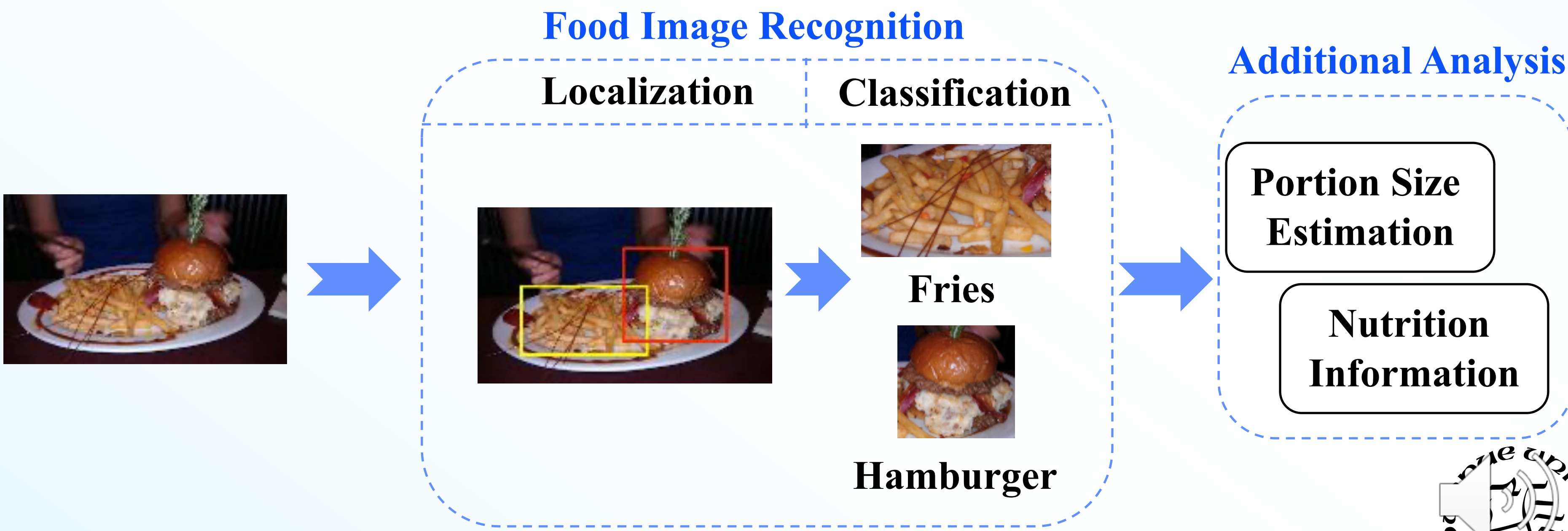
Visual Aware Hierarchy Based Food Recognition

Runyu Mao, Jiangpeng He, Zeman Shao, Sri Kalyan Yarlagadda, and Fengqing Zhu

Video and Image Processing Laboratory
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana, U.S.A.

Food Image Recognition

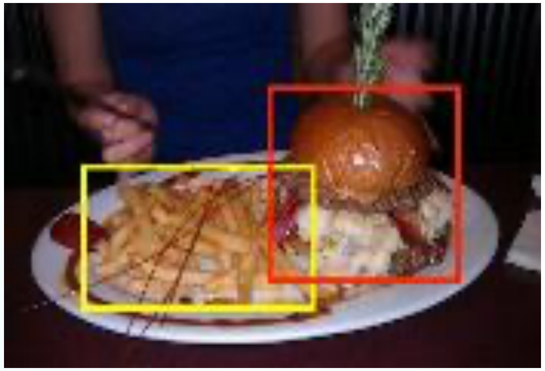
- Conventional dietary assessment methods rely on participant memories which is tedious and error-prone
- Image-based approaches have been integrated into mobile and wearable devices to automatic this process
- Food image recognition provides information on both locations and types of foods in the image



Why Food Localization

- **Most eating occasion images contain multiple foods**

- **For single food image, food localization can eliminate irrelevant background pixels**



Multi-Food Image

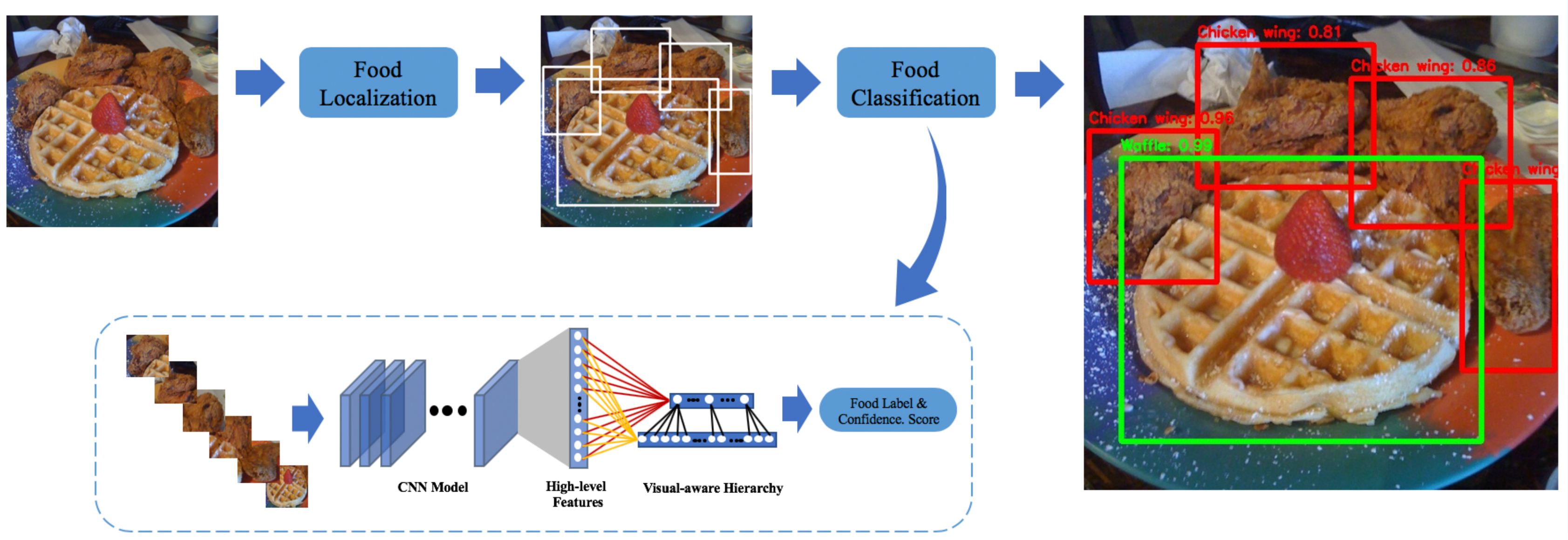
Single-Food Image

Food Classification

- **Food classification is challenging due to the inter-class similarity and intra-class variability**

almond milk **cheese** **cottage cheese** **yogurt**

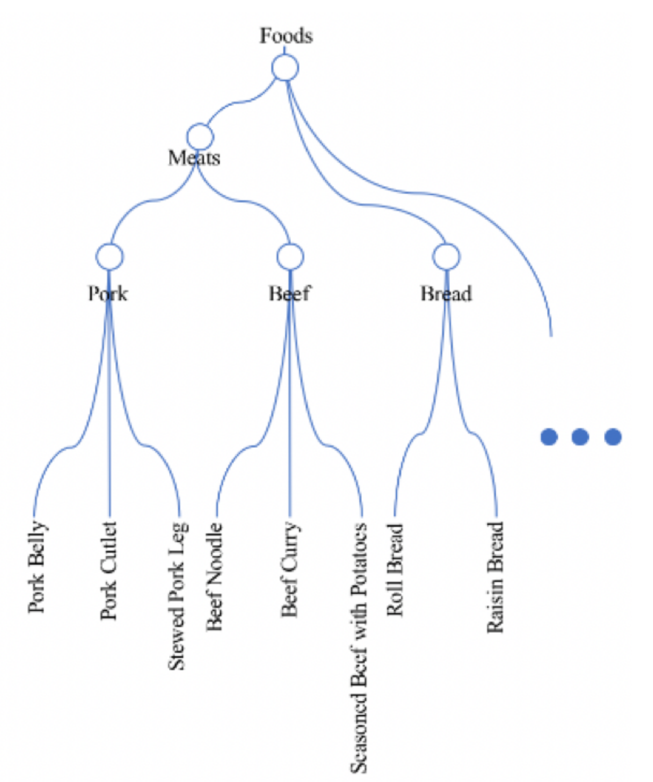
Proposed Food Recognition System



- **A two-step recognition consists of food localization and food classification**
  - Two deep models in sequence
  - Food localization: Faster R-CNN proposes food regions with bounding boxes
  - Food classification: embedding visual aware hierarchy to improve the classification performance

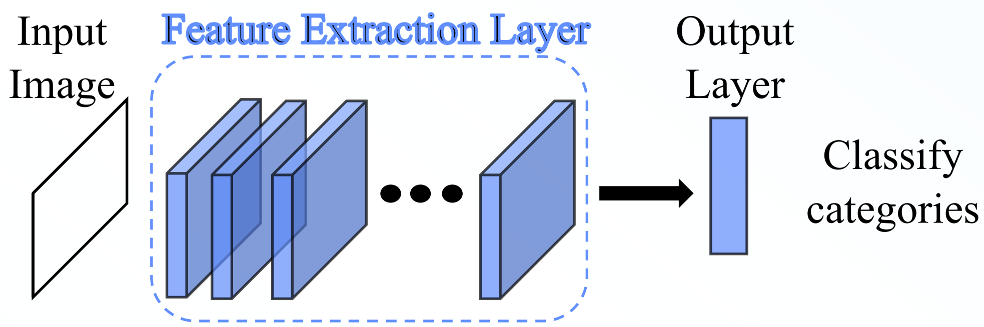
Visual-Aware Hierarchy

- **Hierarchical structure depicts the visual relationship between classes**
- **Visually similar categories are merged as a single cluster**
- **Better mistake: the prediction made by classifier and the true category belong to same cluster**
- **Automatically generate for different datasets**

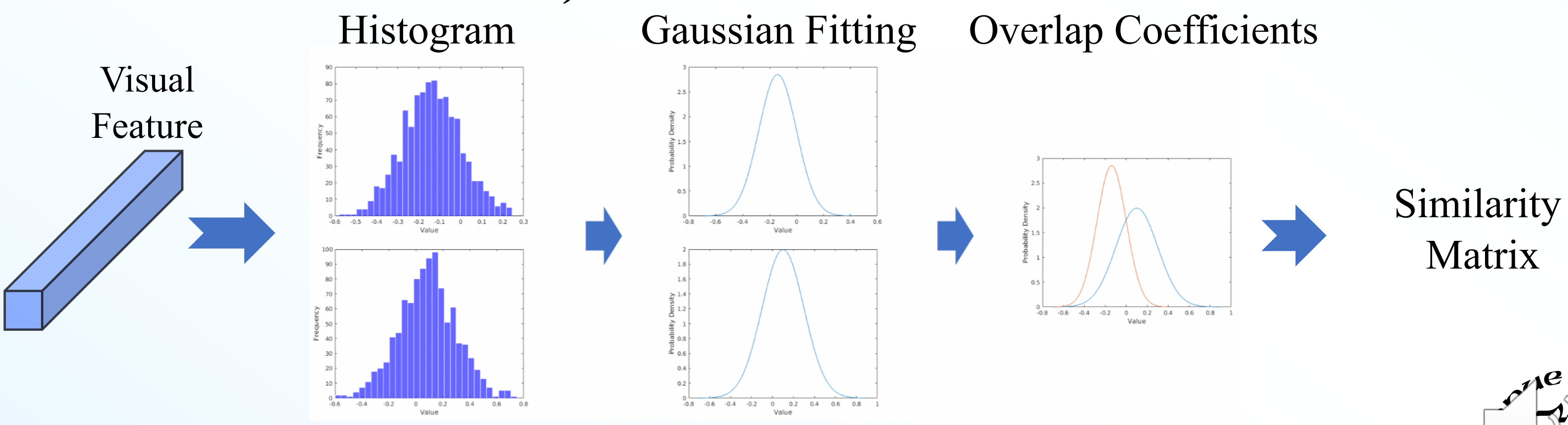


Visual Feature Similarity

- **Flat train the CNN model on the food image dataset**



- **Feature extraction layer outputs visual feature (1x1024 for DenseNet-121)**



Food Clustering and Hierarchical Structure

- **Affinity Propagation (AP) for clustering**
  - Based on similarity matrix
  - No need to estimate the number of clusters
- **Two matrices are used to propagate the information**
  - Responsibility matrix (r)
  - Availability matrix (a)

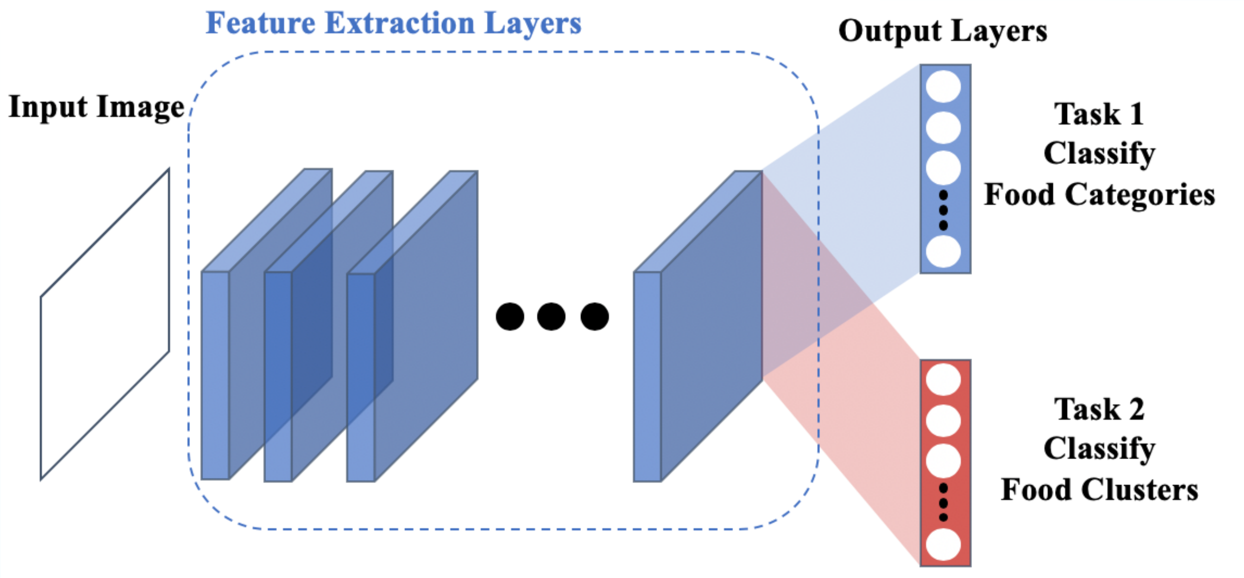
$$r(i,k) \leftarrow s(i,k) - \max_{k' \neq k} \{a(i,k') + s(i,k')\}$$

$$a(i,k) \leftarrow \min\left(0, r(k,k) + \sum_{i' \notin \{i,k\}} \max(0, r(i',k))\right) \text{ for } i \neq k \text{ and}$$

$$a(k,k) \leftarrow \sum_{i' \neq k} \max(0, r(i',k)).$$

Multi-Task Model

- **Multi-task model is used to embed the hierarchical structure**



- **Multi-task loss:**

$$L(\mathbf{w}) = \sum_{t=1}^{T} \lambda_t \sum_{i=1}^{N_t} -logp(y_i^{(t)}|\mathbf{x}_i, \mathbf{w}_0, \mathbf{w}^{(t)})$$

VIPER FoodNet (VFN) Dataset

- **Data-driven method highly depends on the quantity and quality of data**
- **VFN dataset contains 82 most frequently consumed food categories from What We Eat In America food category classification [1]**
- **Images are collected from public online sources with contextual information and close to real-life scenario**
  - ***e.g.*, fries and hamburger are typically consumed together**
- **VFN has 14,991 food images with 22,423 bounding boxes**

[1] H. Eicher-Miller and C.J. Boushey, "How Often and How Much? Differences in Dietary Intake by Frequency and Energy Contribution Vary among U.S. Adults in NHANES 2007–2012," Nutrients, vol. 9, no. 1, pp. 86, Jan 2017.

VIPER FoodNet (VFN) Dataset

- **Compared to other food image datasets for recognition**
  - Free-living for unconstrained image capturing environment
  - Controlled settings for fixed lighting conditions, dinnerware such as plates, glasses and silverwares

	UNIMIB2015	UNIMIB2016	UEC-100	UEC-256	VFN
Category	15	73	100	256	82
Image	2,000	1,027	12,740	28,897	14,991
% of Multi-food	100%	100%	9.2%	6.4%	26.1%
Study Type	Controlled	Controlled	Free-living	Free-living	Free-living

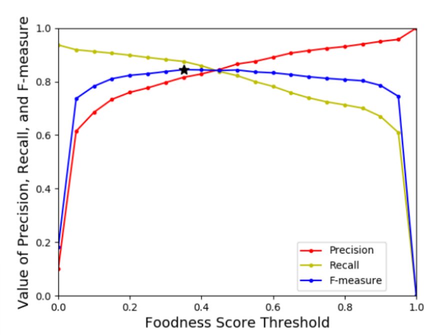
Experiments - Datasets

- **Our method is validated on 4 public datasets and our VFN dataset**
  - ETHZ-101 and UPMC-101 do not have bounding box information
  - UEC-100, UEC-256 focus on Japanese and Chinese food, and provide bounding box annotation
  - VFN contains American foods and provides bounding box annotation

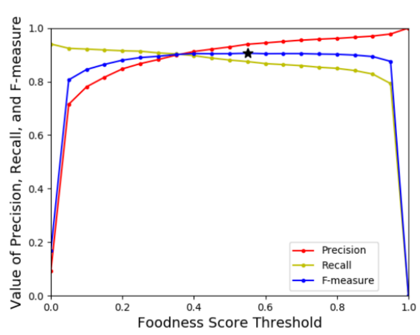
	ETHZ-101	**UPMC-101**	**UEC-100**	**UEC-256**	**VFN**
Category	101	101	100	256	82
Image	101,000	90,840	12,740	28,897	14,991
Bounding Box	--	--	14,361	31,395	22,423
Multi-food image	--	--	1,175	1,854	3,915
Portion of Multi-food image	--	--	9.2%	6.4%	26.1%

Experiments – Food Localization

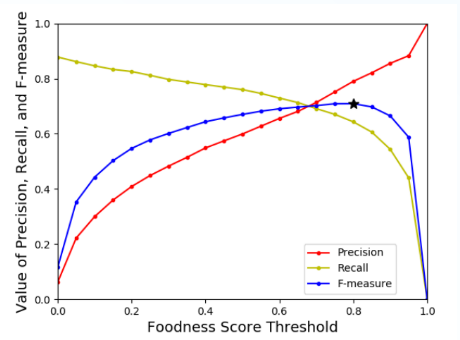
- **Train Faster RCNN on each dataset and select the highest confidence score on validation sets**



UEC-100



UEC-256



VFN

- **Evaluate our model on test sets**

	UEC-100	UEC-256	VFN
Confidence Threshold	0.35	0.55	0.80
Precision	0.8159	0.9388	0.6372
Recall	0.8376	0.9065	0.7064

Experiments – Food Classification

- **Automatically clustering food categories for different food image datasets and build two-level hierarchy**

	ETHZ-101	UPMC-101	UEC-100	UEC-256	VFN
# of Category	101	101	100	256	82
# of Cluster	17	18	15	33	14

- **Top-1 accuracy for category and cluster**

	Top-1 (Flat)	Top-1 (Hierarchical)	Cluster Top-1 (Flat)	Cluster Top-1 (Hierarchical)
ETHZ-101	75.31%	79.78%	85.06%	87.82%
UPMC-101	64.83%	69.26%	74.26%	78.73%
UEC-100	78.23%	80.81%	88.95%	90.20%
UEC-256	67.08%	72.36%	78.39%	83.37%
VFN	65.20%	71.81%	79.86%	84.81%

Experiments - Food Recognition

- **Combined food localization and classification and tested on three datasets that have bounding boxes**
- **Precision, recall, accuracy and mean Average Precision (mAP) are used to evaluate the recognition performance**

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \qquad Accuracy = \frac{TP}{TP + FP + FN}$$

	Precision	Recall	Accuracy	mAP
UEC-100	63.09%	66.54%	47.90%	60.63%
UEC-256	65.60%	61.24%	46.35%	56.73%
VFN	56.55%	45.46%	33.69%	40.06%

Visual Aware Hierarchy Based Food Recognition

Runyu Mao, Jiangpeng He, Zeman Shao,
Sri Kalyan Yarlagadda, and Fengqing Zhu

Video and Image Processing Laboratory
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana, U.S.A.