MADiMa2016

2nd International Workshop on Multimedia Assisted Dietary Management @ACM MM 2016

Michele Merler, Hui Wu, Rosario Uceda-Sosa
Quoc-Bao Nguyen, John R. Smith

IBM TJ Watson Research Center
mimerler@us.ibm.com

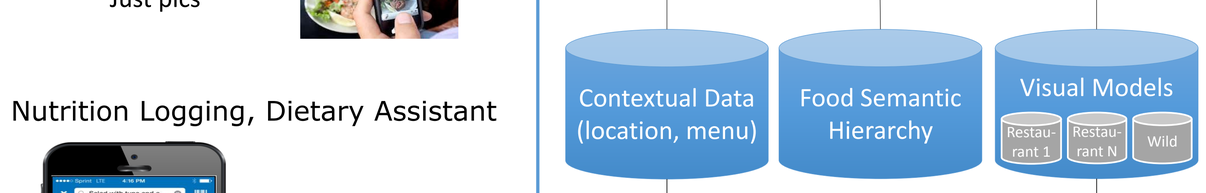
IBM

snap *eat* *repEat*

a Food Recognition Engine for Dietary Logging

Motivation

- **Exercise**, **sleep** and **nutrition** monitoring is essential for optimizing athletic **performance**
- Need to reduce friction (**manual**, **inaccurate**) to make nutrition monitoring fast and easy
- **Visual food recognition** greatly simplifies logging of meals using **context** and **content**
- Provides **accurate tracking of diet** and planning nutritional intake for achieving goals

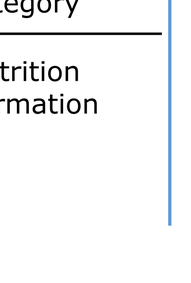


History | Logging | Planning

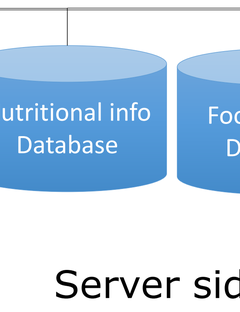
Watson Vision

Context:
- Geo-Location
- Time of day
- Restaurant name
- Historical meals

Content:
- Photo
- Text
- Interaction



Unknown Photo → Food Match & Nutrition Info ←



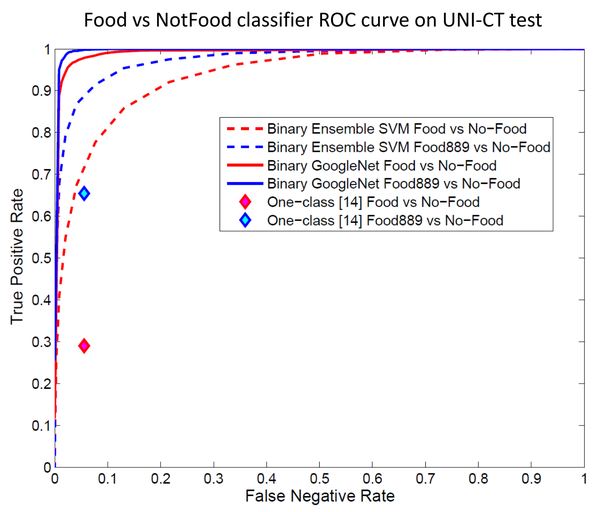
Food matching:
- Fast, accurate
- Multi-modal
- Scalable

Food database:
- Food photos
- Nutrition info
- Menus
- User data

Visual Recognition – Food vs Not-Food

Model: GoogleNet pretrained on Imagenet and finetuned on a dataset with 2.M training and 660K test images

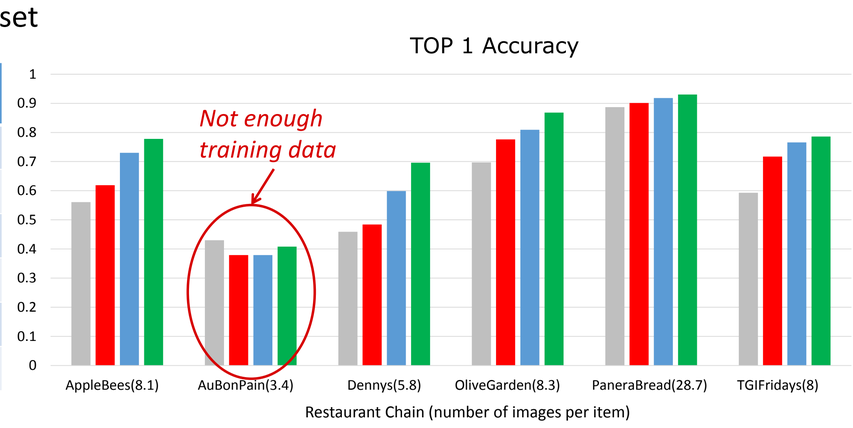
Method	One-Class SVM [Farinella et al. MaDiMa15]	Binary Ensemble SVM	Binary Fine-Tuned GoogleNet
Food889 True Positives Rate	0.6543	0.8685	**0.9711**
Flickr Food True Positives Rate	0.4300	0.6744	**0.9417**
Flickr No-Food True Negative Rate	0.9444	0.9589	**0.9817**
Overall Accuracy	0.9202	0.9513	**0.9808**
660K Test Set	-	0.8877	**0.9895**



Food Recognition in Context

- **K-NN**: based on fc7 features from AlexNet
- **AlexNet**: finetuned on restaurant chain training set
- **GoogLeNet**: finetuned on Restaurant chains training set
- **GoogLeNet$_{Food}$**: two finetuning steps, first n subset of Food vs Not-food dataset, then Restaurant chains training set

Restaurant	# Classes	# Images
Applebee's	50	405
Au Bon Pain	43	146
Denny's	56	325
Olive Garden	55	457
Panera Bread	79	2,267
TGI Fridays	54	432

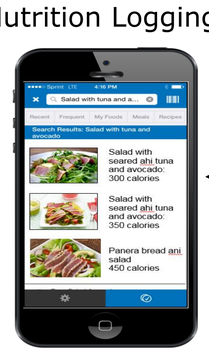


System Architecture

Snap Meal Photos

1. **In Context** pics, restaurant, menu
2. **In-the-wild** Just pics

Nutrition Logging, Dietary Assistant



Recognized food category

Nutrition information

Client side

REST API

Food Visual Recognition and Analysis

Contextual Data (location, menu)

Food Semantic Hierarchy

Visual Models: Restau-rant 1, Restau-rant N, Wild

Nutritional info Database

Food Images Database

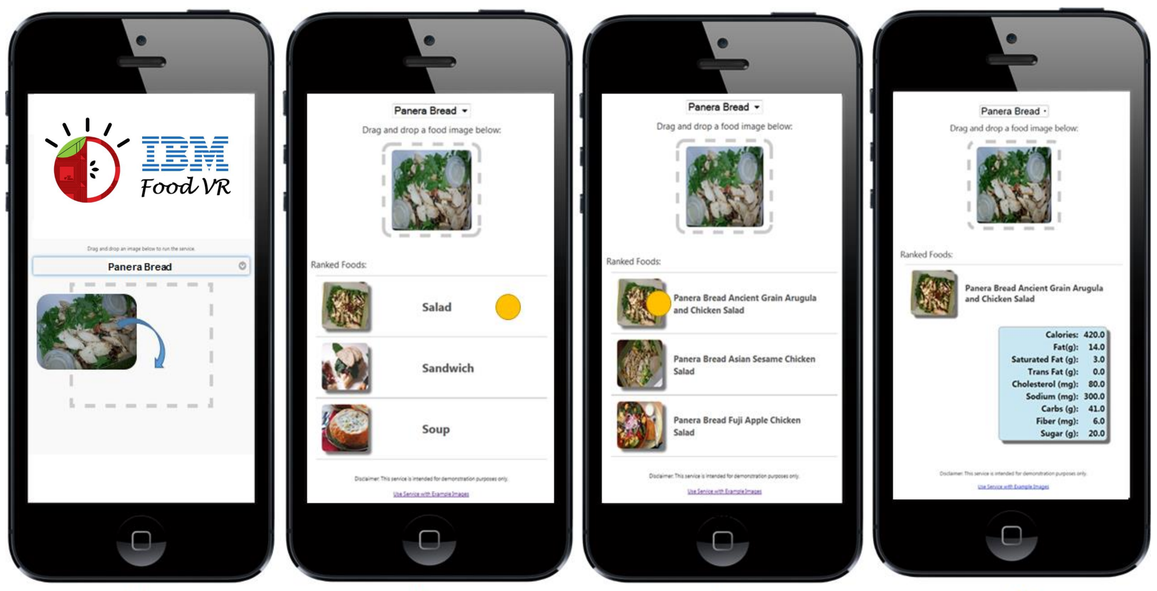
Server side

Food Recognition "in the wild"

Web and Social Media Crawling → Unnecessary images removal
- Duplicates
- Empty images
- Small images

→ Filter and rank by classifier (Food vs. not Food) → Crowdsourced human verifications

Visual Interface



Created Largest Visual Food Recognition Dataset

	Dataset	Number of Classes	Number of Images/Class	Number of Images	Food Ontology
IBM	**UEC Food 256 [22]**	256	89	31,651	None
IBM	**Geolocalized [40]**	3,852	30	117,504	None
NOT-IBM	**Food-101 [7]**	101	1000	101,100	None
NOT-IBM	**ETHZ Food 101 [37]**	101	1000	101,100	None
IBM	**Food 500**	**508**	**290**	**148,408**	**Yes**
IBM	**Food 3,000 (ongoing)**	3000	500	1.5M	Yes

Model: GoogleNet pretrained on Imagenet and finetuned on given dataset

Dataset	Accuracy (top 1)
Food 101 [Martinel ICCV15]	79
Food 101 (ours)	69.64
Food 500 (ours)	40.37

Most Confused Categories