Selecting Patches, Matching Species:

Shu Kong

CS, ICS, UCI

2016-4-6

Selecting Patches, Matching Species: Fossil Pollen Identification ....

Shu Kong

CS, ICS, UCI

2016-4-6

Selecting Patches, Matching Species: Fossil Pollen Identification by Spatially Aware Coding

Shu Kong

CS, ICS, UCI

2016-4-6

Outline

1. Background
2. Strong baselines
3. Our framework
4. Exemplar selecting for discriminative dictionary
5. Spatially aware coding for matching
6. Implementation details
7. Experimental study
8. Conclusion

Why pollen grains?

Pollen --

- is one of the most ubiquitous of terrestrial fossils

Why pollen grains?

Pollen --

- is one of the most ubiquitous of terrestrial fossils
- has an extraordinarily rich record

Why pollen grains?

Pollen --

- is one of the most ubiquitous of terrestrial fossils
- has an extraordinarily rich record
- has been used to test hypotheses and a diverse array of disciplines.

Why pollen grains?

such as…

- paleoecological and paleoclimatological investigation across hundreds to millions of years

Why pollen grains?

such as…

- paleoecological and paleoclimatological investigation across hundreds to millions of years
- implement the identification of plant speciation and extinction events

Why pollen grains?

such as...

- paleoecological and paleoclimatological investigation across hundreds to millions of years
- implement the identification of plant speciation and extinction events
- calculate the correlation and biostratigraphic dating of rock sequences

Why pollen grains?

such as...

- paleoecological and paleoclimatological investigation across hundreds to millions of years
- implement the identification of plant speciation and extinction events
- calculate the correlation and biostratigraphic dating of rock sequences
- conduct studies of long-term anthroppogenic impacts on plant communities and the study of plant-pollinator relationships

And…

recognizing pollen grains **at species level** is significant to the reconstruction of paleoenvironments and discrimination of paleoecologically and apleoclimatically significant taxa

And data?

high-throughput microscopic imaging allows for ready acquisition of large numbers of images of modern or fossilized pollen samples

But…

While high-throughput microscopic imaging allows for ready acquisition of large numbers of images of modern or fossilized pollen samples, do you want to identify and count by eye the number of grains of each species?

Left: A researcher (doctoral candidate) does a pollen count on a lake-sediment sample from South America. Acids are used to dissolve away the rock material and then the pollen is stained before it is viewed under the microscope. The number of pollen grains for each of several different species of plants is tallied.

But...

While high-throughput microscopic imaging allows for ready acquisition of large numbers of images of modern or fossilized pollen samples, do you want to identify and count by eye the number of grains of each species?

NO -- It is painstaking work, and requires substantial expertise and training.

ah...

While high-throughput microscopic imaging allows for ready acquisition of large numbers of images of modern or fossilized pollen samples, do you want to identify and count by eye the number of grains of each species?

NO -- It is painstaking work, and requires substantial expertise and training.

We don't want to do it by ourselves.

Then...

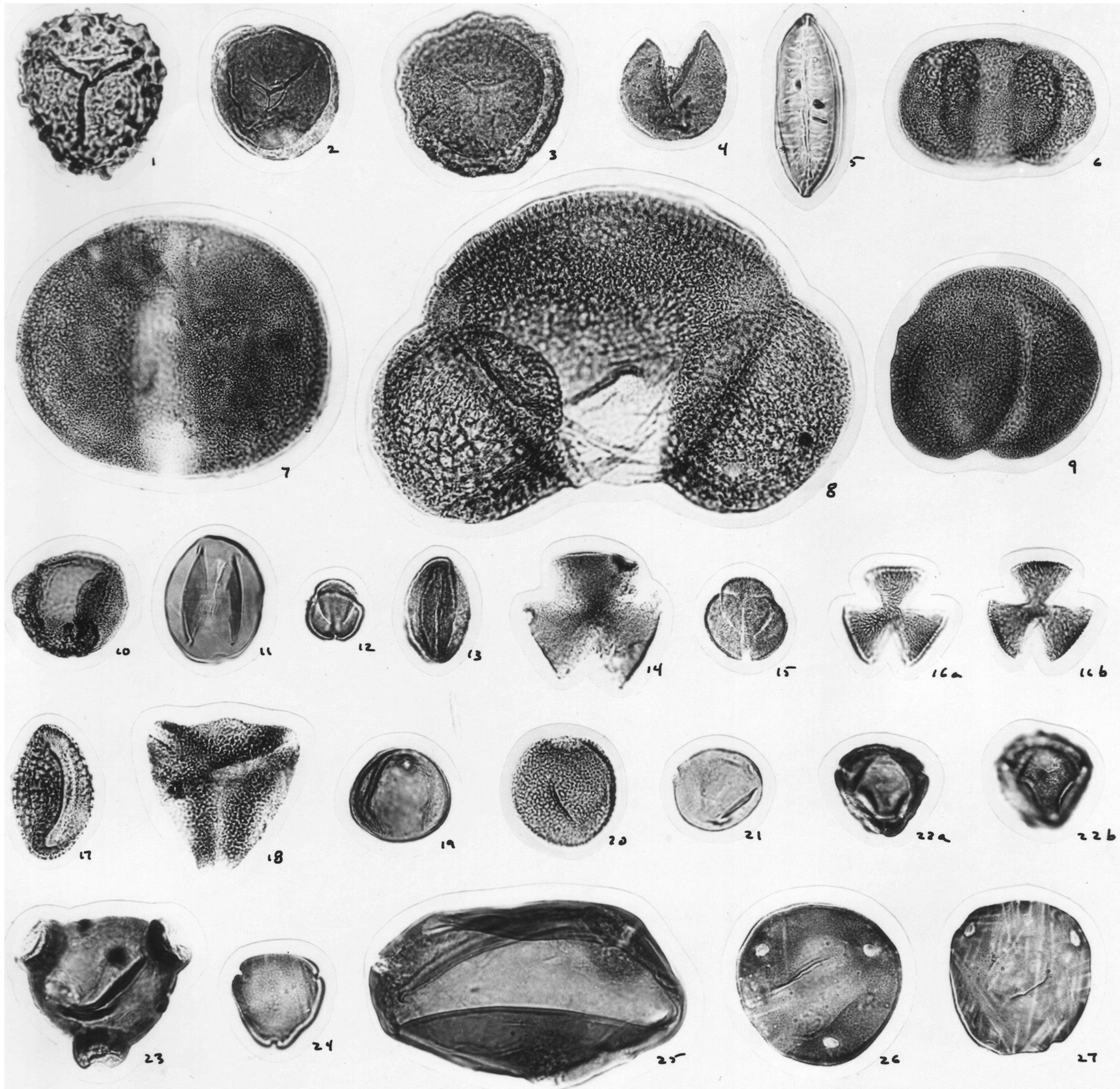
While high-throughput microscopic imaging allows for ready acquisition of large numbers of images of modern or fossilized pollen samples, <u>do you want to identify and count by eye the number of grains of each species?</u>

NO -- It is painstaking work, and requires substantial expertise and training.

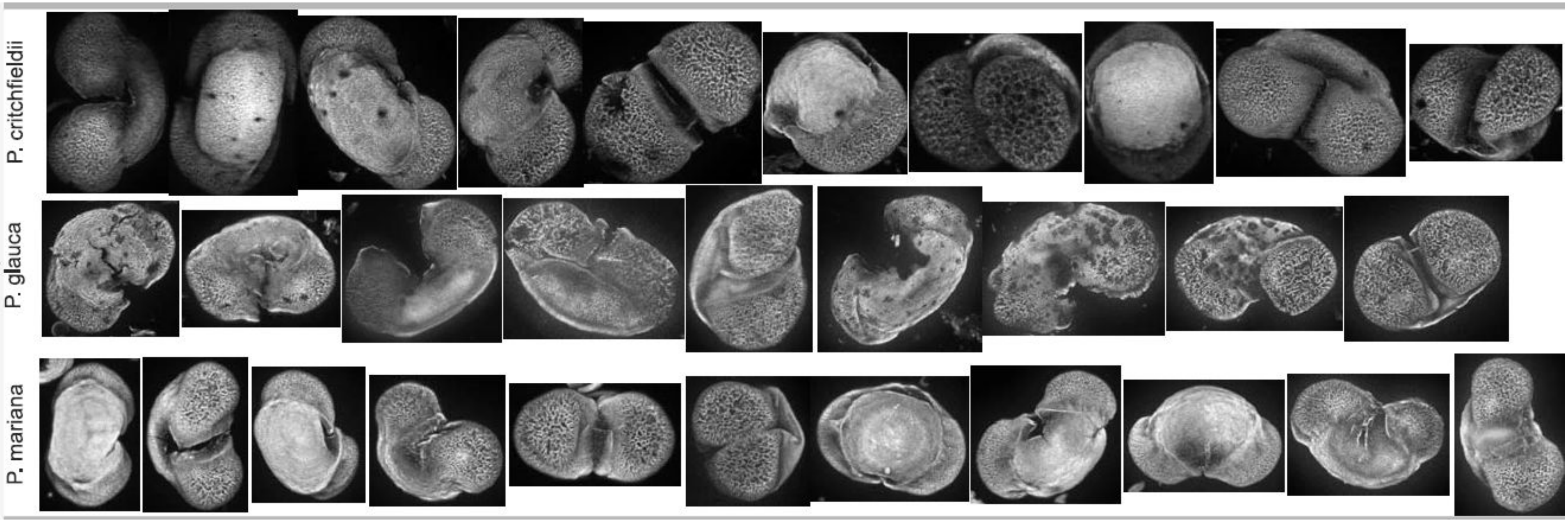
We don't want to do it by ourselves.

We would like to automate through machine learning, computer vision, etc.

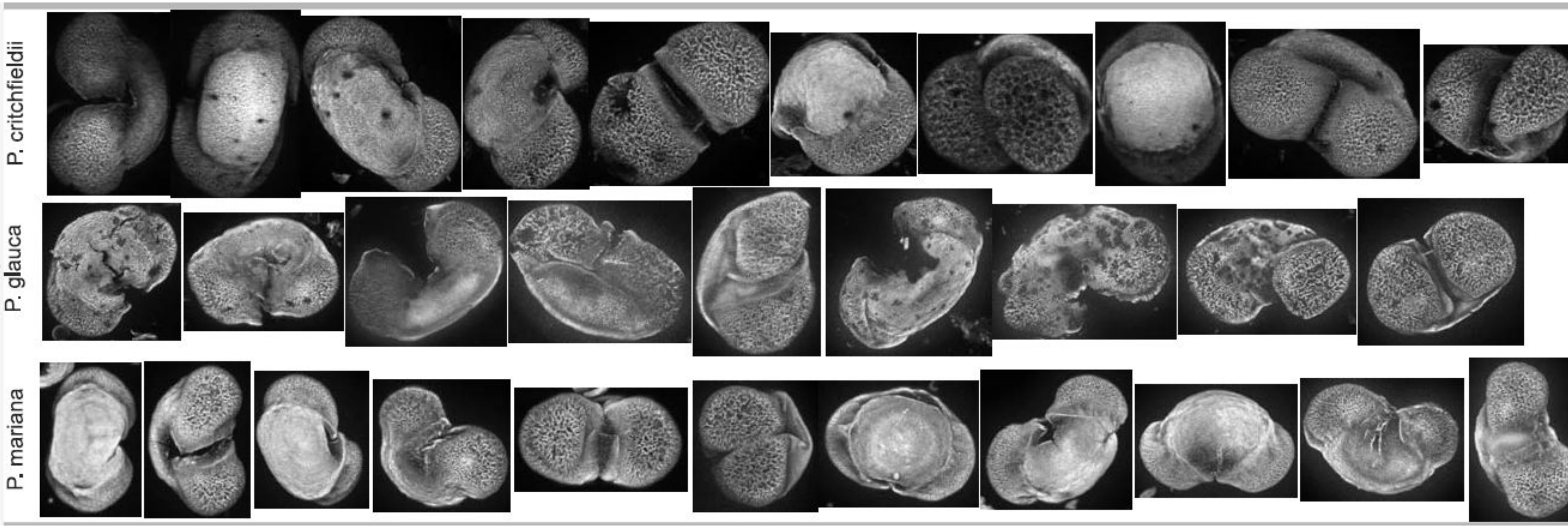
um.........................


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16a
16b
17
18
19
20
21
22a
22b
23
24
25
26
27

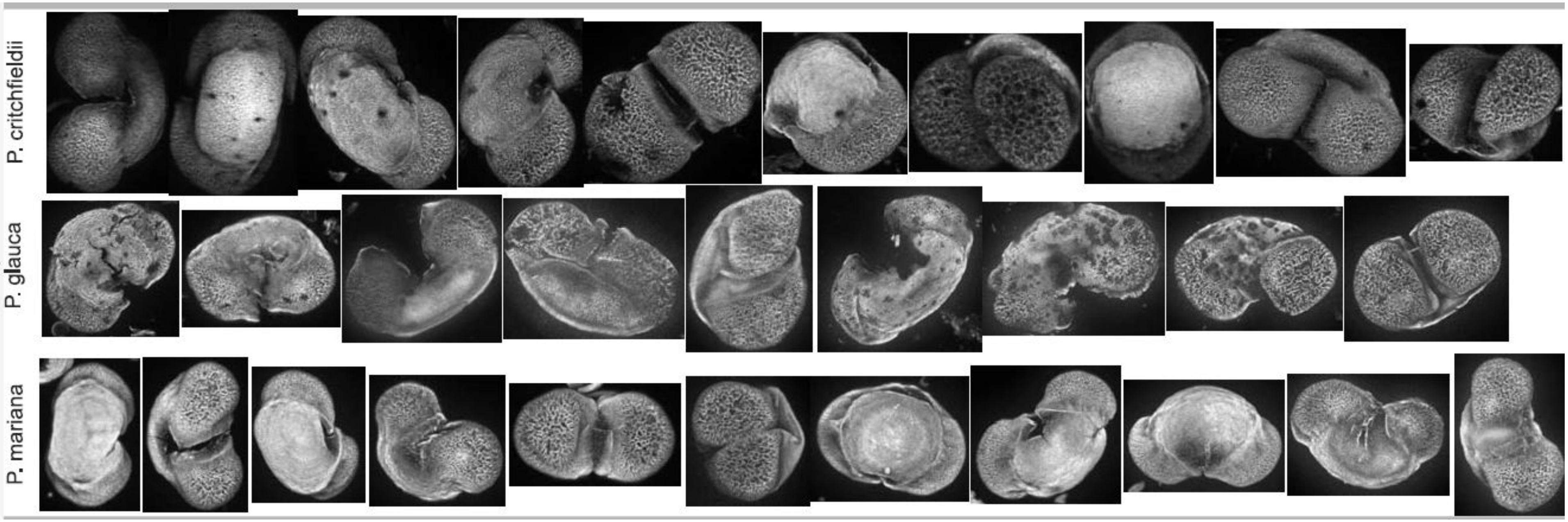
um.........................



um...it's nontrivial

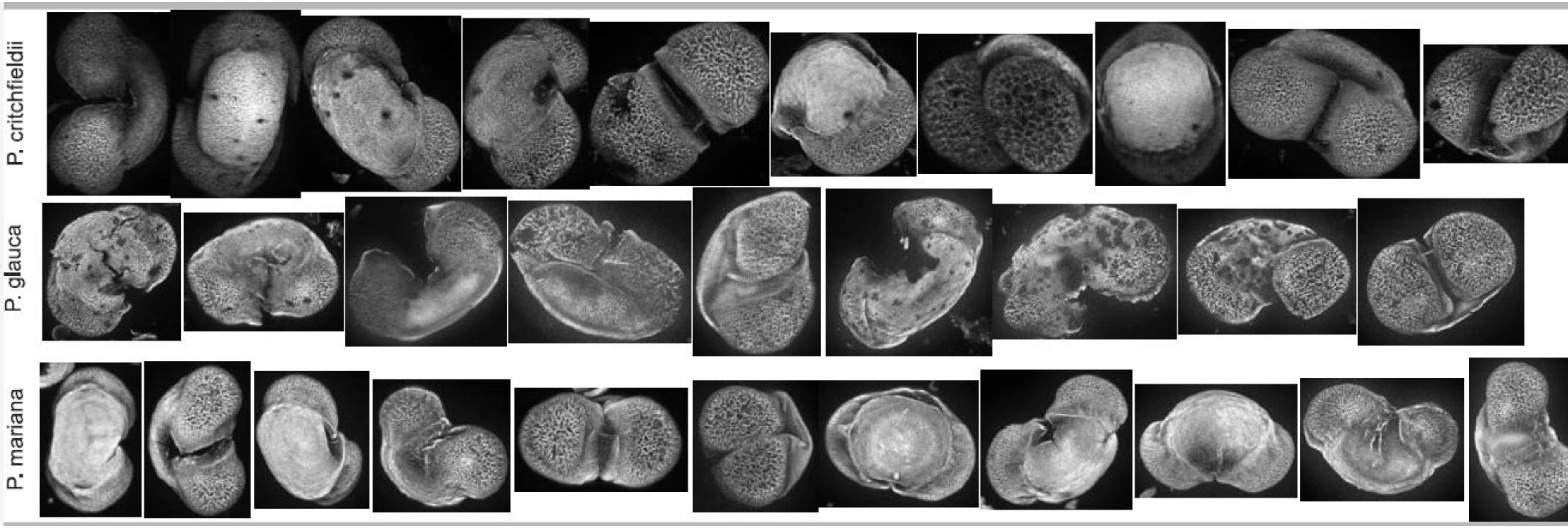


nontrivial task



1. arbitrary viewpoint of the pollen grains imaged

nontrivial task



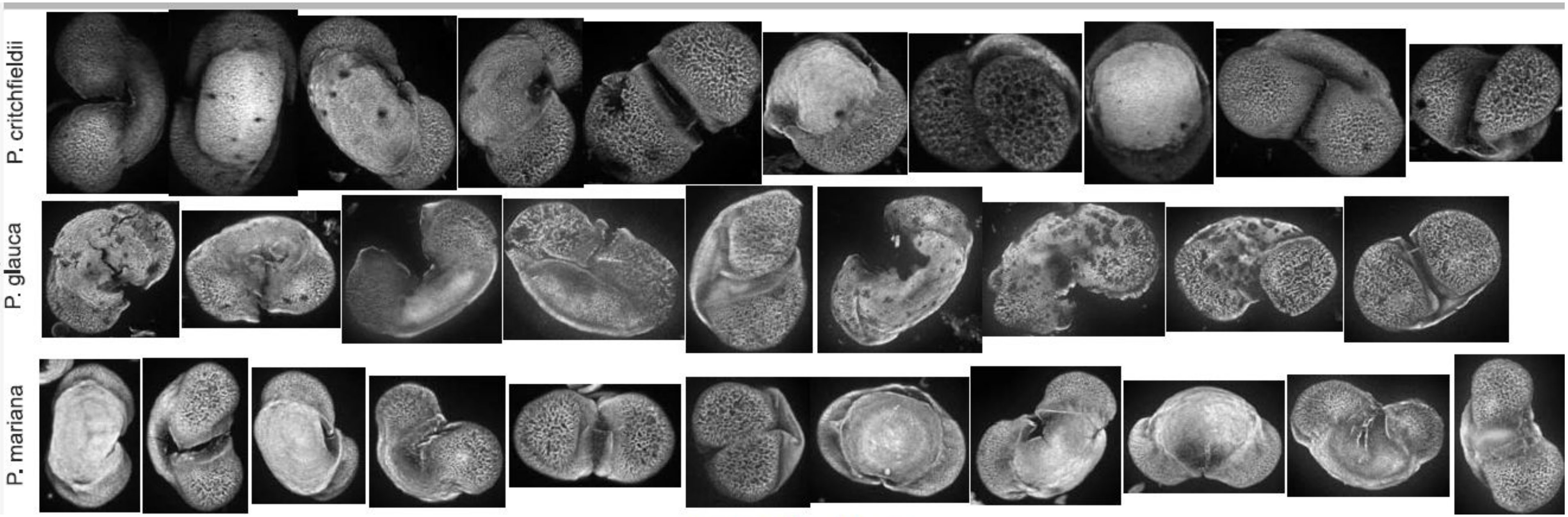
1. arbitrary viewpoint of the pollen grains imaged

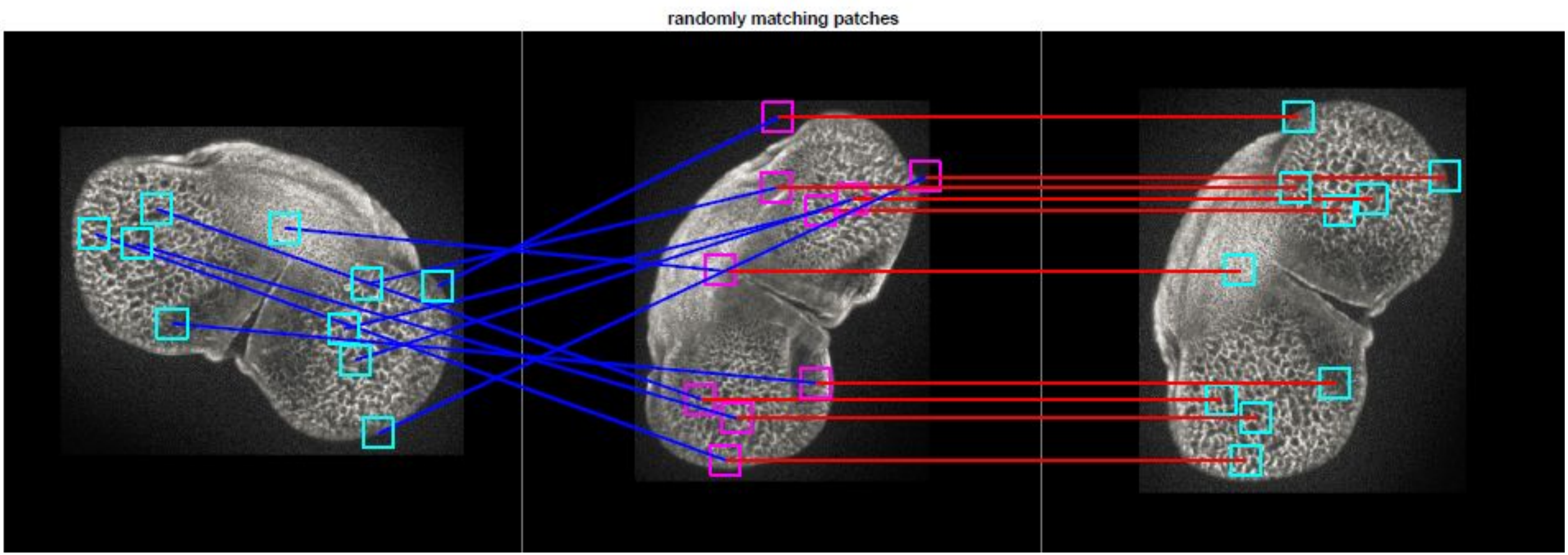
2. very limited amounts of expert-labeled training data

Table 1. Statistics of our fossil pollen grain dataset.

	#train	#test	#total
P. critchfieldii	65	43	108
P. glauca	65	355	420
P. mariana	65	287	352
Summary	195	685	880

well, for arbitrary viewpoint

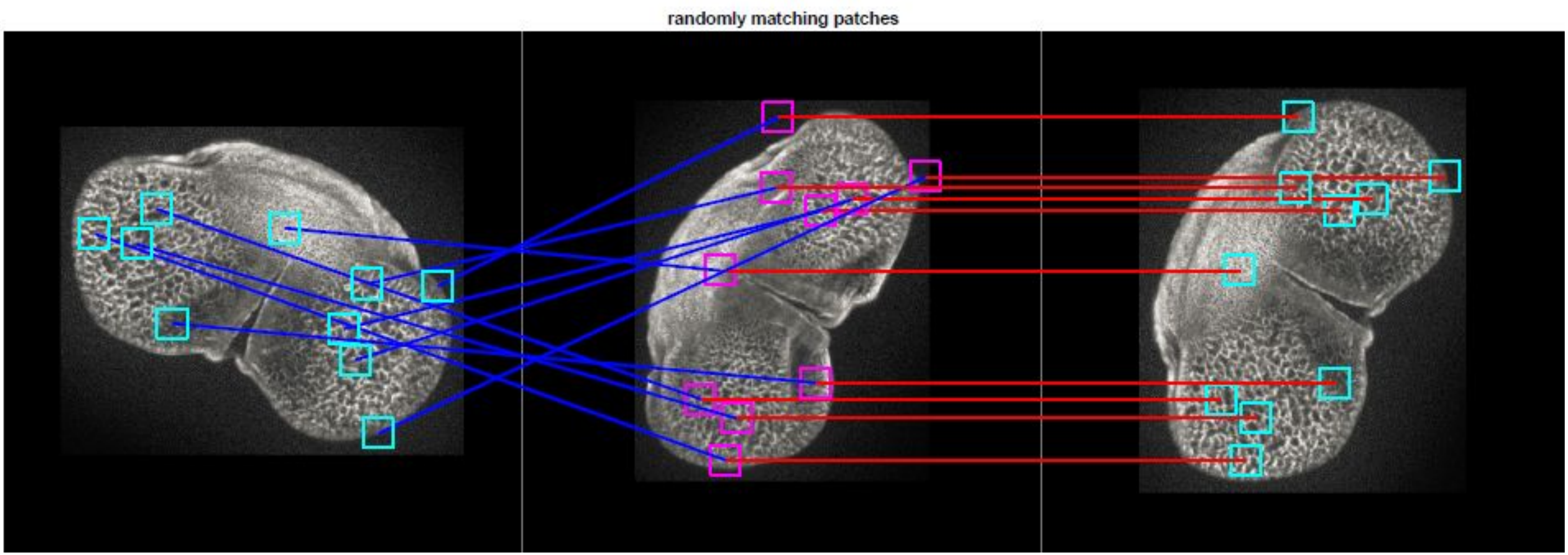




well, for arbitrary viewpoint

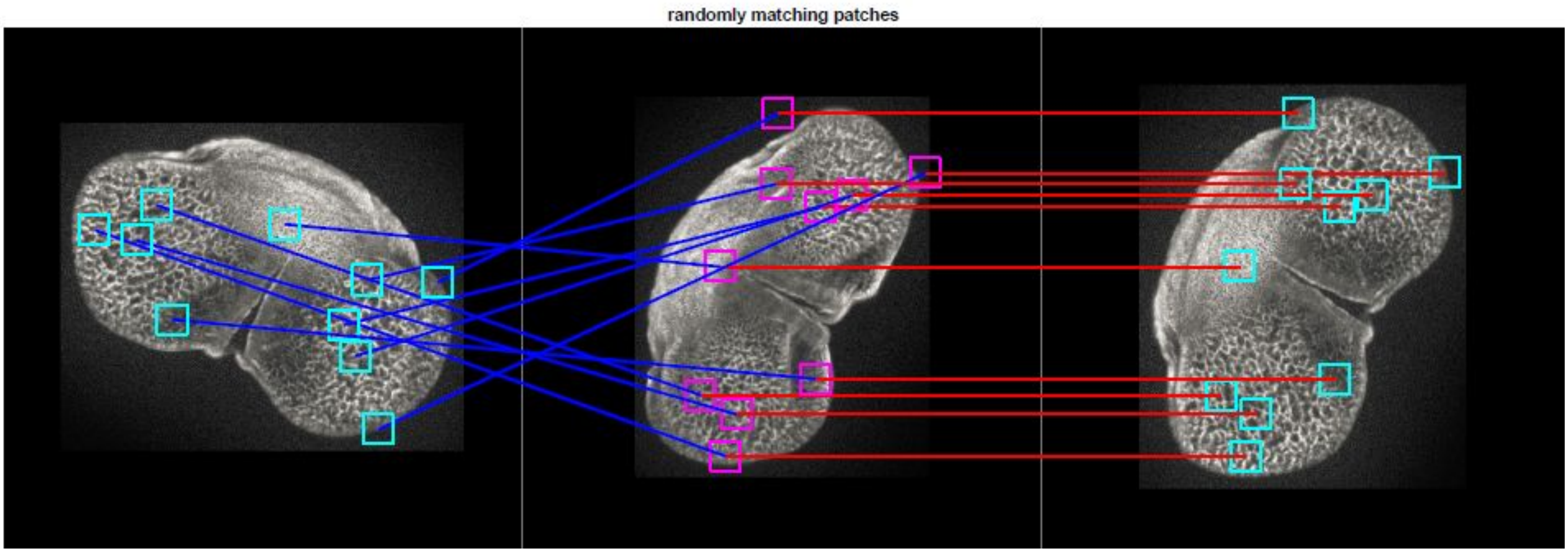
$$\min_\theta \|A - R_\theta(B)\|$$

where $R_\theta(B)$ is an operator that rotates image B by θ degrees.



hey... let's find some canonical viewpoints

practical to align images according to canonical viewpoints

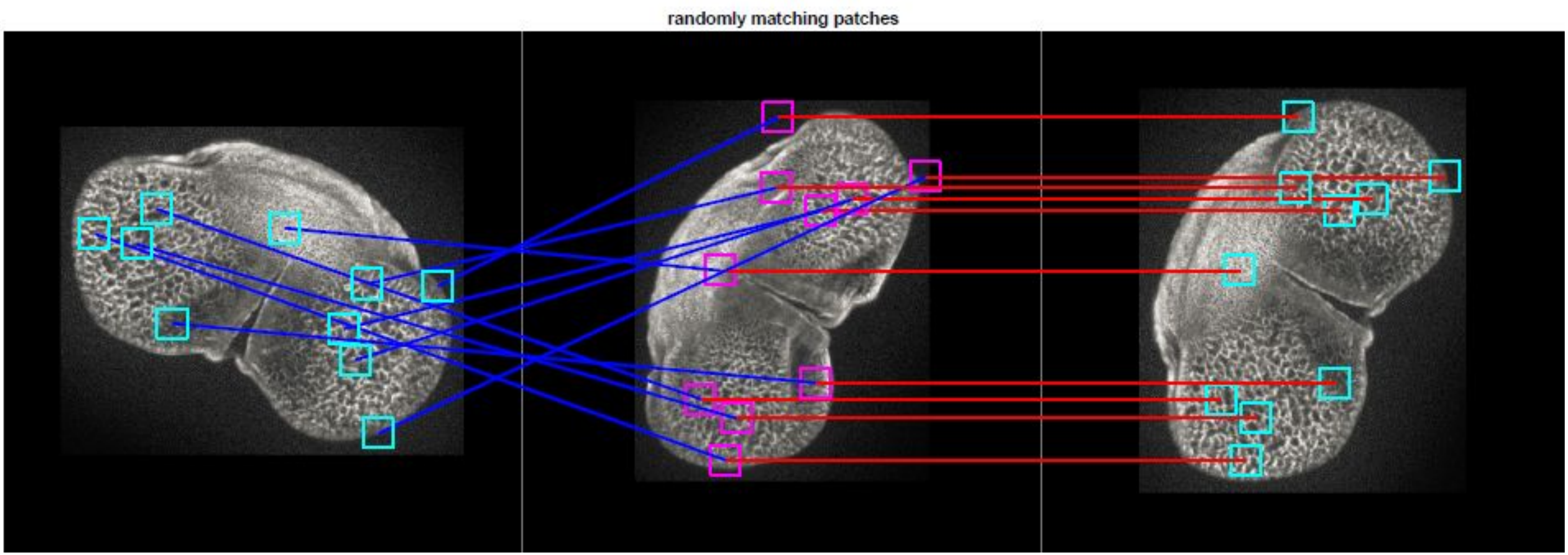


hey... let's find some canonical viewpoints

practical to align images according to canonical viewpoints

perform k-medoids clustering on a similarity graph of training set

$$similarity(A, B) = \frac{1}{\min_\theta \|A - R_\theta(B)\|}$$



Here it is

resize images to 40x40 pixel resolution

bin 360 degrees, and also flip images

k-medoids

Here it is

resize images to 40x40 pixel resolution

bin 360 degrees, and also flip images

k-medoids

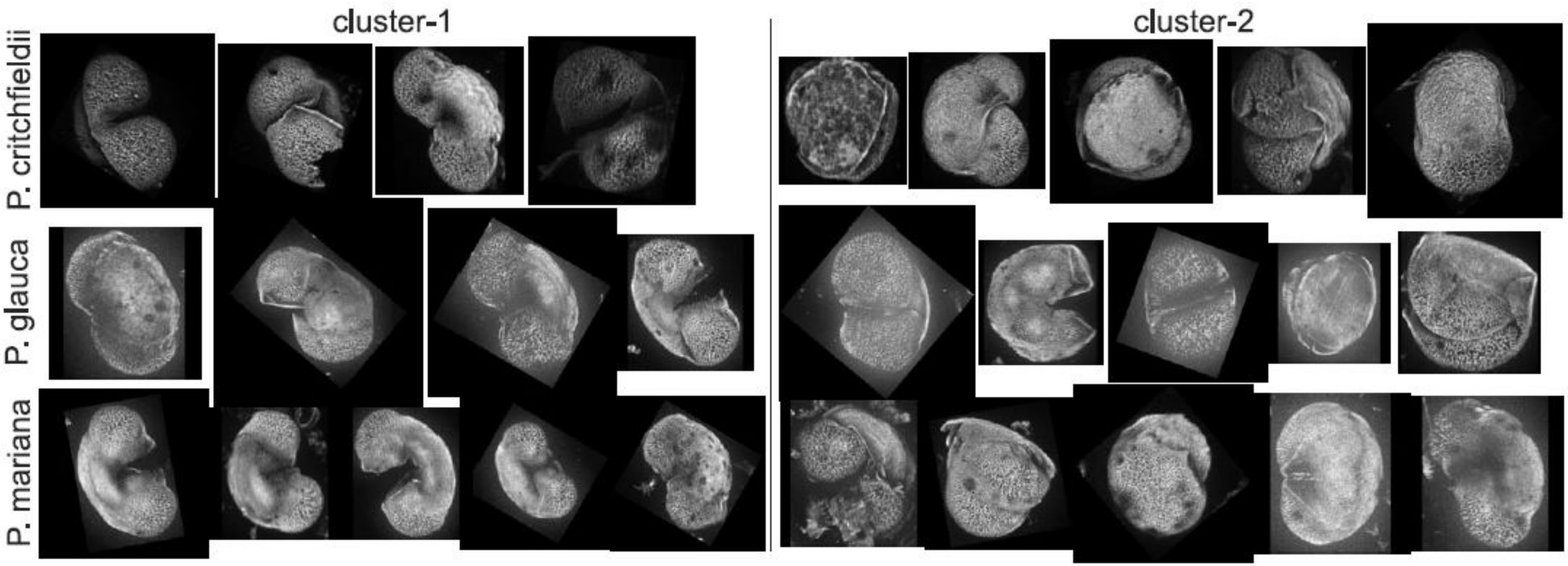


Figure 3. Rotated images according to two canonical viewpoints determined by k-medoids clustering.

Here it is

resize images to 40x40 pixel resolution

bin 360 degrees, and also flip images

k-medoids

once in-plate rotation is removed, better performance is observed

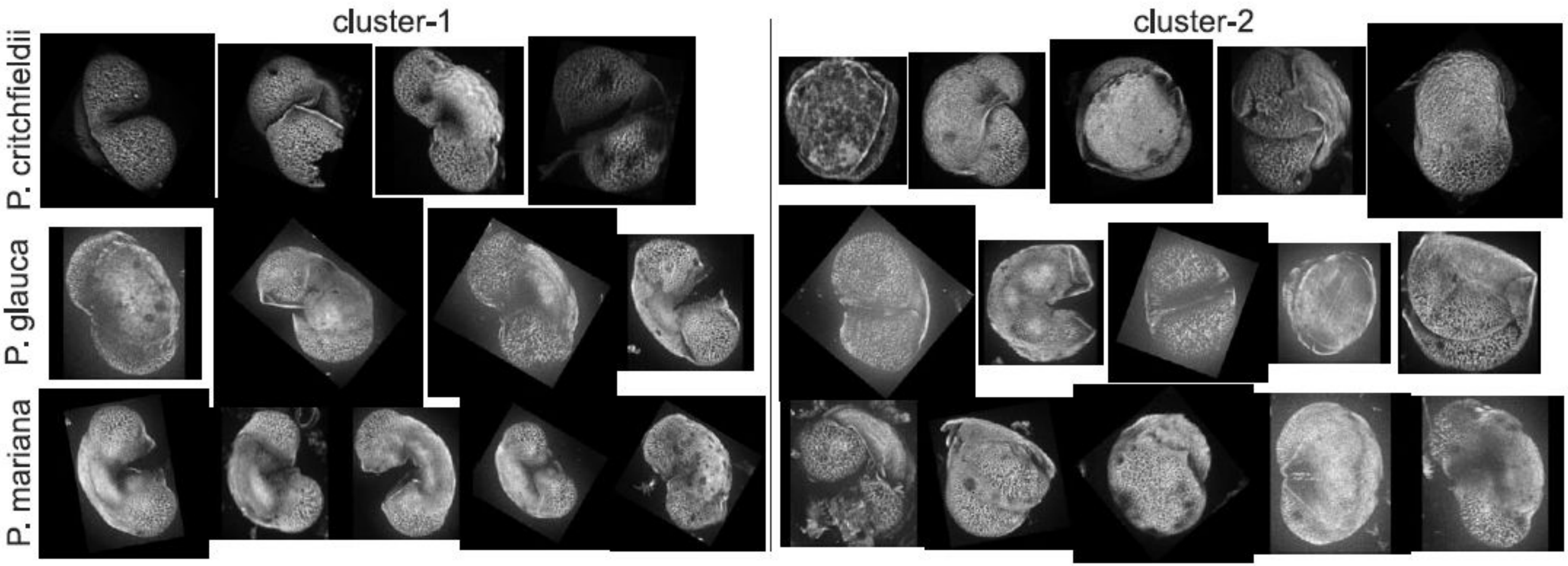


Figure 3. Rotated images according to two canonical viewpoints determined by k-medoids clustering.

Baseline methods

1. SRC
2. VGG19+FV+SVM
3. VGG19+SVM

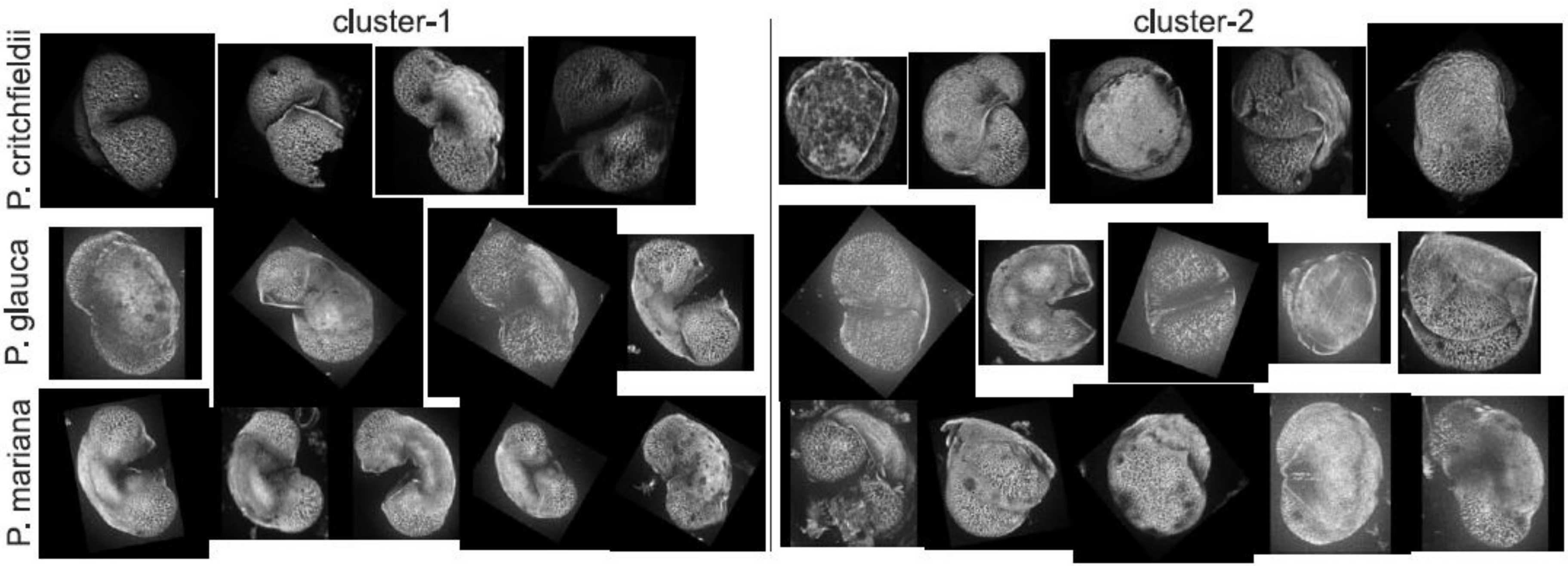


Figure 3. Rotated images according to two canonical viewpoints determined by k-medoids clustering.

Baseline methods

SRC	VGG19+SVM	FV+SVM
62.04	65.11	61.46

1. SRC
2. VGG19+FV+SVM
3. VGG19+SVM

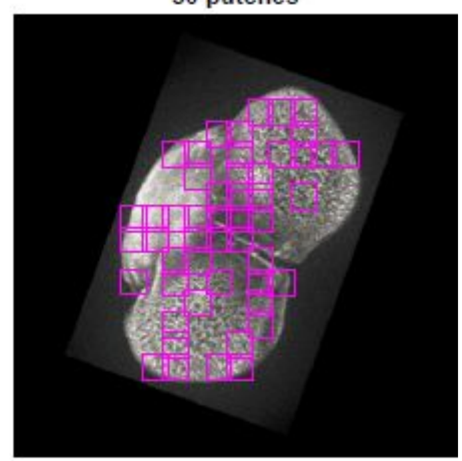
Baseline methods

SRC	VGG19+SVM	FV+SVM
62.04	65.11	61.46

1. SRC
2. VGG19+FV+SVM
3. VGG19+SVM

SRC uses patches from training set as dictionary. It sums the reconstruction error for testing patches, and also exploits the spatial information of the patches.

50 patches



happy with it?

SRC	VGG19+SVM	FV+SVM
62.04	65.11	61.46

1. SRC
2. VGG19+FV+SVM
3. VGG19+SVM

SRC uses patches from training set as dictionary. It sums the reconstruction error for testing patches, and also exploits the spatial information of the patches.

- random patches without spatial information: 57.12%
- selected patches with spatial information: 62.04%

happy with it?

SRC	VGG19+SVM	FV+SVM
62.04	65.11	61.46

1. SRC
2. VGG19+FV+SVM
3. VGG19+SVM

SRC uses patches from training set as dictionary. It sums the reconstruction error for testing patches, and also exploits the spatial information of the patches.

- random patches without spatial information: 57.12%
- selected patches with spatial information: 62.04%
- **besides, global pooling+SVM: 77.62%**

Our robust framework

1. Well-selected patches as dictionary perform better than random patches.

Our robust framework

1. Well-selected patches as dictionary perform better than random patches. --> **exemplar selection**

Our robust framework

1. Well-selected patches as dictionary perform better than random patches. --> **exemplar selection**

2. incorporating spatial information of the patches

Our robust framework

1. Well-selected patches as dictionary perform better than random patches. --> **exemplar selection**

2. incorporating spatial information of the patches -> **spatially aware coding**

Our robust framework

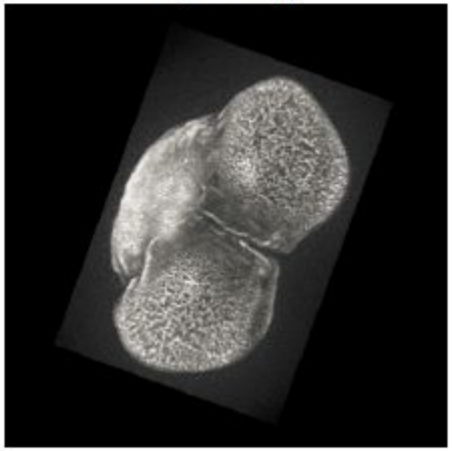
1. Well-selected patches as dictionary perform better than random patches. --> **exemplar selection**

2. incorporating spatial information of the patches -> **spatially aware coding**

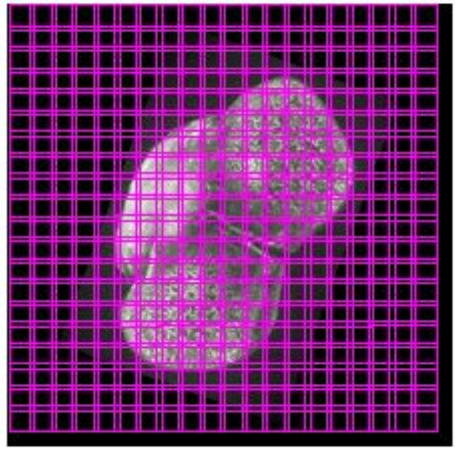
3. pooling+SVM is better than reconstruction-based scheme

Exemplar Selection

original image

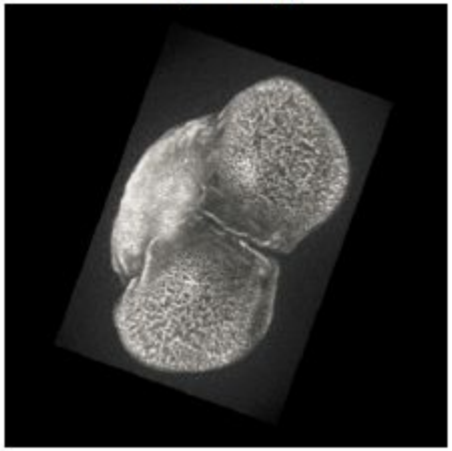


dense patches

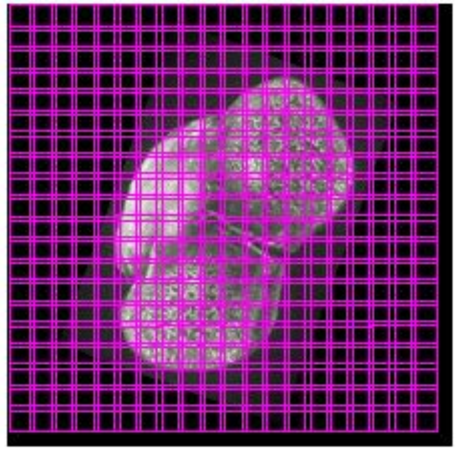


Exemplar Selection

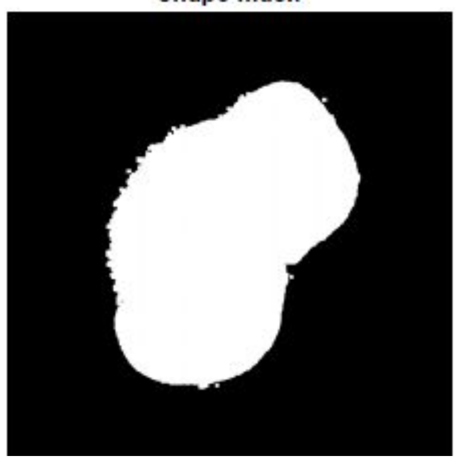
original image



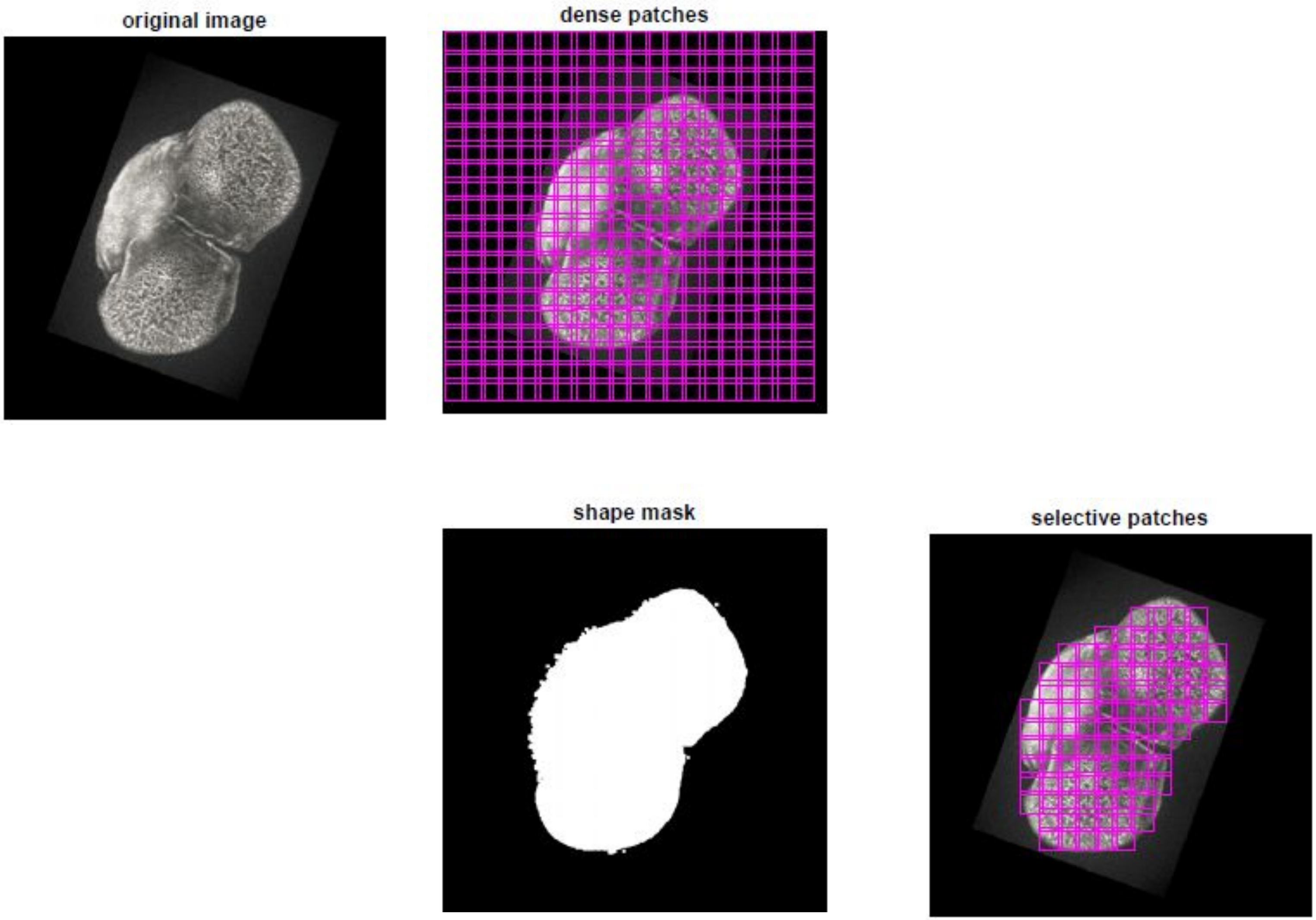
dense patches



shape mask

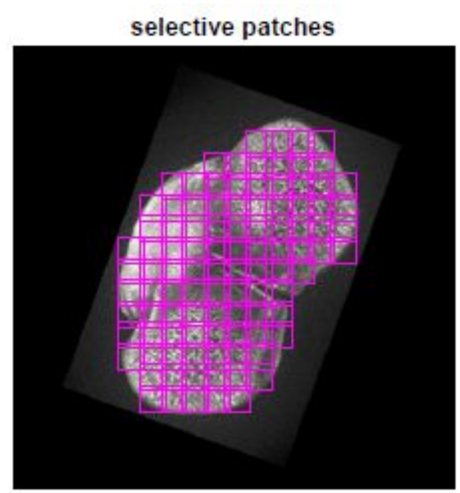


Exemplar Selection



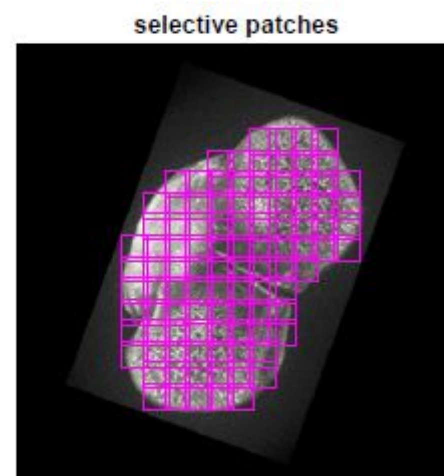
Exemplar Selection

From a finite set of patches, $\mathcal{V}$, we'd like to select M patches



Exemplar Selection

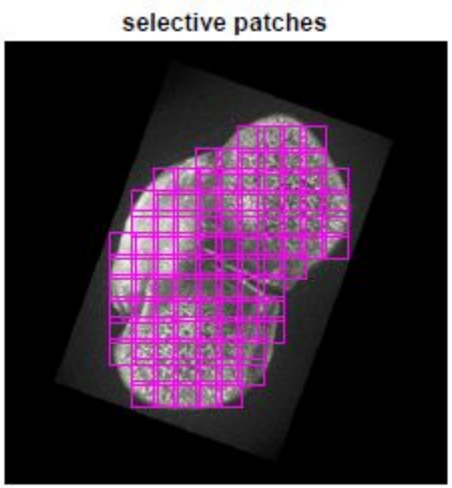
From a finite set of patches, $\mathcal{V}$, we'd like to select M patches, which should be/have



Exemplar Selection

From a finite set of patches, $\mathcal{V}$, we'd like to select M patches, which should be/have

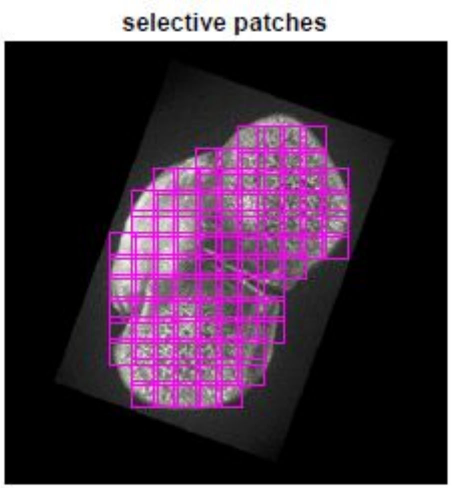
1. Representative in feature space



Exemplar Selection

From a finite set of patches, $\mathcal{V}$, we'd like to select M patches, which should be/have

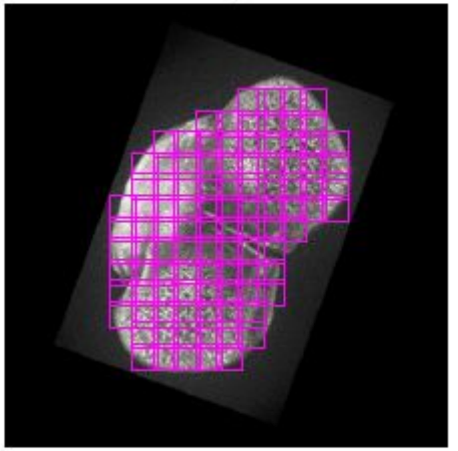
1. Representative in feature space
2. Spatially distributed in input space



Exemplar Selection

From a finite set of patches, $\mathcal{V}$, we'd like to select M patches, which should be/have

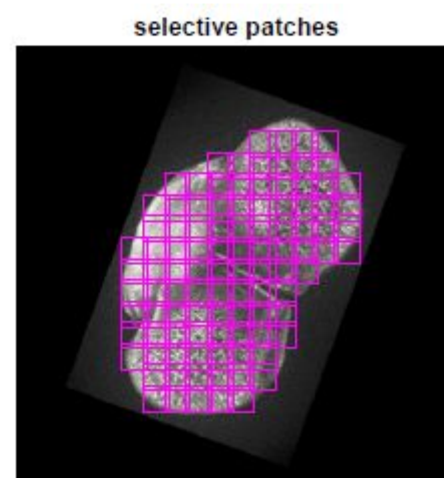
1. Representative in feature space
2. Spatially distributed in input space
3. Discriminative power



Exemplar Selection

From a finite set of patches, $\mathcal{V}$, we'd like to select M patches, which should be/have

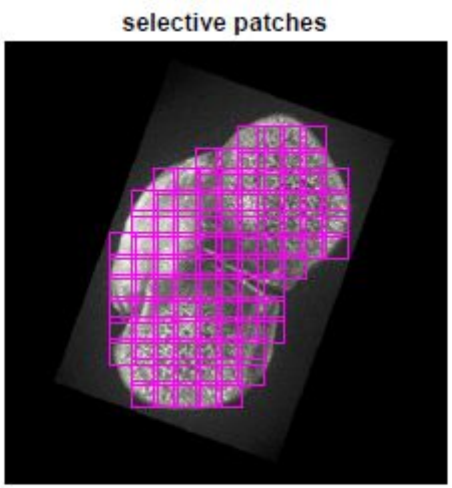
1. Representative in feature space
2. Spatially distributed in input space
3. Discriminative power
4. Class balance



Exemplar Selection

From a finite set of patches, $\mathcal{V}$, we'd like to select M patches, which should be/have

1. Representative in feature space
2. Spatially distributed in input space
3. Discriminative power
4. Class balance
5. Cluster compactness



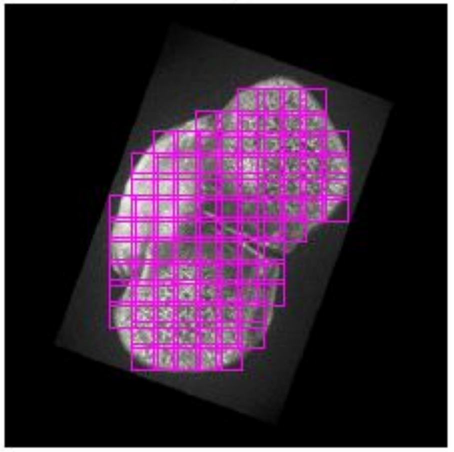
Exemplar Selection

From a finite set of patches, $\mathcal{V}$, we'd like to select M patches, which should be/have

1. Representative in feature space
2. Spatially distributed in input space
3. Discriminative power
4. Class balance
5. Cluster compactness

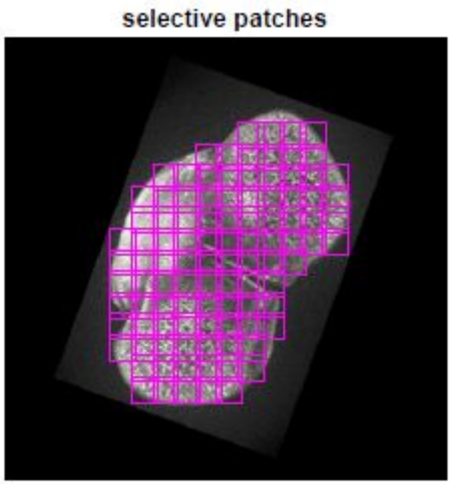
We index the selected patches by A

selective patches



Exemplar Selection -- Rule 1

Representative in feature space

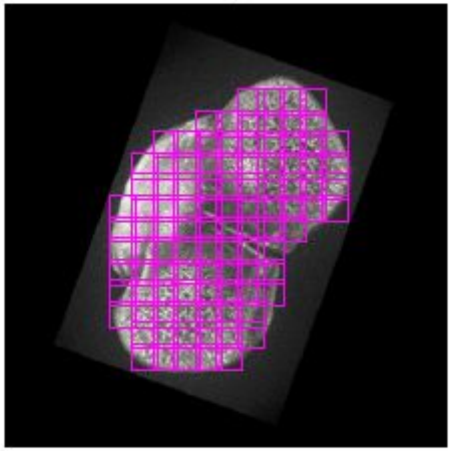


Exemplar Selection -- Rule 1

Representative in feature space

$\mathbf{S} \in \mathbb{R}^{M \times M}$ where $\mathbf{S}_{ij}$ is the similarity (a non-negative value) between patch i and patch j. Our aim is to select a subset $A \subseteq \mathcal{V}$ consisting of patches that are representative in the sense that every patch in $\mathcal{V}$ is similar to some patch in the set A. We define the score of a set exemplars A as:

$$\mathcal{F}_R(A) = \sum_{j \in \mathcal{V}} \max_{i \in A} \mathbf{S}_{ij}, \qquad (1)$$



Distract a bit

Maximizing the following set function is NP-hard.

$$\mathcal{F}_R(A) = \sum_{j \in \mathcal{V}} \max_{i \in A} \mathbf{S}_{ij}$$

Distract a bit

Maximizing the following set function is NP-hard.

$$\mathcal{F}_R(A) = \sum_{j \in \mathcal{V}} \max_{i \in A} \mathbf{S}_{ij}$$

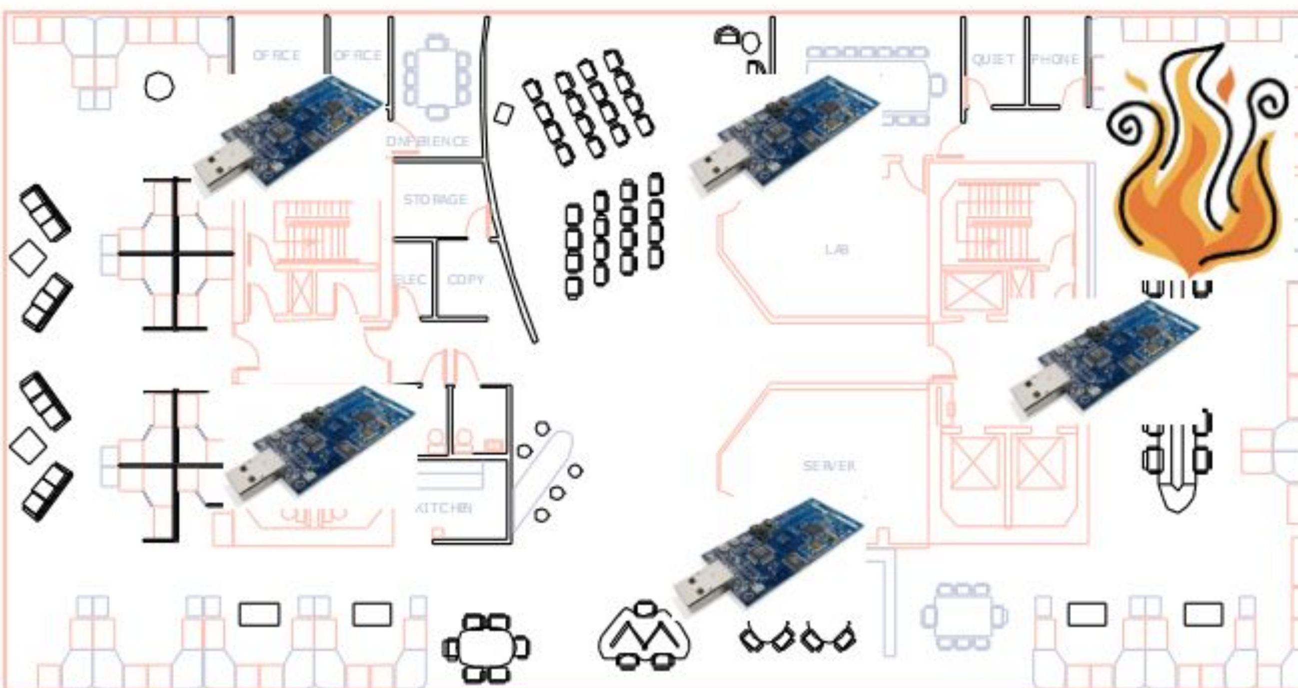
A more general, well-known problem is the facility location problem

Distract a bit

Maximizing the following set function is NP-hard.

$$\mathcal{F}_R(A) = \sum_{j \in \mathcal{V}} \max_{i \in A} \mathbf{S}_{ij}$$

A more general, well-known problem is the facility location problem, placing sensors to monitor temperature.



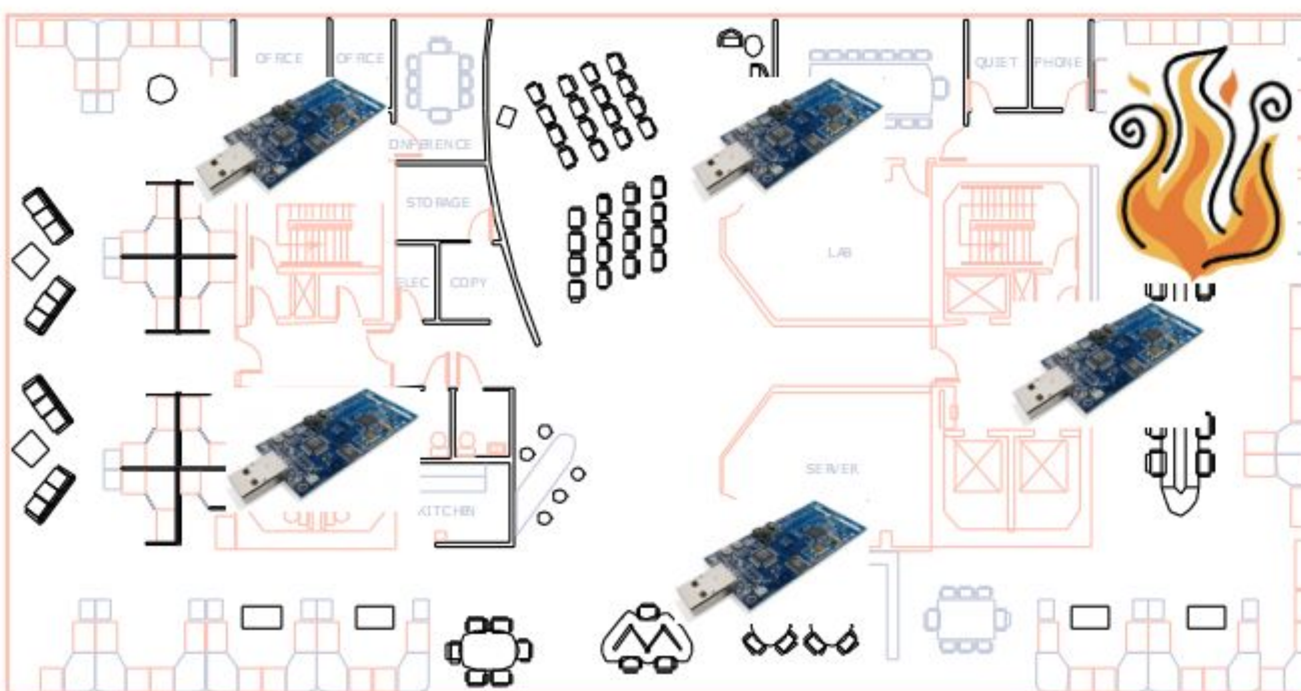
Distract a bit

$$\mathcal{F}_R(A) = \sum_{j \in \mathcal{V}} \max_{i \in A} \mathbf{S}_{ij}$$

facility location problem

$$A^* = \max_A \left\{ \mathcal{F}(A) \equiv \sum_{j \in \mathcal{V}} \max_{i \in A} \mathbf{S}_{ij} - \sum_{i \in A} c_i \right\}$$

maximizing the above is NP-hard.



Distract a bit

facility location problem

$$A^* = \max_A \left\{ \mathcal{F}(A) \equiv \sum_{j \in \mathcal{V}} \max_{i \in A} \mathbf{S}_{ij} - \sum_{i \in A} c_i \right\}$$

maximizing the above is NP-hard.

Good property, the function is

1. monotonically increasing, $\mathcal{F}(A) \leq \mathcal{F}(B)$ for all $A \subseteq B$.

Distract a bit

facility location problem

$$A^* = \max_A \{\mathcal{F}(A) \equiv \sum_{j \in \mathcal{V}} \max_{i \in A} \mathbf{S}_{ij} - \sum_{i \in A} c_i\}$$

maximizing the above is NP-hard.

Good property, the function is

1. monotonically increasing, $\mathcal{F}(A) \leq \mathcal{F}(B)$ for all $A \subseteq B$.
2. submodular, or diminishing return property

$\mathcal{F}(A \cup a) - \mathcal{F}(A) \geq \mathcal{F}(A \cup \{a, b\}) - \mathcal{F}(A \cup b)$, for all $A \subseteq \mathcal{V}$ and $a, b \in \mathcal{V}/A$.

Good properties

1. monotonically increasing, $\mathcal{F}(A) \leq \mathcal{F}(B)$ for all $A \subseteq B$.

$$\mathcal{F}(A) \equiv \sum_{j \in \mathcal{V}} \max_{i \in A} \mathbf{S}_{ij}$$

Good properties

1. monotonically increasing, $\mathcal{F}(A) \leq \mathcal{F}(B)$ for all $A \subseteq B$.
2. submodular, or diminishing return property

$\mathcal{F}(A \cup a) - \mathcal{F}(A) \geq \mathcal{F}(A \cup \{a, b\}) - \mathcal{F}(A \cup b)$, for all $A \subseteq \mathcal{V}$ and $a, b \in \mathcal{V}/A$.

$$\mathcal{F}(A) \equiv \sum_{j \in \mathcal{V}} \max_{i \in A} \mathbf{S}_{ij}$$

To be greedy is good

The two properties make a greedy algorithm give a near-optimal solution with (1-1/e)-approximation bound.

Algorithm 1 Greedy Selection Algorithm

Input: $\mathcal{V}, \mathcal{F}, K$
Output: a subset A with $|A| \leq K$

```
initialize A = ∅, k = 0
while k ≤ K do
    for all i ∈ V/A do
        compute Δ(i) = F(A ∪ {i}) − F(A)
    end for
    i* = arg max_{i∈V/A} Δ(i)
    if Δ(i*) < 0 then
        return A
    else
        A = A ∪ {i*}
        k = k + 1
    end if
end while
return A
```

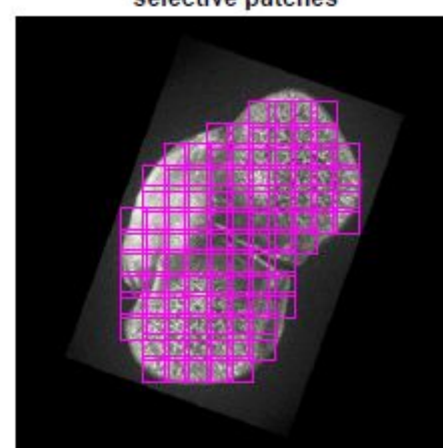
M. Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In Optimization Techniques, pages 234–243. Springer, 1978

Exemplar Selection -- Rule 1

Representative in feature space

$\mathbf{S} \in \mathbb{R}^{M \times M}$ where $\mathbf{S}_{ij}$ is the similarity (a non-negative value) between patch i and patch j. Our aim is to select a subset $A \subseteq \mathcal{V}$ consisting of patches that are representative in the sense that every patch in $\mathcal{V}$ is similar to some patch in the set A. We define the score of a set exemplars A as:

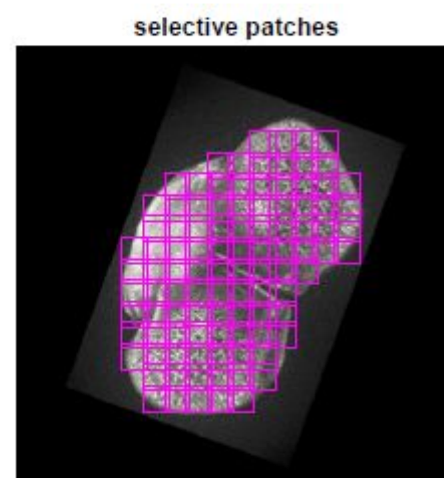
$$\mathcal{F}_R(A) = \sum_{j \in \mathcal{V}} \max_{i \in A} \mathbf{S}_{ij}, \tag{1}$$



Exemplar Selection -- Rule 2

Spatially distributed in input space

$$\mathcal{F}_S(A) = \sum_{j \in \mathcal{V}} \max_{i \in A} \mathbf{L}_{ij}$$

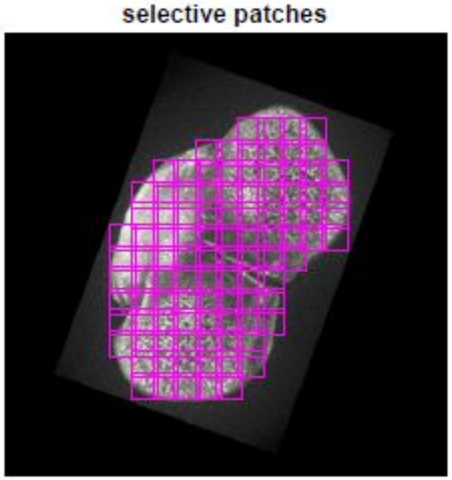


Exemplar Selection -- Rule 3

Discriminative power

$$\mathcal{F}_D(A) = \sum_{i \in A} \frac{\max_c N_c^i}{\sum_c N_c^i} - |A|, \qquad (5)$$

where N_c^i is the number of exemplars from the c^{th} class that are assigned to the i^{th} cluster. As shown in [10], Eq. 5 is also a monotonically increasing submodular function.

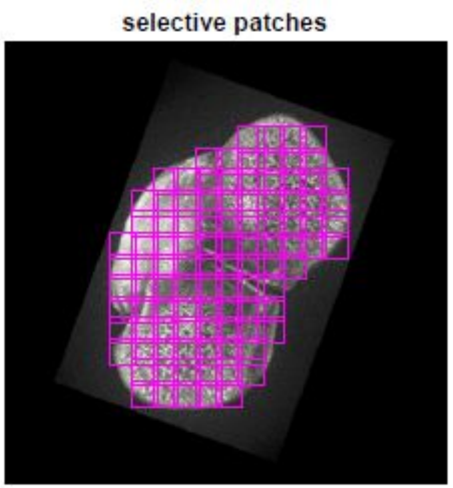


Exemplar Selection -- Rule 4

Class balance: We further adopt the balancing term introduced in [11] to balance the number of exemplars belonging to different classes:

$$\mathcal{F}_B(A) = \sum_c \log(|A_c| + 1) \qquad (6)$$

where A_c is the subset of exemplars in A belonging to class c. The proof can be found in [11] that the above term is monotonically increasing and a submodular function.



Exemplar Selection -- Rule 5

Cluster compactness

$$\mathcal{F}_C(A) = -\sum_{i \in A} p(i)\log(p(i)) - |A| \qquad (7)$$

where $p(i) = \frac{|C_i|}{|\mathcal{V}|}$ is the prior probability of a patch belonging to the i^{th} exemplar cluster. This is a monotonically increasing submodular function as shown in [18].

Exemplar Selection -- Objective

$$\begin{aligned}\mathcal{F}(A) \equiv \sum_{j=1}^{M} \max_{i \in A} \mathbf{S}_{ij} + \lambda_S \sum_{j=1}^{M} \max_{i \in A} \mathbf{L}_{ij} \\ + \lambda_D \left(\sum_{i \in A} \frac{\max_c N_c^i}{\sum_c N_c^i} - |A| \right) \\ + \lambda_B \sum_c \log(|A_c| + 1) \\ + \lambda_C \left(- \sum_{i \in A} p(i) \log(p(i)) - |A| \right)\end{aligned} \tag{8}$$

where $\{\lambda_S, \lambda_D, \lambda_B, \lambda_C\}$ are hyperparameters that weigh the relative contribution of each term. We note that $\mathcal{F}(\varnothing) = 0$. As each term is a monotonically increasing submodular function, our objective summing up all the five terms is also a monotonically increasing submodular function. There-

Exemplar Selection -- be greedy

There-

fore

Algorithm 1 Greedy Selection Algorithm

Input: $\mathcal{V}, \mathcal{F}, K$
Output: a subset A with $|A| \leq K$

```
initialize A = ∅, k = 0
while k ≤ K do
    for all i ∈ V/A do
        compute Δ(i) = F(A ∪ {i}) − F(A)
    end for
    i* = arg max_{i∈V/A} Δ(i)
    if Δ(i*) < 0 then
        return A
    else
        A = A ∪ {i*}
        k = k + 1
    end if
end while
return A
```

Exemplar Selection -- be greedy and lazy

a lazy greed algorithm

Algorithm 2 Lazy Greedy Selection Algorithm

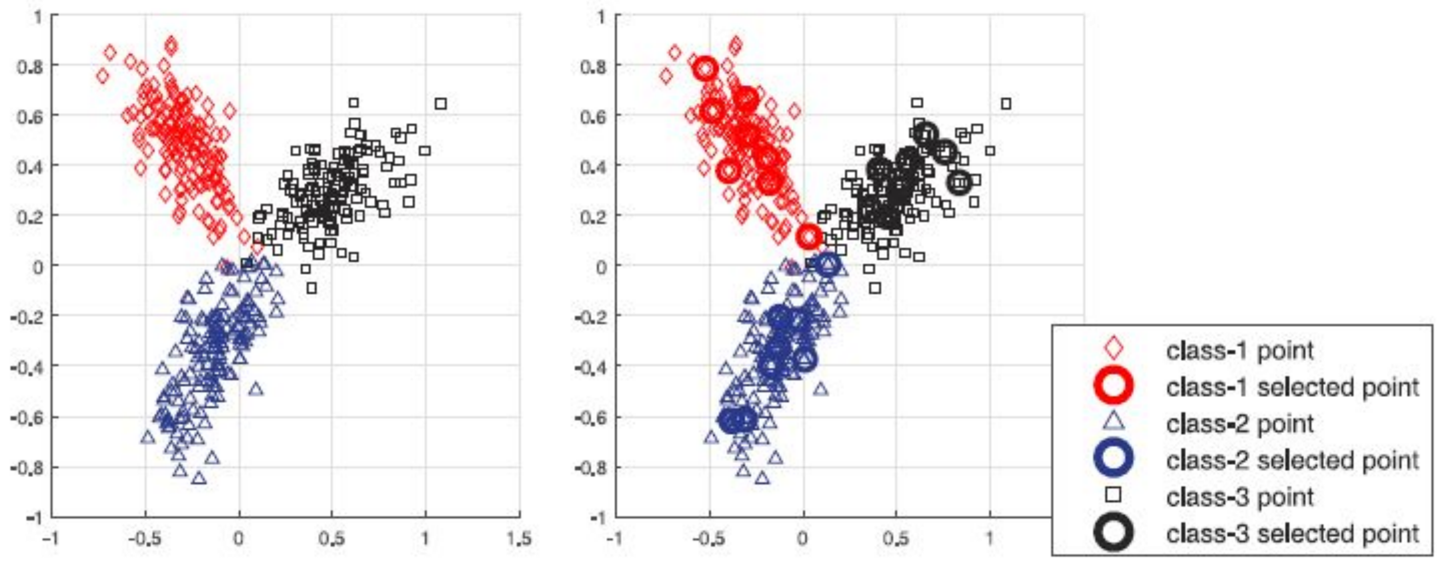
Input: $\mathcal{V}, \mathcal{F}, K$
Output: a subset A with $|A| \leq K$

```
initialize A = ∅, iteration k = 0
for all i ∈ V, compute Δ(i) = F({i})
while k ≤ K do
    i* = arg max_{i∈V/A} Δ(i)
    compute Δ(i*) = F(A ∪ {i*}) − F(A)
    if Δ(i*) ≥ max_{i∈V/A} Δ(i) then
        if Δ(i*) < 0 then
            return A
        else
            A = A ∪ {i*}
            k = k + 1
        end if
    end if
end while
```

M. Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In Optimization Techniques, pages 234–243. Springer, 1978

Exemplar selection -- toy data

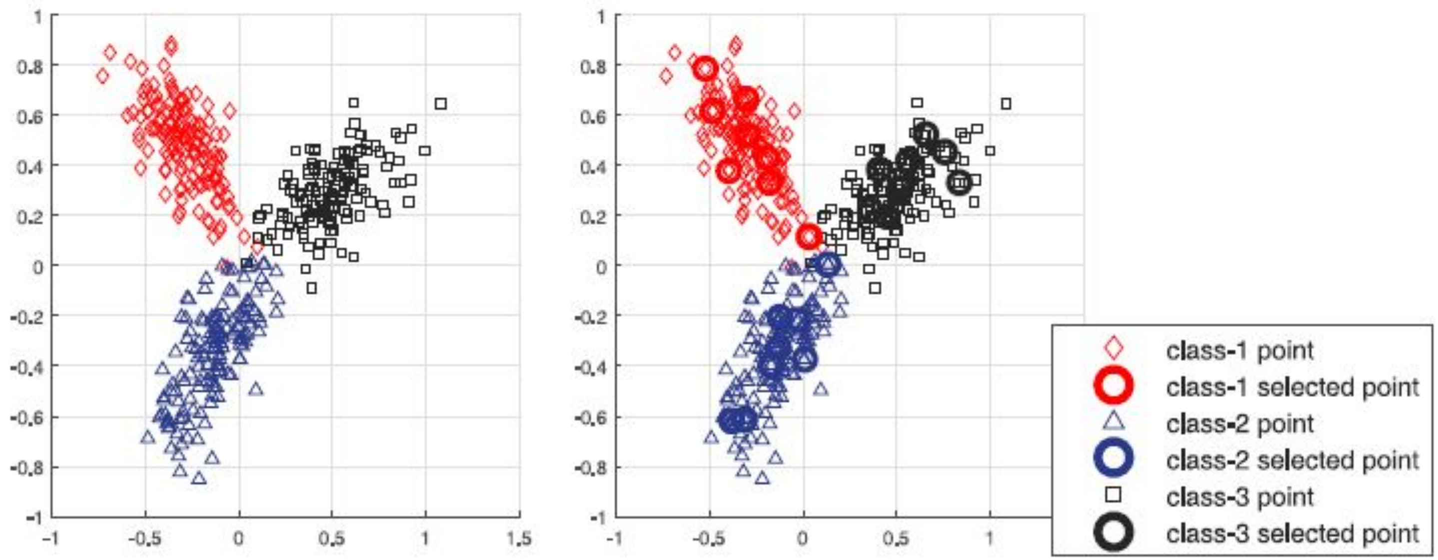
2D data simplify it by merging features and physical coordinates



Exemplar selection -- toy data

2D data simplify it by merging features and physical coordinates

cover the data points from each class

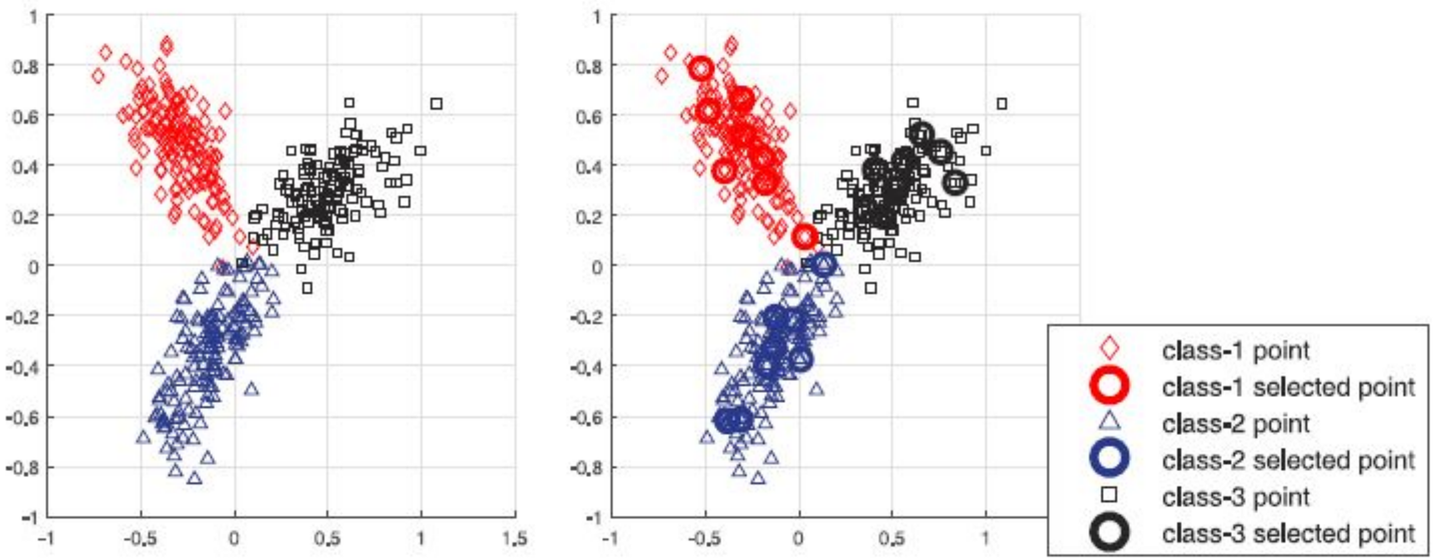


Exemplar selection -- toy data

2D data simplify it by merging features and physical coordinates

cover the data points from each class

maintain discriminative power by sampling near class boundaries



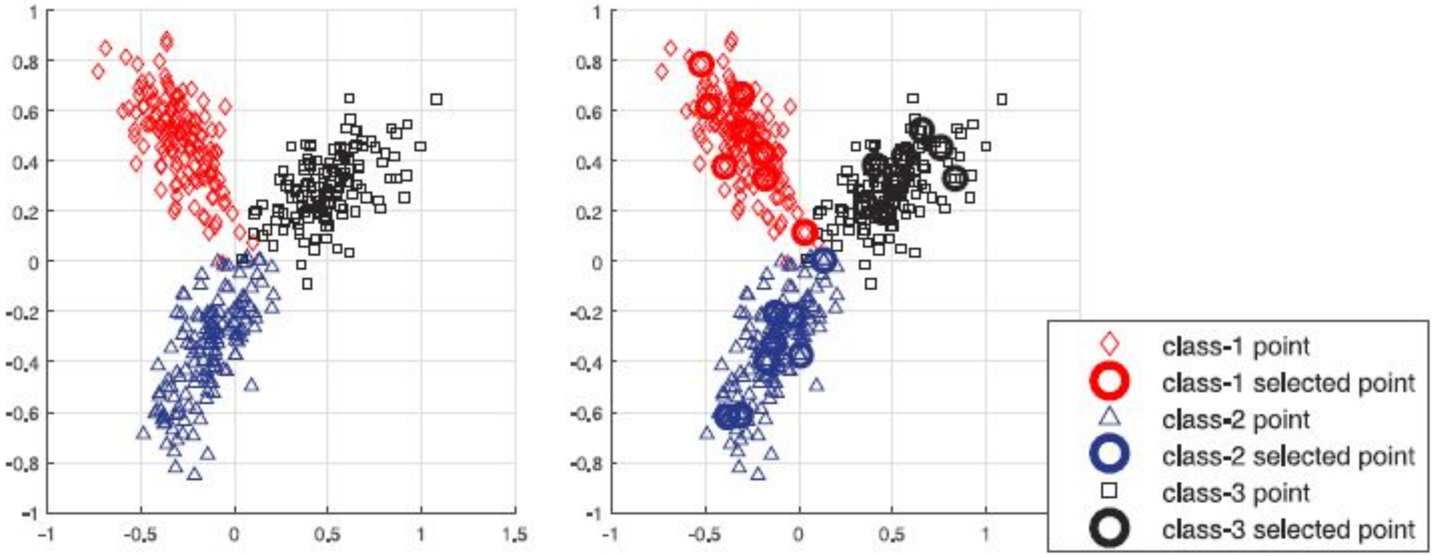
Exemplar selection -- toy data

2D data simplify it by merging features and physical coordinates

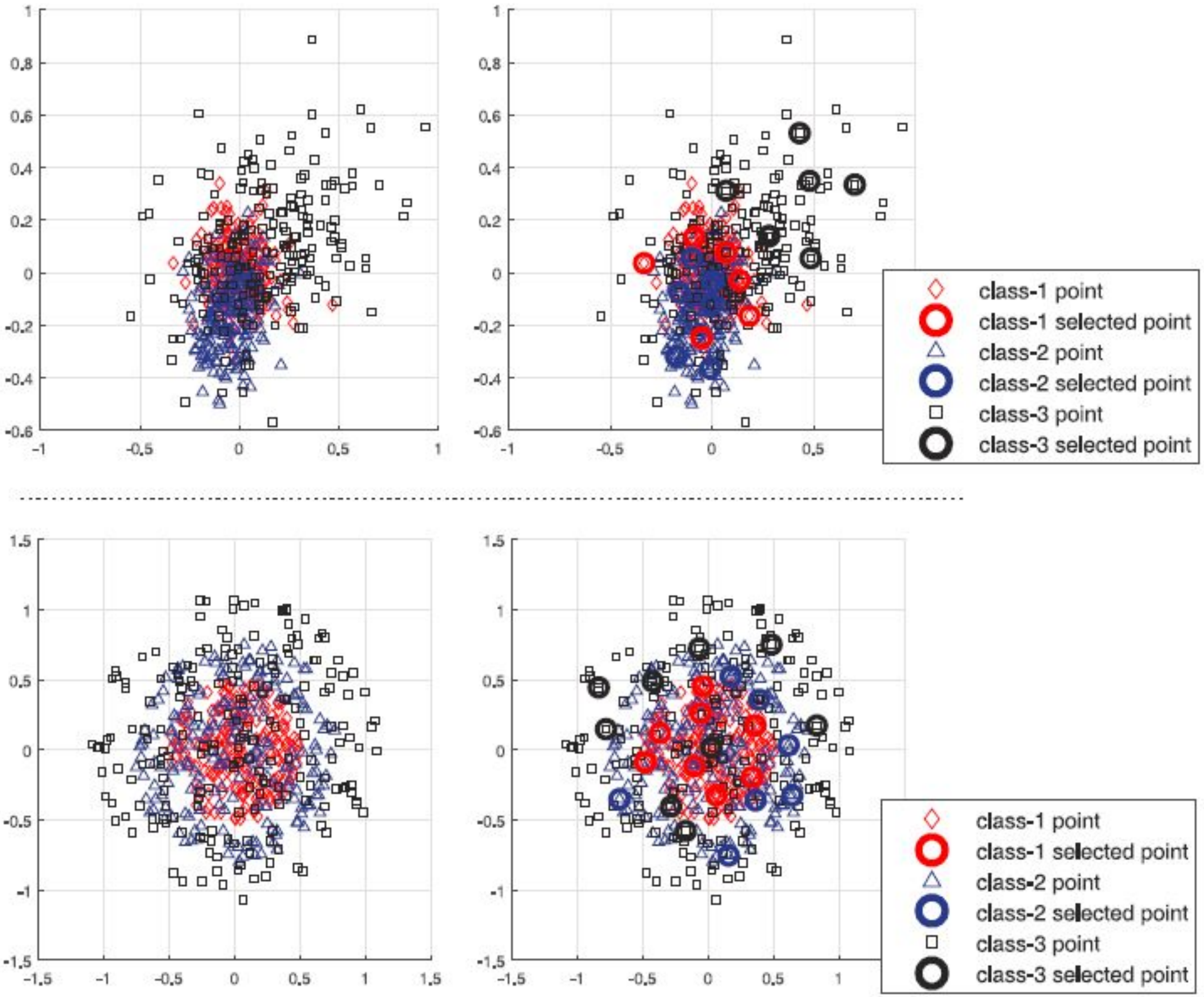
cover the data points from each class

maintain discriminative power by sampling near class boundaries

avoid high inter-class overlap



Exemplar selection -- toy data



Exemplar selection for dictionary

assemble the selected exemplar patches for a discriminative dictionary

Exemplar selection for dictionary

assemble the selected exemplar patches for a discriminative dictionary

The spatial information is also saved as a part of the dictionary.

Spatially aware coding

Sparse coding $$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_0$$

Our robust framework

1. Well-selected patches as dictionary perform better than random patches. --> **exemplar selection**

2. incorporating spatial information of the patches -> **spatially aware coding**

3. pooling+SVM is better than reconstruction-based scheme

Spatially aware coding

Sparse coding $$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_0$$

weighted sparse coding to model spatially aware coding

Spatially aware coding

Sparse coding $$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_0$$

weighted sparse coding to model spatially aware coding

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda_1\|\mathsf{diag}(\mathbf{w})\mathbf{a}\|_1$$

Spatially aware coding

Sparse coding $$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_0$$

weighted sparse coding to model spatially aware coding

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda_1\|\mathsf{diag}(\mathbf{w})\mathbf{a}\|_1$$

or

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda_2\|\mathsf{diag}(\mathbf{w})\mathbf{a}\|_2^2 + \lambda_1\|\mathbf{a}\|_1.$$

Our robust framework

1. Well-selected patches as dictionary perform better than random patches. --> **exemplar selection**

2. incorporating spatial information of the patches -> **spatially aware coding**

3. pooling+SVM is better than reconstruction-based scheme

Spatially aware coding -- fast alternative

$\mathbf{D} \in \mathbb{R}^{p \times m},\ p \geq m.$

$\mathbf{a}^* = \mathrm{argmin}_{\mathbf{a}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2$

Spatially aware coding -- fast alternative

$$\mathbf{D} \in \mathbb{R}^{p \times m}, \; p \geq m.$$

$$\mathbf{a}^* = \operatorname{argmin}_{\mathbf{a}} \|\mathbf{x} - \mathbf{Da}\|_2^2$$

$$\mathbf{a}^* = \mathbf{\Omega}\mathbf{x}, \text{where } \mathbf{\Omega} \equiv (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T$$

Spatially aware coding -- fast alternative

$$\mathbf{D} \in \mathbb{R}^{p\times m},\ p \geq m.$$

$$\mathbf{a}^* = \mathrm{argmin}_{\mathbf{a}} \|\mathbf{x} - \mathbf{Da}\|_2^2$$

$$\mathbf{a}^* = \mathbf{\Omega}\mathbf{x}, \text{where } \mathbf{\Omega} \equiv (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T$$

$$\mathbf{a}^* = \underset{\mathbf{a}}{\mathrm{argmin}} \|\mathbf{x} - \mathbf{Da}\|_2^2 + \lambda\|\mathbf{a}\|_0$$

$$\mathbf{a}^* = \underset{\mathbf{a}}{\mathrm{argmin}} \|\mathbf{\Omega x} - \mathbf{a}\|_2^2 + \lambda_1\|\mathbf{a}\|_1$$

Spatially aware coding -- fast alternative

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_0$$

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\boldsymbol{\Omega}\mathbf{x} - \mathbf{a}\|_2^2 + \lambda_1 \|\mathbf{a}\|_1$$

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda_1 \|\mathsf{diag}(\mathbf{w})\mathbf{a}\|_1$$

Spatially aware coding -- fast alternative

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_0$$

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{\Omega}\mathbf{x} - \mathbf{a}\|_2^2 + \lambda_1\|\mathbf{a}\|_1$$

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda_1\|\mathsf{diag}(\mathbf{w})\mathbf{a}\|_1$$

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{\Omega}\mathbf{x} - \mathbf{a}\|_2^2 + \lambda_1\|\mathsf{diag}(\mathbf{w})\mathbf{a}\|_1$$

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

Spatially aware coding -- fast alternative

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_0$$

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\boldsymbol{\Omega}\mathbf{x} - \mathbf{a}\|_2^2 + \lambda_1 \|\mathbf{a}\|_1$$

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda_1 \|\mathrm{diag}(\mathbf{w})\mathbf{a}\|_1$$

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\boldsymbol{\Omega}\mathbf{x} - \mathbf{a}\|_2^2 + \lambda_1 \|\mathrm{diag}(\mathbf{w})\mathbf{a}\|_1$$

$$\text{where } \boldsymbol{\Omega} \equiv (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T$$

$$a_i^* = \mathrm{sgn}(u_i) \cdot \max(0, |u_i| - \lambda_1 w_i), \text{where } \mathbf{u} = \boldsymbol{\Omega}\mathbf{x}$$

Spatially aware coding -- fast alternative

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_0$$

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{\Omega}\mathbf{x} - \mathbf{a}\|_2^2 + \lambda_1 \|\mathbf{a}\|_1$$

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda_1 \|\mathsf{diag}(\mathbf{w})\mathbf{a}\|_1$$

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{\Omega}\mathbf{x} - \mathbf{a}\|_2^2 + \lambda_1 \|\mathsf{diag}(\mathbf{w})\mathbf{a}\|_1$$

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

$$a_i^* = \operatorname{sgn}(u_i) \cdot \max(0, |u_i| - \lambda_1 w_i), \text{ where } \mathbf{u} = \mathbf{\Omega}\mathbf{x}$$

SACO-I (Spatially Aware Sparse Coding, version-I)

Spatially aware coding -- fast alternative

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda_2 \|\mathsf{diag}(\mathbf{w})\mathbf{a}\|_2^2 + \lambda_1 \|\mathbf{a}\|_1.$$

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\boldsymbol{\Omega}\mathbf{x} - \mathbf{a}\|_2^2 + \lambda_1 \|\mathbf{a}\|_1$$

$$\begin{aligned}
\boldsymbol{\Omega} \equiv& (\mathbf{D}^T\mathbf{D} + \lambda_2 \mathsf{diag}(\mathbf{w})^2)^{-1}\mathbf{D}^T \\
\mathbf{u} =& \boldsymbol{\Omega}\mathbf{x} \\
a_i^* =& \operatorname{sgn}(u_i) \cdot \max(0, |u_i| - \lambda_1) \\
\mathbf{a}^* =& [a_1^*, \ldots, a_i^*, \ldots, a_m^*]^T.
\end{aligned}$$

SACO-II

Implementation details

global average pooling on the sparse codes

linear SVM

Experimental study -- features

dense SIFT descriptor, SACO-I yields 54.40%

Experimental study -- features

dense SIFT descriptor, SACO-I yields 54.40%

VGG19 features

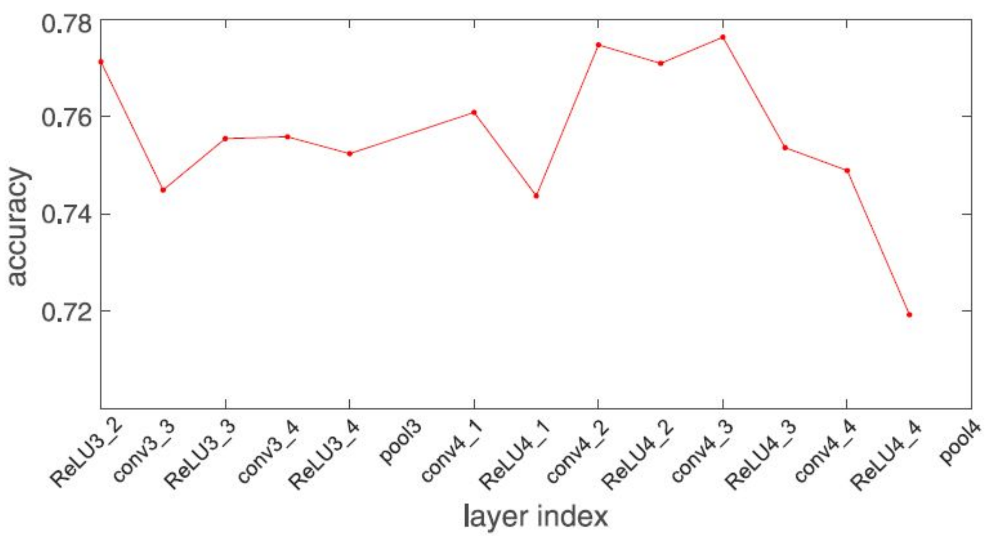


Figure 5. Classification accuracy vs. layer index in VGG19 model. We use features extracted from $conv4\_3$ in the remainder of our experiments.

Experimental study -- features

dense SIFT descriptor, SACO-I yields 54.40%

VGG19 features

SACO-I yields 77.62 at layer conv4_3

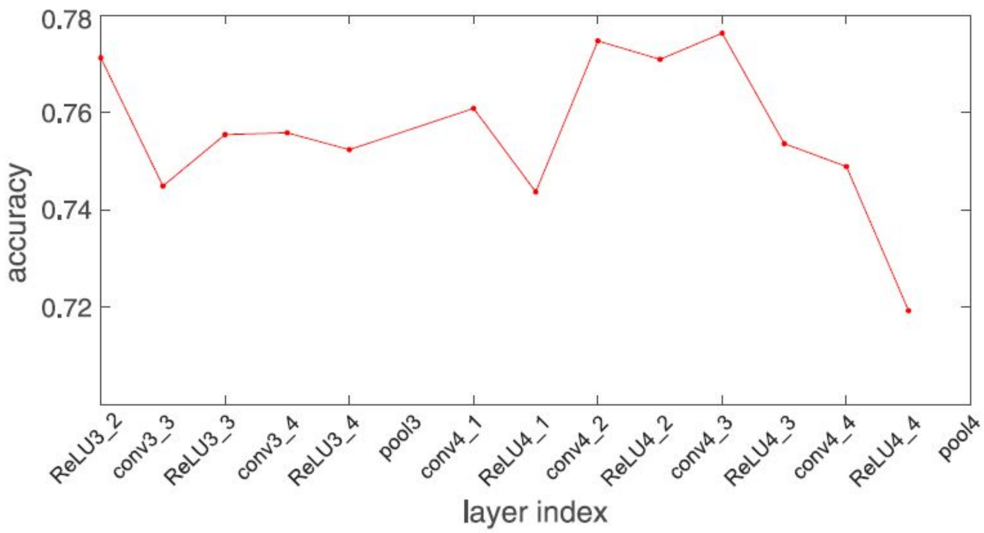


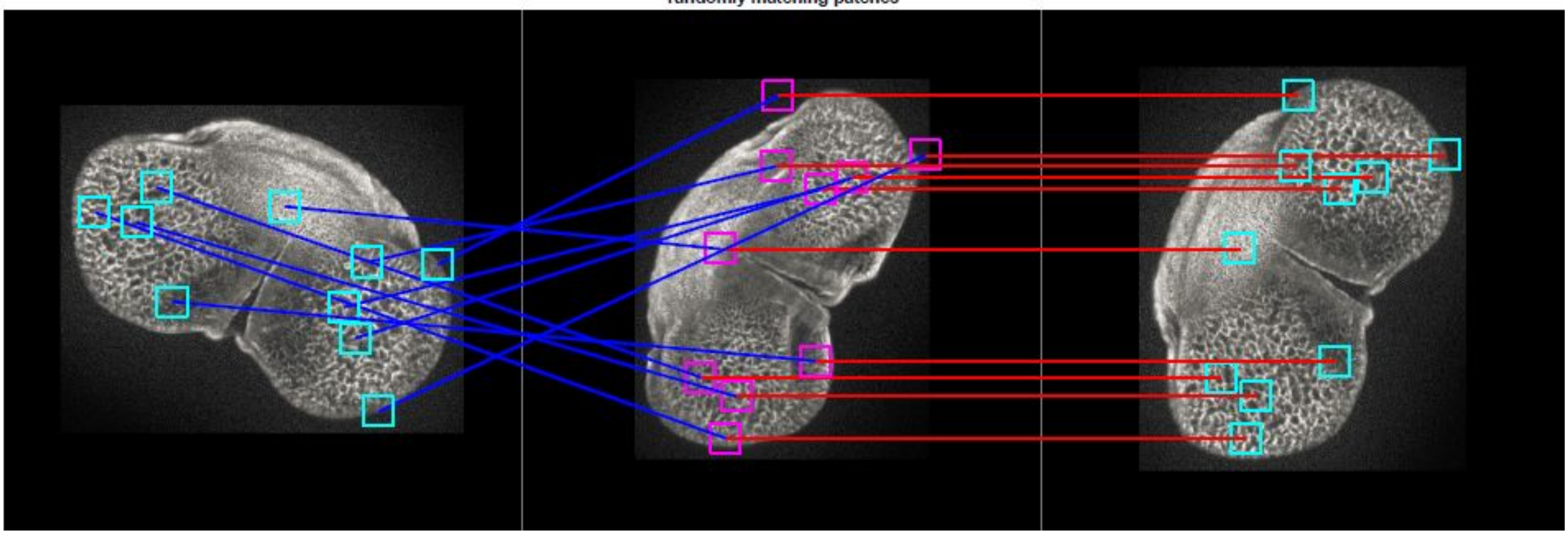
Figure 5. Classification accuracy vs. layer index in VGG19 model. We use features extracted from $conv4\_3$ in the remainder of our experiments.

Experimental study -- features

dense SIFT descriptor, SACO-I yields 54.40%

VGG19 features

SACO-I yields 77.62 at layer conv4_3, RF: 52x52



Experimental study -- dictionary choice

Selected exemplar patches vs. random patches

performance as a function of size

Experimental study -- dictionary choice

Selected exemplar patches vs. random patches

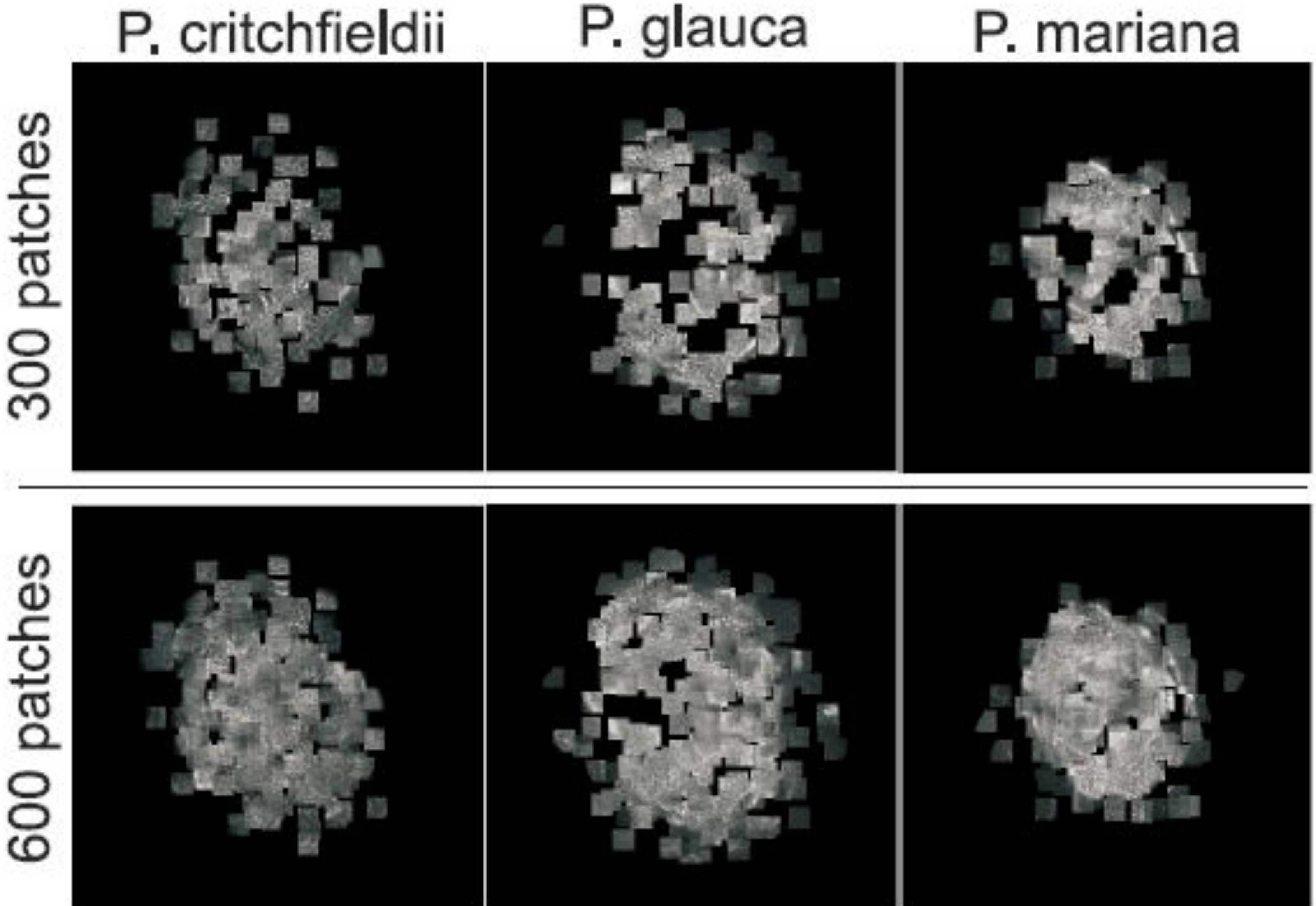
performance as a function of size

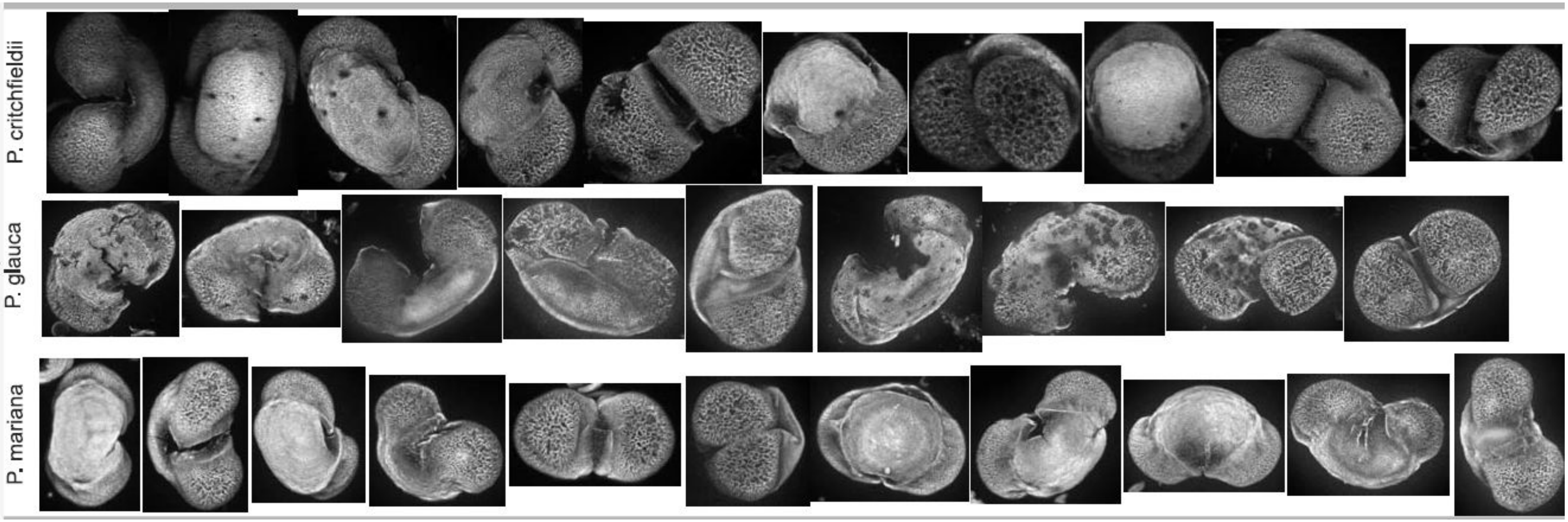
dictionary size	300	512	600
Random Selection	77.66	76.49	77.23
Discriminative Selection	81.75	81.60	82.34

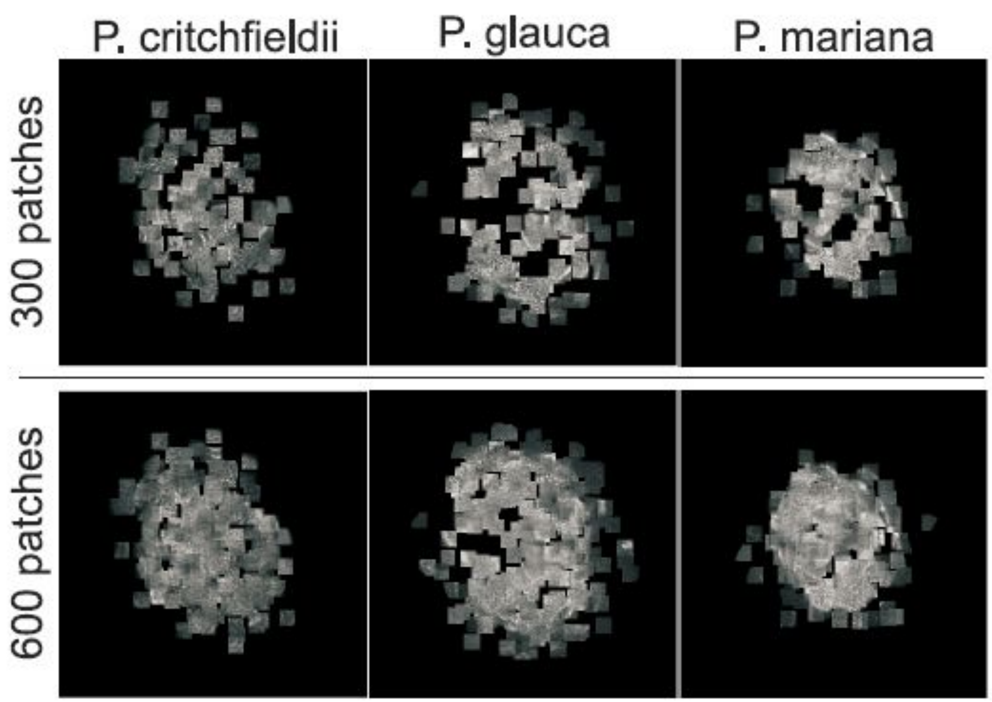
Table 2. Classification accuracy (%) for different sized dictionaries constructed by our discriminative exemplar selection algorithm. Our method consistently outperforms a baseline that selects patches at random from the training set.

Experimental study -- dictionary visualization

Selected exemplar patchesof size




P. critchfieldii
P. glauca
P. mariana


P. critchfieldii
P. glauca
P. mariana
300 patches
600 patches

Experimental study -- baseline comparison

Compared to the strong baselines

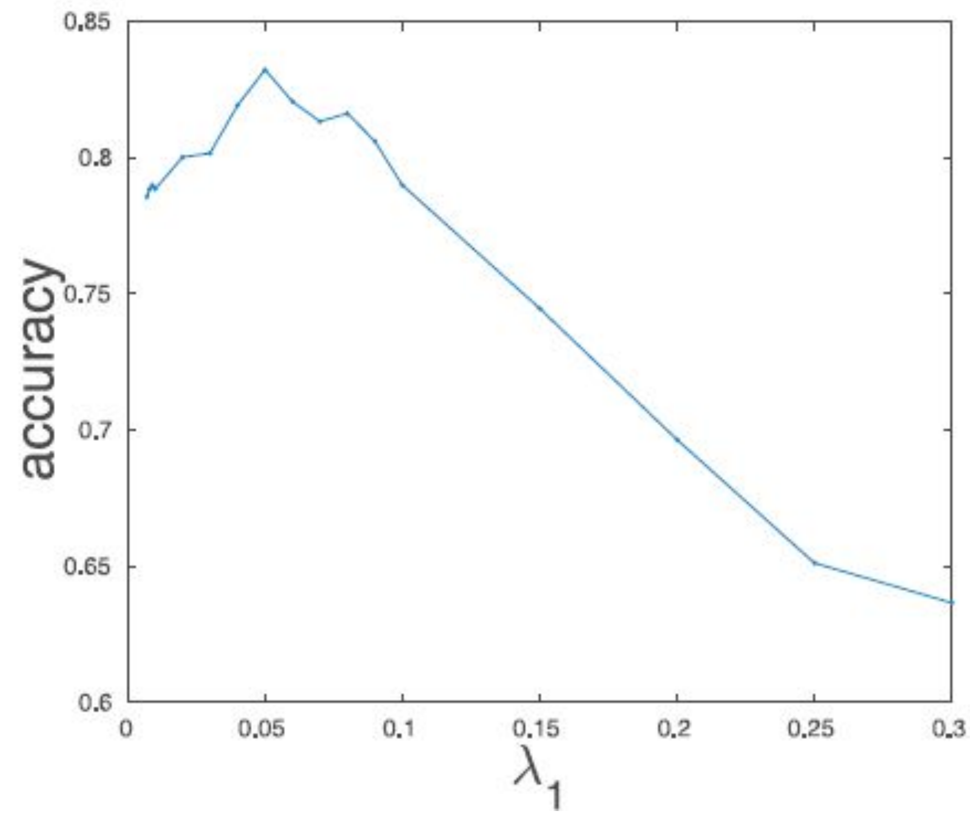
SRC	VGG19+SVM	FV+SVM	SACO-I	SACO-II
62.04	65.11	61.46	83.21	86.13

Table 3. Performance of baselines and our SACO methods measured by classification accuracy (%).

Experimental study -- parameter

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\boldsymbol{\Omega}\mathbf{x} - \mathbf{a}\|_2^2 + \lambda_1 \|\mathbf{a}\|_1$$

$$\begin{aligned}
\boldsymbol{\Omega} &\equiv (\mathbf{D}^T\mathbf{D} + \lambda_2 \mathsf{diag}(\mathbf{w})^2)^{-1}\mathbf{D}^T \\
\mathbf{u} &= \boldsymbol{\Omega}\mathbf{x} \\
a_i^* &= \operatorname{sgn}(u_i) \cdot \max(0, |u_i| - \lambda_1) \\
\mathbf{a}^* &= [a_1^*, \ldots, a_i^*, \ldots, a_m^*]^T.
\end{aligned}$$



Experimental study -- parameter

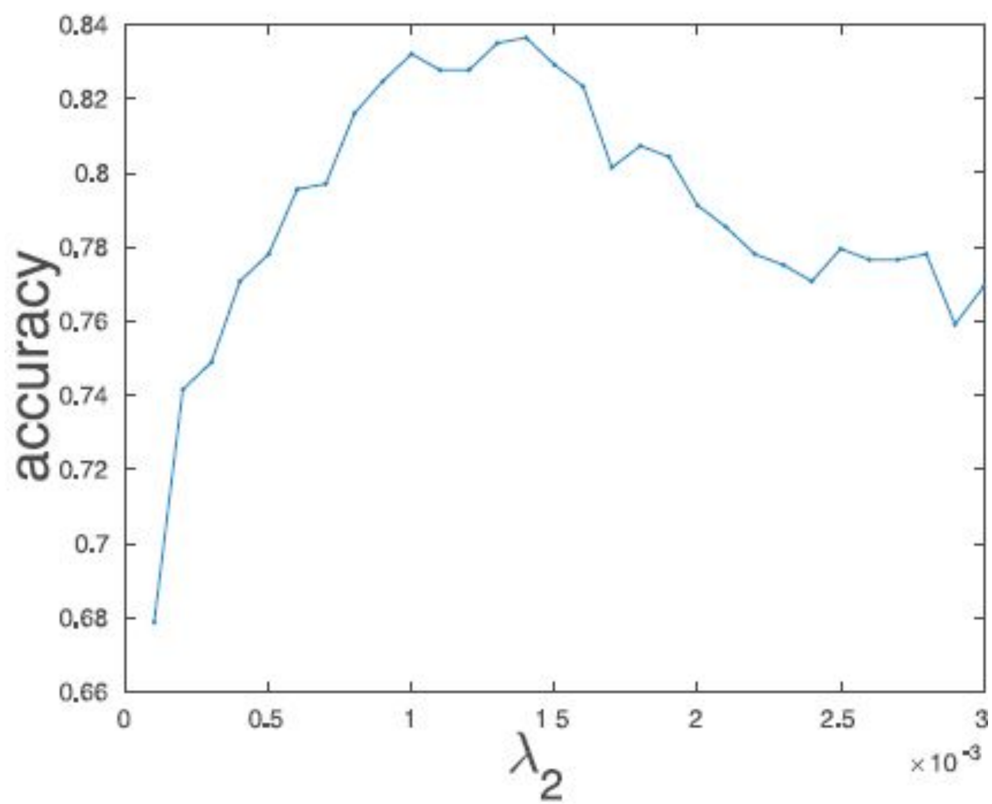
$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\boldsymbol{\Omega}\mathbf{x} - \mathbf{a}\|_2^2 + \lambda_1 \|\mathbf{a}\|_1$$

$$\boldsymbol{\Omega} \equiv (\mathbf{D}^T\mathbf{D} + \lambda_2 \mathsf{diag}(\mathbf{w})^2)^{-1}\mathbf{D}^T$$

$$\mathbf{u} = \boldsymbol{\Omega}\mathbf{x}$$

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

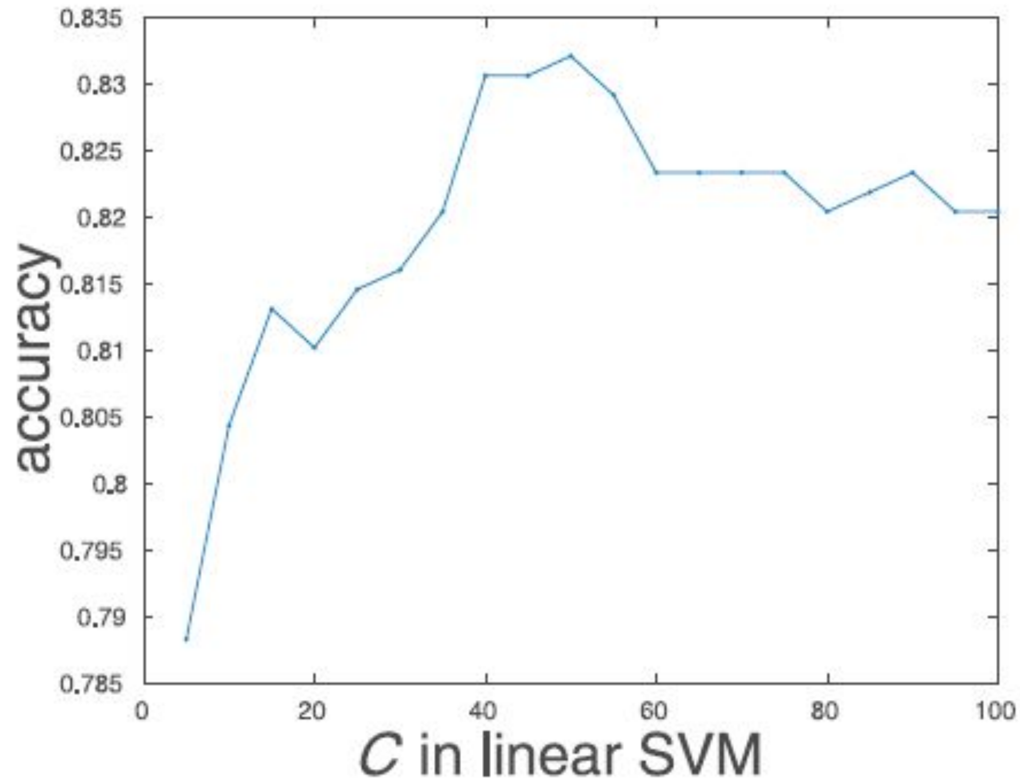
$$\mathbf{a}^* = [a_1^*, \ldots, a_i^*, \ldots, a_m^*]^T.$$



Experimental study -- parameter

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \|\boldsymbol{\Omega}\mathbf{x} - \mathbf{a}\|_2^2 + \lambda_1 \|\mathbf{a}\|_1$$

$$\begin{aligned}
\boldsymbol{\Omega} &\equiv (\mathbf{D}^T\mathbf{D} + \lambda_2 \mathsf{diag}(\mathbf{w})^2)^{-1}\mathbf{D}^T \\
\mathbf{u} &= \boldsymbol{\Omega}\mathbf{x} \\
a_i^* &= \operatorname{sgn}(u_i) \cdot \max(0, |u_i| - \lambda_1) \\
\mathbf{a}^* &= [a_1^*, \ldots, a_i^*, \ldots, a_m^*]^T.
\end{aligned}$$



Extension -- on real testing set

The results are from validation set. We also test it on the real testing set.

Extension -- on real testing set

The results are from validation set. We also test it on the real testing set.

"The results broadly match what we have based on traditional morphometric measurements."

-- Surangi Punyasena

Extension -- cross-domain

learning from modern pollen grains from two species, P. glauca and P. mariana

modern glauca pollen grain

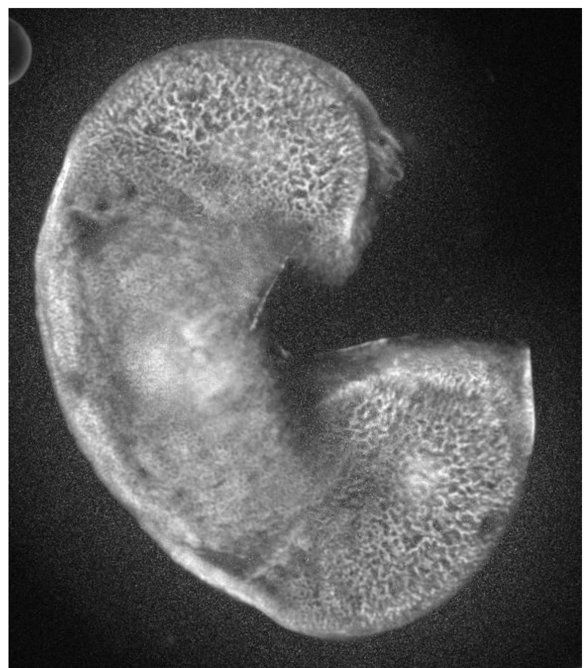
Extension -- cross-domain

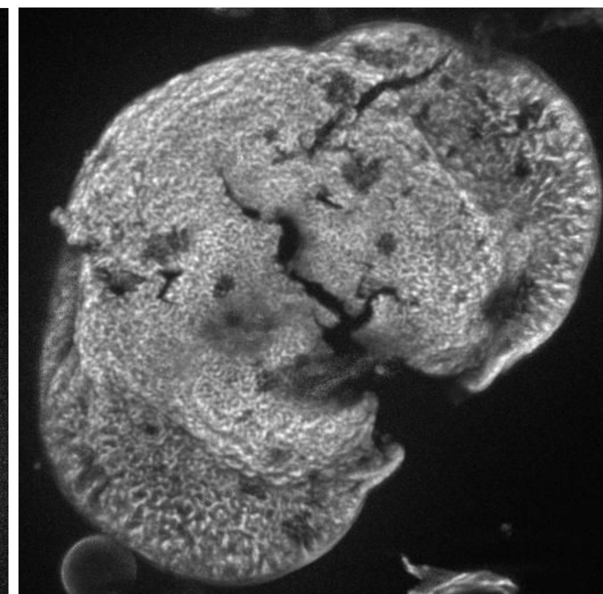
learning from modern pollen grains from two species, P. glauca and P. mariana

testing on fossil ones, which have been destroyed over time and are small in number.

modern glauca pollen grain

fossil glauca pollen grain





Extension -- cross-domain

learning from modern pollen grains from two species, P. glauca and P. mariana

testing on fossil ones, which have been destroyed over time and are small in number.

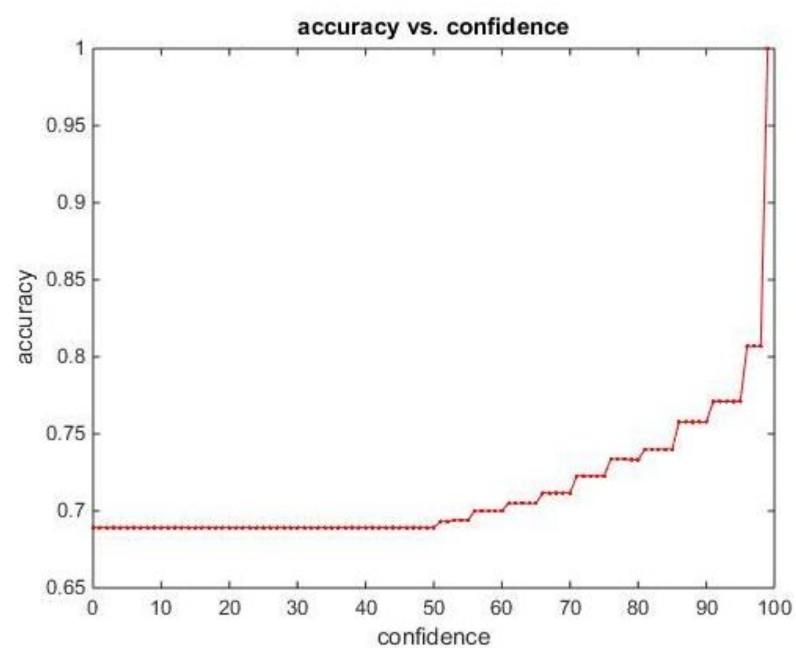
68.9% this is old results without exemplar selection, using outdated SRC classifier without global pooling and SVM learning.

Extension -- cross-domain

learning from modern pollen grains from two species, P. glauca and P. mariana

testing on fossil ones, which have been destroyed over time and are small in number.

68.9% this is old results without exemplar selection, using outdated SRC classifier without global pooling and SVM learning.



Conclusion

- robust system of practical use in new area

Conclusion

- robust system of practical use in new area
- first experiment of matching fossil pollen grains through modern ones at species level

Conclusion

- robust system of practical use in new area
- first experiment of matching fossil pollen grains through modern ones at species level
- New technical directions to explore, embedding selection in neural net, how to exploit confidence score for better training, etc.

Thanks