^T) - \text{tr}(U_- U_-^T X_i X_i^T) \} + b)$$
explicitly computing bilinear features
more efficient useful when \(hw>m\)

our model (LRBP-II)
Holistic representation based method

classifier co-decomposition -- learning a common factor and class-specific parameters of smaller size

$$\min_{U_k, P} \sum_{k=1}^{K} \| U_k - PV_k \|_F^2$$

$$U_k = [U_{+k}, U_{-k}] \in \mathbb{R}^{c \times r}$$

$$P \in \mathbb{R}^{c \times m}$$

$$V_k \in \mathbb{R}^{m \times r}$$

$$m < c$$
Holistic representation based method

classifier co-decomposition -- learning a common factor and class-specific parameters of smaller size

\[
\min_{\mathbf{V}_k, \mathbf{P}} \sum_{k=1}^{K} \| \mathbf{U}_k - \mathbf{P} \mathbf{V}_k \|_F^2
\]

\[
\mathbf{U}_k = [\mathbf{U}_{+k}, \mathbf{U}_{-k}] \in \mathbb{R}^{c \times r}
\]

\[
\mathbf{P} \in \mathbb{R}^{c \times m}
\]

\[
\mathbf{V}_k \in \mathbb{R}^{m \times r}
\]

\[
m < c
\]

**Theorem 2** The optimal solution of \( \mathbf{P} \) to Equation 11 spans the subspace of the singular vectors corresponding to the largest \( m \) singular values of \([\mathbf{U}_1, \ldots, \mathbf{U}_K]\).
Holistic representation based method

classifier co-decomposition -- learning a common factor and class-specific parameters of smaller size

$$\min_{V_k, P} \sum_{k=1}^{K} \| U_k - PV_k \|_F^2$$

\[ U_k = [U_{+k}, U_{-k}] \in \mathbb{R}^{c \times r} \]

\[ P \in \mathbb{R}^{c \times m} \]

\[ V_k \in \mathbb{R}^{m \times r} \]

\[ m < c \]
Studying the two hyperparameters

- low dimension $m$
- low rank $r$
Studying the two hyperparameters -- \( m \) and \( r \)
Studying the two hyperparameters -- $m$ and $r$
Holistic representation based method

Studying the two hyperparameters -- $m$ and $r$
Studying the two hyperparameters -- $m$ and $r$

Figure 5: Classification accuracy on CUB-200 dataset [31] vs. reduced dimension ($m$) and rank ($r$).

Figure 6: Reconstruction fidelity of classifier parameters measured by peak signal-to-noise ratio versus reduced dimension ($m$) and rank ($r$).

Figure 7: The learned parameter size versus reduced dimension ($m$) and rank ($r$).
Studying the two hyperparameters -- $m$ and $r$

**Figure 5:** Classification accuracy on CUB-200 dataset [31] vs. reduced dimension ($m$) and rank ($r$).

**Figure 6:** Reconstruction fidelity of classifier parameters measured by peak signal-to-noise ratio versus reduced dimension ($m$) and rank ($r$).

**Figure 7:** The learned parameter size versus reduced dimension ($m$) and rank ($r$).

If 200 classes, then param size is reduced from $200 \times 512 \times 512$ to $(200 \times 8 \times 100 + 100 \times 512)$ ($\sim 5.2 \times 10^7$ single precision) to $(2.1 \times 10^5$ single precision)
### Holistic representation based method

#### Details on the complexity

<table>
<thead>
<tr>
<th></th>
<th>Full Bilinear</th>
<th>Random Maclaurin</th>
<th>Tensor Sketch</th>
<th>LRBP-I</th>
<th>LRBP-II</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Feature computation</strong></td>
<td>$O(hw c^2)$</td>
<td>$O(hw cd)$</td>
<td>$O(hw(c + d \log d))$</td>
<td>$O(hwmc)$</td>
<td>$O(hwmc + hw m^2)$</td>
</tr>
<tr>
<td><strong>Classification comp.</strong></td>
<td>$O(K c^2)$</td>
<td>$O(K d)$</td>
<td>$O(K d)$</td>
<td>$O(Krm hw)$</td>
<td>$O(Krm^2)$</td>
</tr>
<tr>
<td><strong>Feature Param</strong></td>
<td>$0$</td>
<td>$2 cd$ [40MB]</td>
<td>$2c$ [4KB]</td>
<td>$cm$ [200KB]</td>
<td>$cm$ [200KB]</td>
</tr>
<tr>
<td><strong>Total (K = 200)</strong></td>
<td>$K c^2$ [200MB]</td>
<td>$2 cd + K d$ [48MB]</td>
<td>$2c + K d$ [8MB]</td>
<td>$cm + K rm$ [0.8MB]</td>
<td>$cm + K rm$ [0.8MB]</td>
</tr>
</tbody>
</table>
Quantitative evaluation on benchmark datasets

Table 3: Summary statistics of datasets.

<table>
<thead>
<tr>
<th></th>
<th># train img.</th>
<th># test img.</th>
<th># class</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUB [31]</td>
<td>5994</td>
<td>5794</td>
<td>200</td>
</tr>
<tr>
<td>DTD [4]</td>
<td>1880</td>
<td>3760</td>
<td>47</td>
</tr>
<tr>
<td>Car [17]</td>
<td>8144</td>
<td>8041</td>
<td>196</td>
</tr>
<tr>
<td>Airplane [21]</td>
<td>6667</td>
<td>3333</td>
<td>100</td>
</tr>
</tbody>
</table>
## Holistic representation based method

### Quantitative evaluation on benchmark datasets

Table 3: Summary statistics of datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># train img.</th>
<th># test img.</th>
<th># class</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUB [31]</td>
<td>5994</td>
<td>5794</td>
<td>200</td>
</tr>
<tr>
<td>DTD [4]</td>
<td>1880</td>
<td>3760</td>
<td>47</td>
</tr>
<tr>
<td>Car [17]</td>
<td>8144</td>
<td>8041</td>
<td>196</td>
</tr>
<tr>
<td>Airplane [21]</td>
<td>6667</td>
<td>3333</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>FC-VGG16</th>
<th>Fisher</th>
<th>Full Bilinear</th>
<th>Random Maclaurin</th>
<th>Tensor Sketch</th>
<th>LRBP (Ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUB [31]</td>
<td>70.40</td>
<td>74.7</td>
<td>84.01</td>
<td>83.86</td>
<td>84.00</td>
<td><strong>84.21</strong></td>
</tr>
<tr>
<td>DTD [4]</td>
<td>59.89</td>
<td>65.53</td>
<td>64.96</td>
<td>65.57</td>
<td>64.51</td>
<td><strong>65.80</strong></td>
</tr>
<tr>
<td>Car [17]</td>
<td>76.80</td>
<td>85.70</td>
<td>91.18</td>
<td>89.54</td>
<td>90.19</td>
<td><strong>90.92</strong></td>
</tr>
<tr>
<td>Airplane [21]</td>
<td>74.10</td>
<td>77.60</td>
<td>87.09</td>
<td>87.10</td>
<td>87.18</td>
<td><strong>87.31</strong></td>
</tr>
</tbody>
</table>

| param. size (CUB) | 67MB | 50MB | 200MB | 48MB | 8MB | 0.8MB |
Holistic representation based method

Qualitative evaluation for understanding the model
Qualitative evaluation for understanding the model

- gradient map --- backpropogating error to input image
Qualitative evaluation for understanding the model

- gradient map --- backpropogating error to input image
- average activation map
Qualitative evaluation for understanding the model

- gradient map --- backpropagating error to input image
- average activation map
- simplying input image by removing superpixels
Holistic representation based method

Qualitative evaluation for understanding the model
Conclusion
Conclusion

1. a more compact and powerful model by coupling bilinear classifier and bilinear feature for fine-grained classification
Conclusion

1. a more compact and powerful model by coupling bilinear classifier and bilinear feature for fine-grained classification
2. a new direction for a weakly supervised visual learning
Conclusion

1. a more compact and powerful model by coupling bilinear classifier and bilinear feature for fine-grained classification
2. a new direction for a weakly supervised visual learning
3. useful for learning interpretable attentions
1. Problem definition
2. Instantiation
3. Challenge and philosophy
4. Fine-grained classification with holistic representation
5. Fine-grained identification by matching local patches
6. Future work and conclusion
patch-match based approach for pollen grain identification
Patch-match based method

patch-match based approach for pollen grain identification problem

Skilled experts trained for years have to identify by eye

image from Surangi W. Punyasena
Why do we care about identifying pollen?

- Pollen grains are ubiquitous and well preserved in the fossil record
Why do we care about identifying pollen?

- Pollen grains are ubiquitous and well preserved in the fossil record.
- Identification of pollen samples allows for analysis of plant biodiversity and evolution, understanding history of long-term climate change, etc.
Patch-match based method

A specific dataset for this exploration

1. arbitrary viewpoint of the pollen grains
A specific dataset for this exploration

1. arbitrary viewpoint of the pollen grains
2. Large intra-class and small inter-class variation
Why not holistic representation?
Why not holistic representation?

1. It is expensive to collect and annotate data.
Why not holistic representation?

1. It is expensive to collect and annotate data.

2. **There are not enough training data using holistic representation.**
Quantitative Result on Fossil Pollen

Why not holistic representation?

1. It is expensive to collect and annotate data.

2. There are not enough training data using holistic representation.

Table 1. Statistics of our fossil pollen grain dataset.

<table>
<thead>
<tr>
<th></th>
<th>#train</th>
<th>#test</th>
<th>#total</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>P. crithfieldii</em></td>
<td>65</td>
<td>43</td>
<td>108</td>
</tr>
<tr>
<td><em>P. glauca</em></td>
<td>65</td>
<td>355</td>
<td>420</td>
</tr>
<tr>
<td><em>P. mariana</em></td>
<td>65</td>
<td>287</td>
<td>352</td>
</tr>
<tr>
<td><strong>Summary</strong></td>
<td>195</td>
<td>685</td>
<td>880</td>
</tr>
</tbody>
</table>
Why not holistic representation?

1. It is expensive to collect and annotate data.
2. There are not enough training data using holistic representation.

Therefore, it's better to match local patches with geometric constraints.

Table 1. Statistics of our fossil pollen grain dataset.

<table>
<thead>
<tr>
<th></th>
<th>#train</th>
<th>#test</th>
<th>#total</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>P. critchfieldii</em></td>
<td>65</td>
<td>43</td>
<td>108</td>
</tr>
<tr>
<td><em>P. glauca</em></td>
<td>65</td>
<td>355</td>
<td>420</td>
</tr>
<tr>
<td><em>P. mariana</em></td>
<td>65</td>
<td>287</td>
<td>352</td>
</tr>
<tr>
<td>Summary</td>
<td>195</td>
<td>685</td>
<td>880</td>
</tr>
</tbody>
</table>

The patch-match method needs images to be aligned.
perform $k$-medoids clustering on an affinity graph of training set,
perform $k$-medoids clustering on an affinity graph of training set,
where pairwise similarity is based on Euclidean distance of pollen grain silhouette
perform $k$-medoids clustering on an affinity graph of training set,
where pairwise similarity is based on Euclidean distance of pollen grain silhouette
our patch-match based method

- patch exemplar selection
- patch match by sparse coding
- training stage
- SVM
- testing stage
discriminative patch selection
From a finite set of patches, $V$, we'd like to select $M$ patches, which should be/have
From a finite set of patches, $V$, we'd like to select $M$ patches, which should be/have

1. representative in feature space
From a finite set of patches, $V$, we'd like to select $M$ patches, which should be/have

1. representative in feature space
2. spatially distributed in input space
From a finite set of patches, $V$, we'd like to select $M$ patches, which should be/have

1. representative in feature space
2. spatially distributed in input space
3. discriminative
From a finite set of patches, $V$, we'd like to select $M$ patches, which should be/have

1. representative in feature space
2. spatially distributed in input space
3. discriminative
4. class balance
From a finite set of patches, $V$, we'd like to select $M$ patches, which should be/have

1. representative in feature space
2. spatially distributed in input space
3. discriminative
4. class balance
5. cluster compactness
From a finite set of patches, \( V \), we'd like to select \( M \) patches, which should be/have

1. representative in feature space
2. spatially distributed in input space
3. discriminative
4. class balance
5. cluster compactness

We index the selected patches by \( A \)
Maximizing the following set function is NP-hard.

$$F_R(A) = \sum_{j \in V} \max_{i \in A} S_{i,j}$$
Maximizing the following set function is NP-hard.

$$F_R(A) = \sum_{j \in V} \max_{i \in A} S_{ij}$$

A more general, well-known problem is the facility location problem, for example optimally placing sensors to monitor temperature.
Identification by patch-match sparse coding

1. Automatic patch exemplar selection (dictionary learning)
based on discriminative and generative criteria
Identification by patch-match sparse coding

1. Automatic patch exemplar selection (dictionary learning) based on discriminative and generative criteria

Automatically selected patches
Identification by patch-match sparse coding

1. Automatic patch exemplar selection (dictionary learning) based on discriminative and generative criteria
Identification by patch-match sparse coding

1. Automatic patch exemplar selection (dictionary learning)

2. Spatially-aware sparse coding (SACO)
   - penalize dictionary elements from distant spatial locations
spatially aware coding (SACO)

\[
\underset{a}{\text{argmin}} \left\| x - Da \right\|_2^2 + \lambda_1 \left\| \text{diag}(w)a \right\|_1
\]

Spatially aware dictionary

Exemplar patches (dictionary)

Test image

Test patch

Spatial weights
feedforward shrinkage function by transforming dictionary patches into convolutional filters

$$\arg\min_{a} \|x - Da\|^2_2 + \lambda_1 \|\text{diag}(w) a\|_1$$
feedforward shrinkage function by transforming dictionary patches into convolutional filters

\[
\arg\min_a \| x - Da \|_2^2 + \lambda_1 \| \text{diag}(w)a \|_1
\]

\[
\| x - Da \|_2^2 \quad \Rightarrow \quad \| \Omega x - a \|_2^2
\]
feedforward shrinkage function by transforming dictionary patches into convolutional filters

$$\arg\min_a \|x - Da\|_2^2 + \lambda_1 \|\text{diag}(w)a\|_1$$

$$\|x - Da\|_2^2 \rightarrow \|\Omega x - a\|_2^2$$

**SACO-I**

$$\Omega \equiv (D^T D)^{-1} D^T$$

$$u = \Omega x$$

$$a_i^* = \text{sgn}(u_i) \cdot \max(0, |u_i| - \lambda_1 w_i)$$

$$a^* = [a_1^*, \ldots, a_i^*, \ldots, a_m^*]^T$$
SACO -- Faster Matching

feedforward shrinkage function by transforming dictionary patches into convolutional filters

\[
a^* = \arg\min_a \| \mathbf{x} - D\mathbf{a}\|^2_2 + \lambda_2 \| \text{diag}(\mathbf{w})\mathbf{a}\|^2_2 + \lambda_1 \| \mathbf{a}\|_1
\]

\[
\| \mathbf{x} - D\mathbf{a}\|^2_2 \quad \rightarrow \quad \| \Omega \mathbf{x} - \mathbf{a}\|^2_2
\]

SACO-II

\[
\Omega \equiv (D^TD + \lambda_2 \text{diag}(\mathbf{w})^2)^{-1}D^T
\]

\[
\mathbf{u} = \Omega \mathbf{x}
\]

\[
a_i^* = \text{sgn}(u_i) \cdot \max(0, |u_i| - \lambda_1)
\]

\[
\mathbf{a}^* = [a_1^*, \ldots, a_i^*, \ldots, a_m^*]^T.
\]
Quantitative Result on Fossil Pollen

Represent patch using CNN feature extractor (VGG19)
Global average pooling of sparse codes by SACO
linear SVM

<table>
<thead>
<tr>
<th>SRC</th>
<th>VGG19+SVM</th>
<th>FV+SVM</th>
<th>SACO-I</th>
<th>SACO-II</th>
</tr>
</thead>
<tbody>
<tr>
<td>62.04</td>
<td>65.11</td>
<td>61.46</td>
<td>83.21</td>
<td>86.13</td>
</tr>
</tbody>
</table>

Table 1. Statistics of our fossil pollen grain dataset.

<table>
<thead>
<tr>
<th></th>
<th>#train</th>
<th>#test</th>
<th>#total</th>
</tr>
</thead>
<tbody>
<tr>
<td>P. critchfieldii</td>
<td>65</td>
<td>43</td>
<td>108</td>
</tr>
<tr>
<td>P. glauca</td>
<td>65</td>
<td>355</td>
<td>420</td>
</tr>
<tr>
<td>P. mariana</td>
<td>65</td>
<td>287</td>
<td>352</td>
</tr>
<tr>
<td>Summary</td>
<td>195</td>
<td>685</td>
<td>880</td>
</tr>
</tbody>
</table>

Substantially outperforms standard CNN and Fisher-vector based approaches!

We apply our approach to modern pollen grain identification.

<table>
<thead>
<tr>
<th>Our method</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P. Glauca</td>
</tr>
<tr>
<td>Predicted</td>
<td></td>
</tr>
<tr>
<td>P. Glauca</td>
<td>0.969</td>
</tr>
<tr>
<td>P. Mariana</td>
<td>0.021</td>
</tr>
</tbody>
</table>

Fossil pollen grains are degraded over time.

Using patches from modern pollen reference to identify fossilized ones.
Identifying Fossil Pollen with Modern Reference

- Use our method to select patches from modern pollen grains
- Use the selected modern patches to identify fossil ones
- We achieve 69% accuracy wrt expert labels.
Outline

1. Problem definition
2. Instantiation
3. Challenge and philosophy
4. Fine-grained classification with holistic representation
5. Fine-grained identification by matching local patches
6. Future work and conclusion
Content after this page is not suitable for people to watch!