Improved Bag of Time Model with Feature Fusion

Li Weng, Qianneng Wang, Xizhe Wang, Bingya Wu

Zhejiang Financial College, China

Outline

- Introduction
- The Bag of Time model
- An improved Bag of Time model
- Experiment results
- Conclusion

Introduction

AI

+

Culture

=

AI4Culture

- AI (machine learning in particular) is more commonly utilized in analysis and promotion of cultural heritage content.

Applications of Heritage Content Analysis

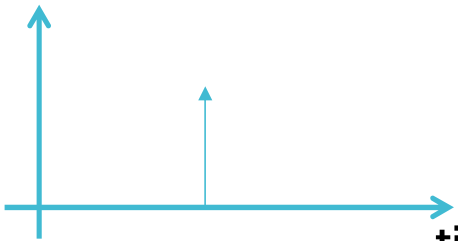
AI4Culture

- Archeology perspective
  - Style classification
  - Author identification
  - Material prediction
  - Time prediction
- Historical perspective
- Aesthetic perspective
- ……

Time Prediction from Cultural Objects

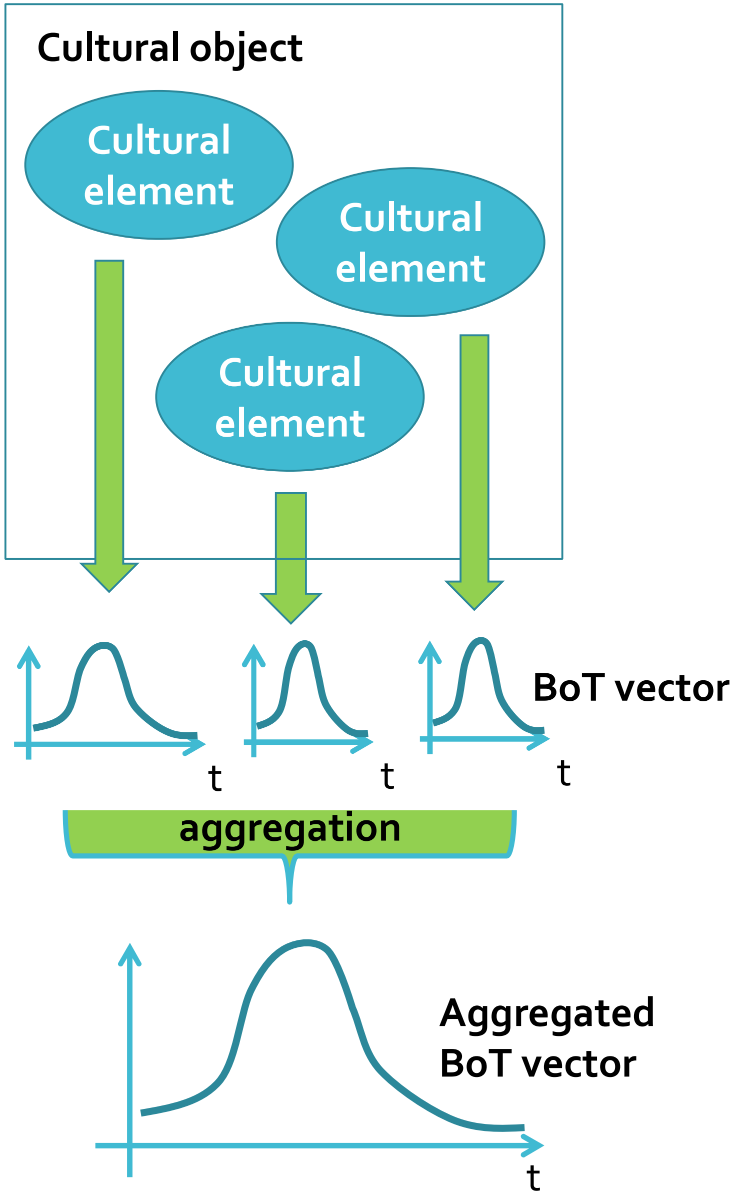
Cultural object

supervised learning (classification / regression)



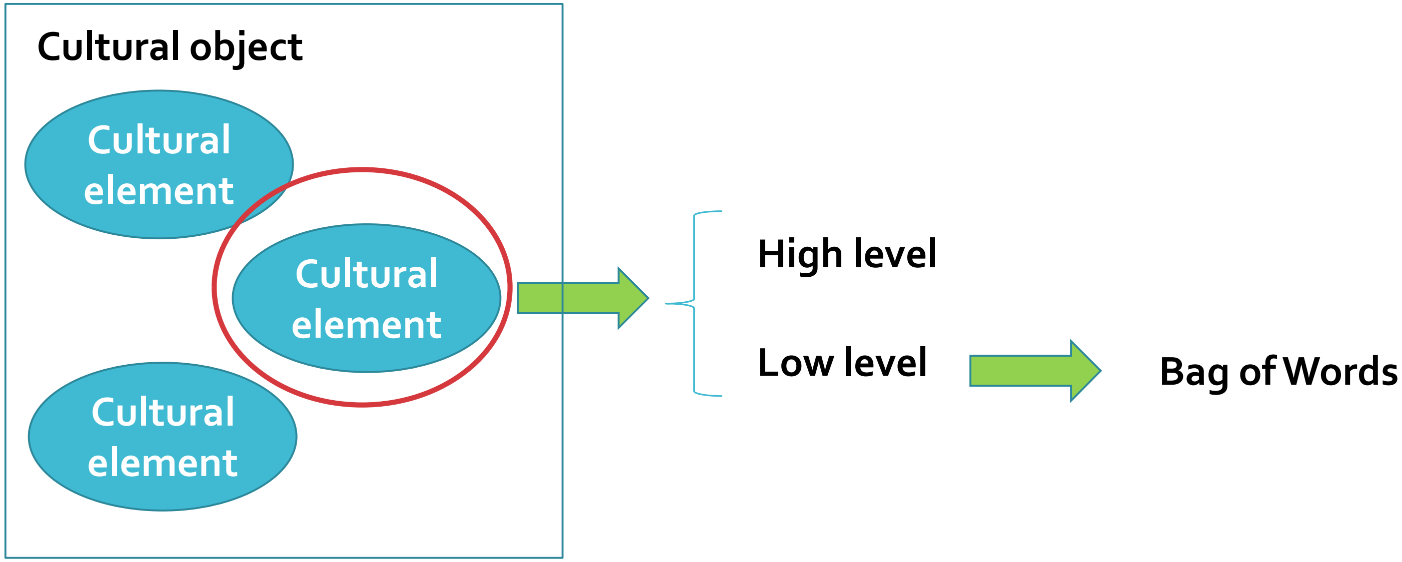
- A cultural object corresponds to a time point, which can be estimated by supervised learning.
- Too simple to reveal insights and model uncertainties.

The Bag of Time Model



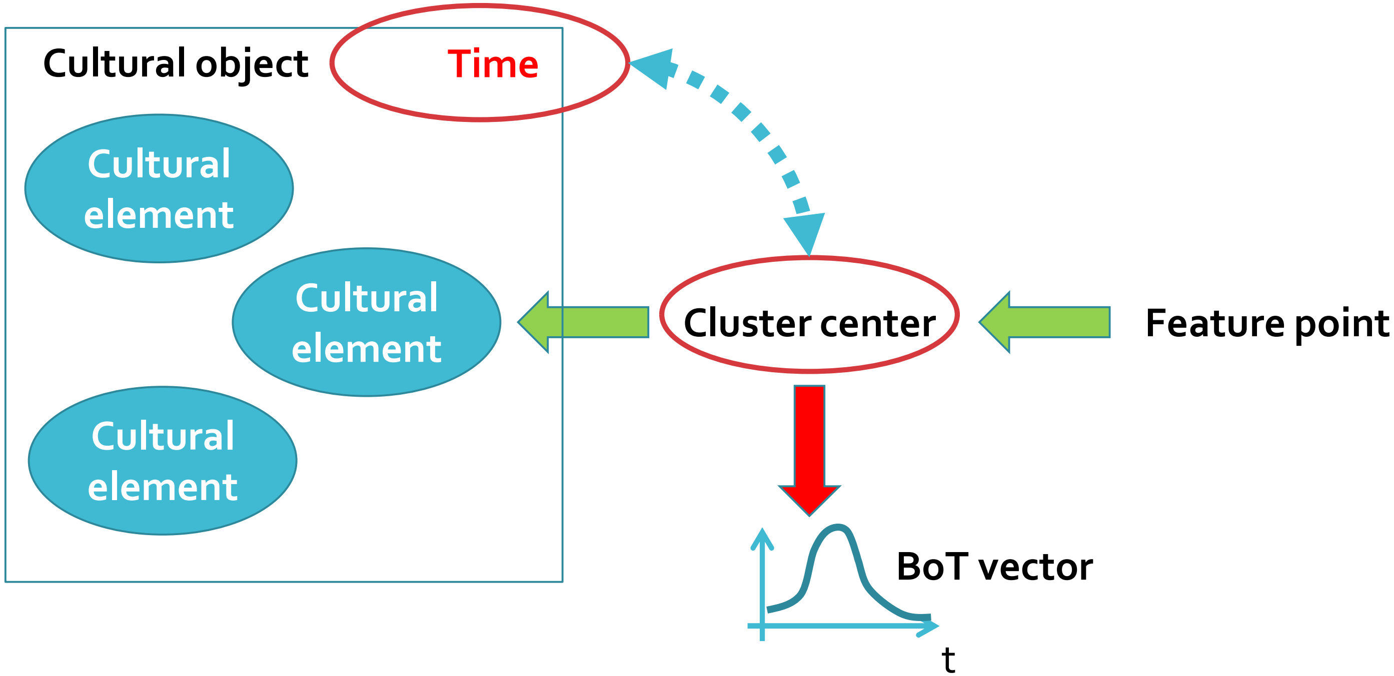
- A cultural object consists of cultural elements.
- Each cultural element represents a time distribution.
- An overall time distribution can be obtained by aggregation.

Define a Cultural Element



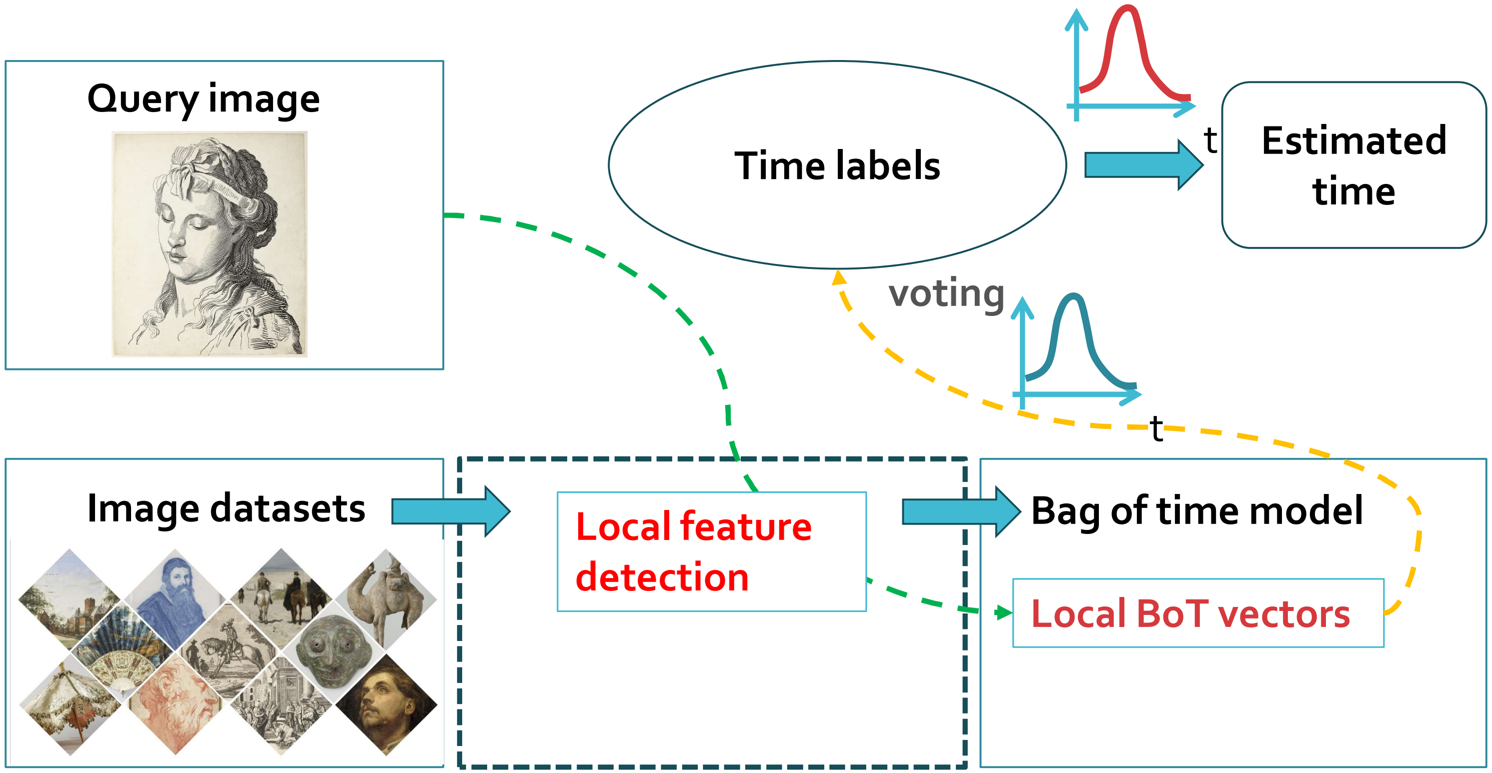
- Cultural elements can be defined on various levels.
- The Bag of Words (BoW) framework offers a low-level representation.

Derive the BoT Model



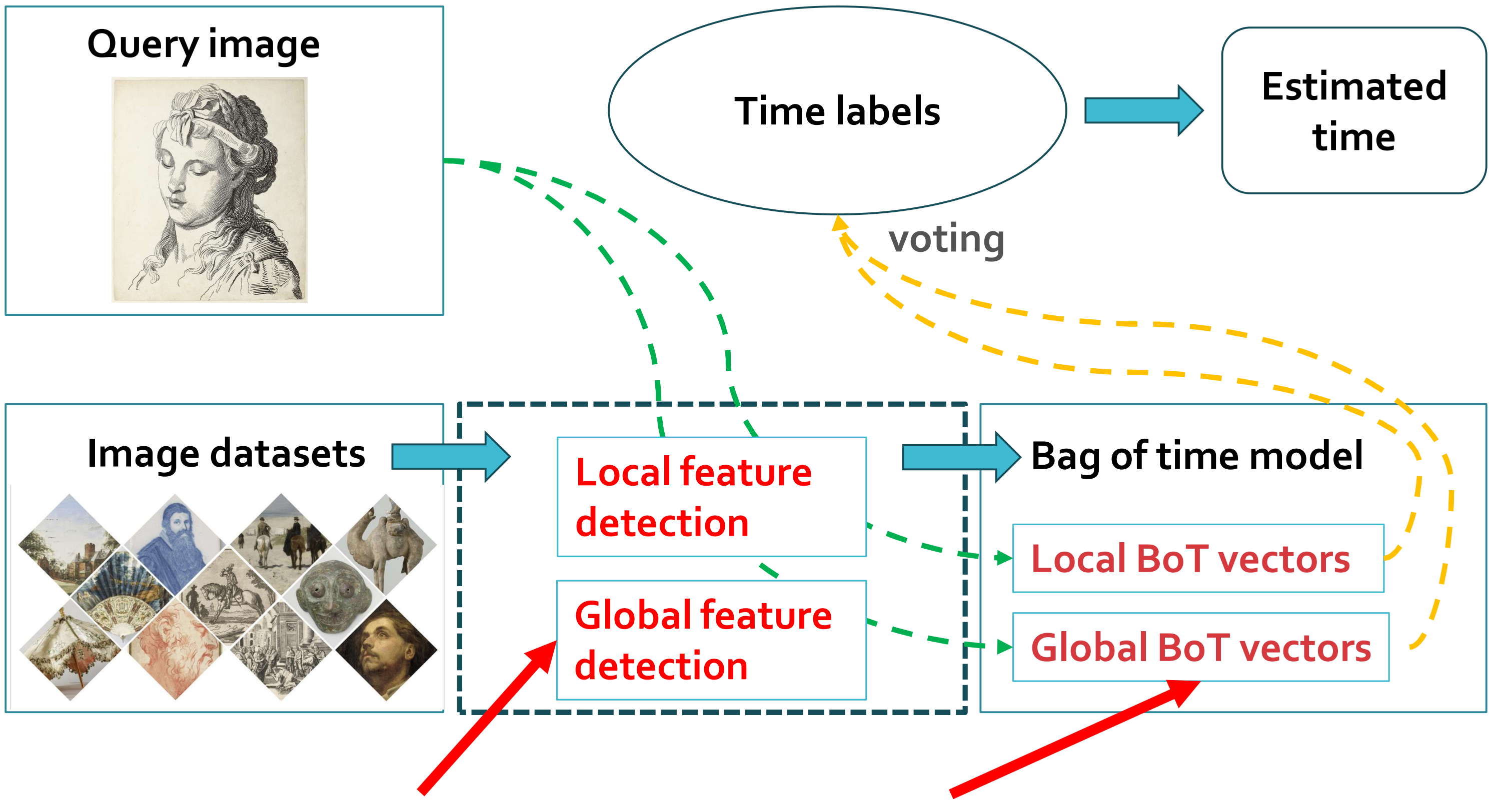
- Each local feature point corresponds to a cluster center.
- Each cluster center corresponds to a cultural element.
- We can estimate a time distribution for each cluster center.
- A BoT model can be built with a training set of images.

Prediction with the Bag of Time Model



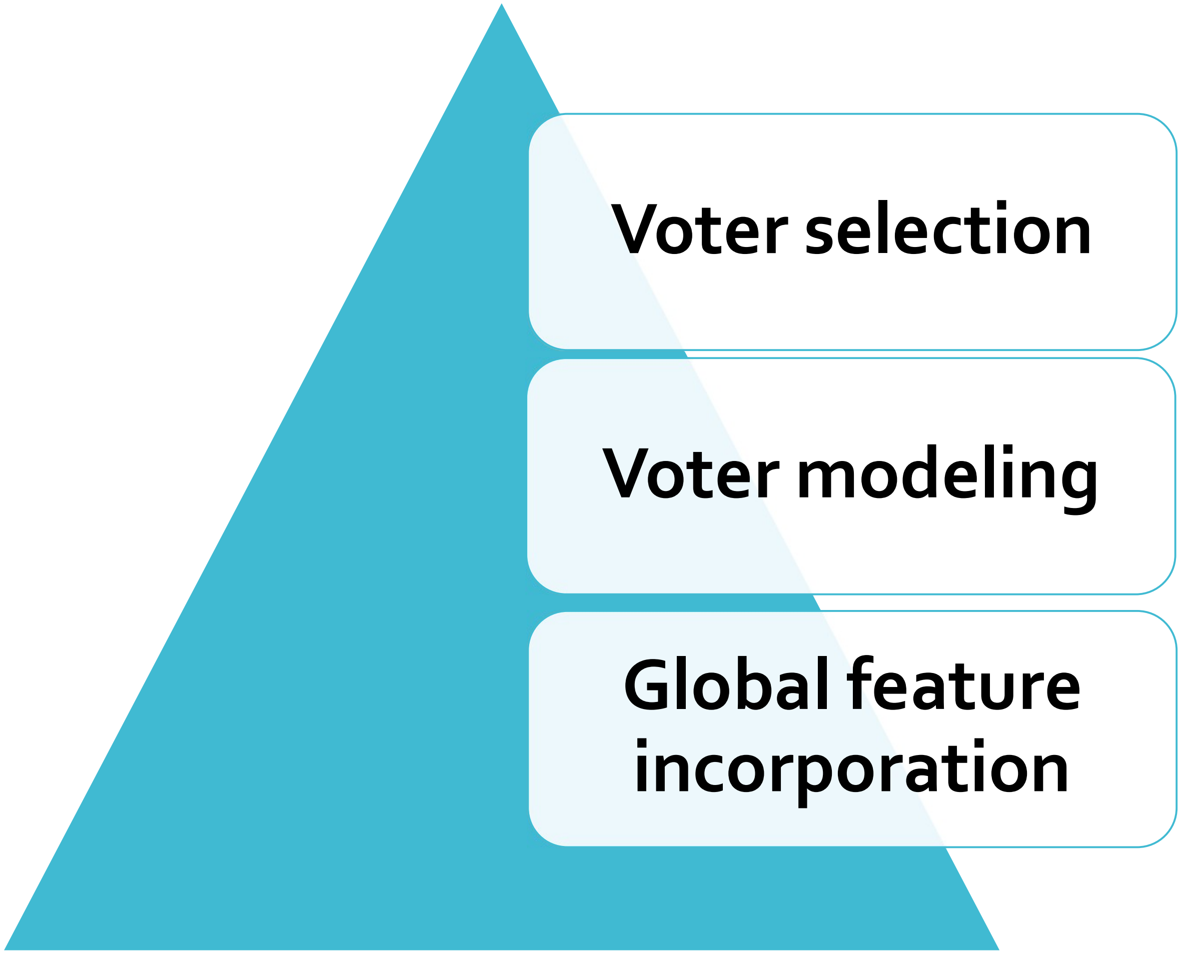
- Compute the aggregated BoT vector for the query image.
- Each feature point casts multiple votes to different time labels.
- The most voted time label is selected.

Prediction with the Improved Bag of Time Model



- In addition to local feature descriptors, each global feature descriptor also casts votes to all time labels.

An Improved Bag of Time Model



Global Feature Incorporation

Aggregation of BoT vectors

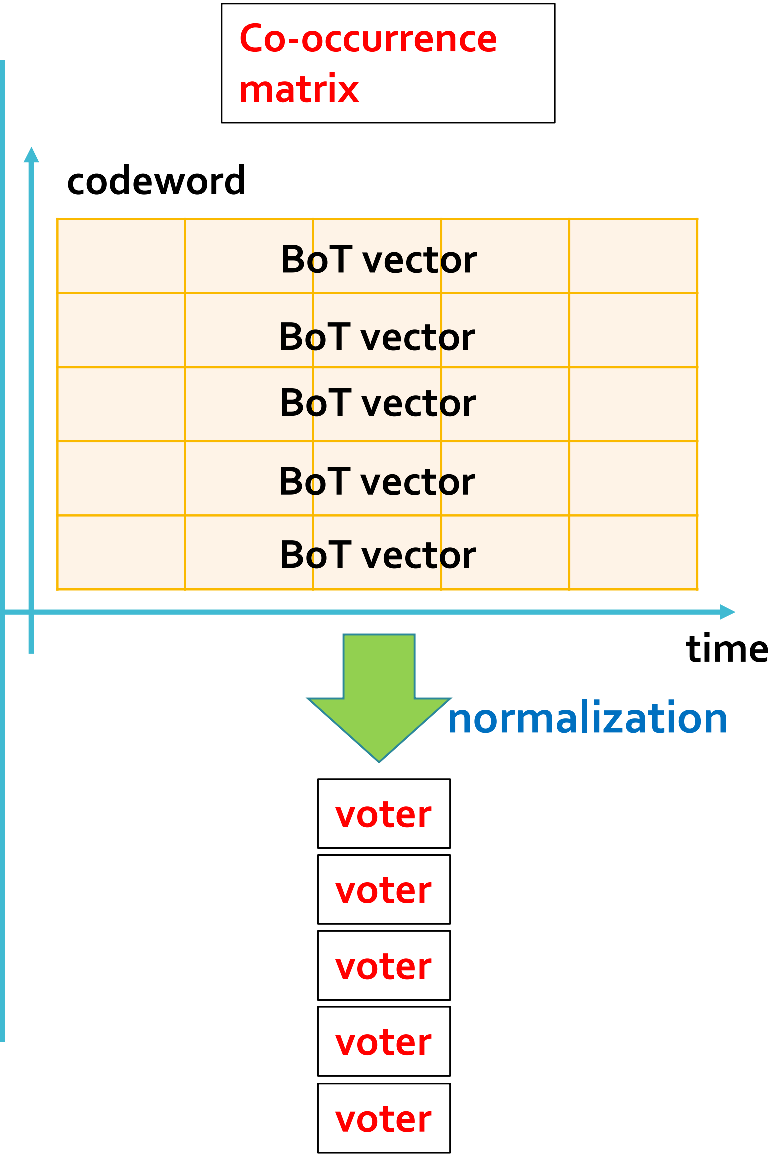
$$a = a^G + W_L * a^L$$

global

local

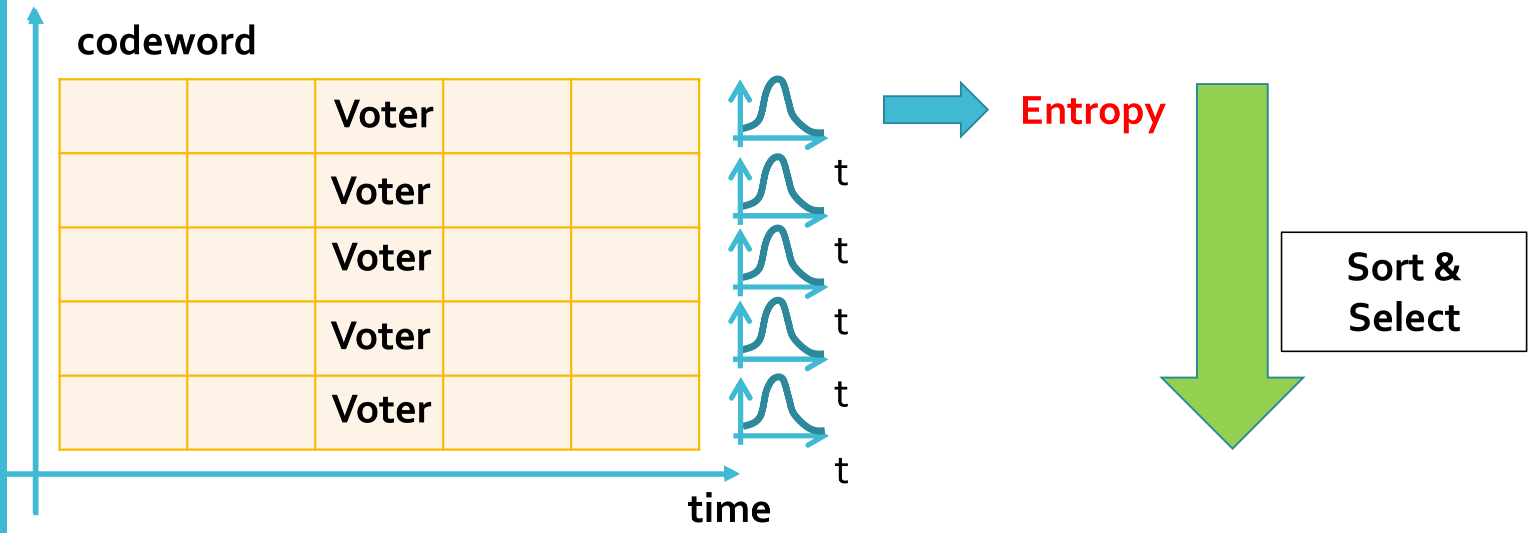
- BoT vectors of global features are given a higher weight.
- The weight is inversely proportional to the number of local feature points.
- It adapts to each image.

Voter Modeling



- **A posteriori voter**
  - Row-wise normalization
- **Likelihood voter**
  - Column-wise normalization
- **Joint probability voter**
  - Matrix-wise normalization

Voter Selection



- Entropy is a well motivated criterion.
- Voters with low entropy (uncertainty) are preferred.

Experiment Overview

dataset	vocabulary size		no. of time labels
	local	global	
Buddha	512, 1024	256, 512	196
Rijksmuseum	2048, 4096	2048, 4096	557

- Two datasets: Buddha (1.2k), Rijksmuseum (100k)
- Local and global features: SIFT, ResNet50
- Different codebook sizes

Effects of Feature Fusion

dataset	vocabulary size		MAE	
	local	global	SIFT	SIFT+ResNet50
Buddha	512	256	372.71	276.00
		512		262.50
	1024	256	372.88	274.60
		512		**256.88**
Rijksmuseum	2048	2048	490.10	264.55
		4096		**249.04**
	4096	2048	492.08	264.81
		4096		251.30

- Different codebook sizes have been tested.
- The MAE (mean absolute error) is significantly reduced (>30%) after using global features.

Effects of Voter Modeling

dataset (vocab. size)	voter type	MAE	
		SIFT	SIFT+ResNet50
Buddha	a posteriori	372.88	256.88
	likelihood	388.44	284.99
(1024, 512)	joint probability	372.41	257.77
Rijksmuseum	a posteriori	490.10	249.04
	likelihood	**466.46**	**143.42**
(2048, 4096)	joint probability	524.81	266.63

- Different codebook sizes have been tested.
- For Buddha, no significant difference is observed.
- For Rijksmuseum, the likelihood voter performs best.

Effects of Voter Selection

dataset (vocab. size)	selection strategy local	selection strategy global	MAE SIFT	MAE SIFT+ResNet50
Buddha	100%	50%	372.88	**256.88**
	50%	100%	372.58	265.63
(1024, 512)	50%	50%	372.58	265.63
	top 32		**361.94**	265.64
	top 16	100%	371.18	267.42
	top 8		397.67	266.85
Rijksmuseum	100%	50%	492.08	264.81
	50%	100%	463.33	272.97
(4096, 2048)	50%	50%	463.33	272.97
	top 32		384.91	220.32
	top 16	100%	**255.72**	162.60
	top 8		325.80	**157.49**

- Different selection strategies have been tested.
- A significant part of voters can be ignored without impacting the MAE.
- Global voters play a dominant role.

Conclusion

- We propose an enhanced Bag-of-Time (BoT) model that improves the task of **time estimation** for cultural heritage images by introducing a **feature fusion strategy** and refining the **voting mechanism**.

- Our method incorporates both **local** and **global** features in a **unified framework**. This **dual-level** representation allows for more **robust modeling** of **temporal cues**, which is beneficial for **heterogeneous** heritage datasets.

- The optimal **formulation** of voters and the **balance** between local and global contributions are still **open problems**. A more supervised approach might work better.

Thank you

- **Li Weng**
- **lweng@zfc.edu.cn**