

# RECIPE RECOGNITION WITH LARGE MULTIMODAL FOOD DATASET

Xin WANG <sup>(1)</sup>, Devinder Kumar <sup>(1)</sup>, Nicolas Thome <sup>(1)</sup>,  
Matthieu Cord <sup>(1)</sup>, Frédéric Precioso <sup>(2)</sup>

(1) Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6

(2) Universités Nice Sophia Antipolis, UMR 7271, I3S, France

CEA workshop, ICME 2015



# Outline

- 1 Context
- 2 New Dataset: UPMC Food-101
- 3 Experiments
- 4 Conclusions & Perspectives

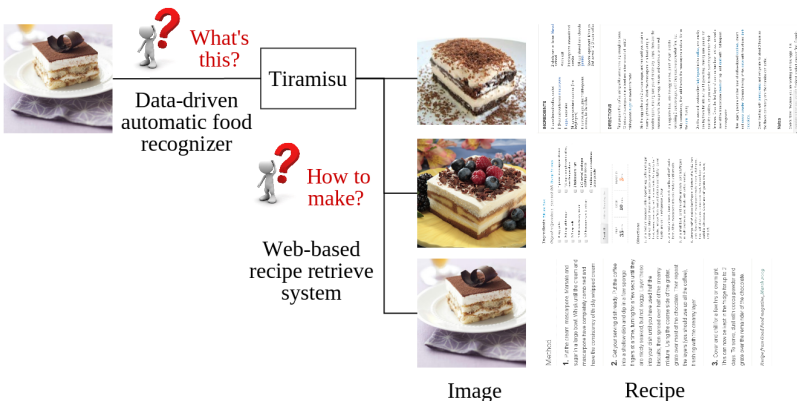
# Outline

- 1 Context
- 2 New Dataset: UPMC Food-101
- 3 Experiments
- 4 Conclusions & Perspectives

# Problem

## Problem

Recipe recognition: food description  $\Rightarrow$  food recipe



**Ingredients**

- 1 1/2 cups mascarpone cheese
- 1/2 cup heavy cream
- 1/2 cup sugar
- 1/2 cup coffee
- 1/2 cup cocoa powder
- 1/2 cup gelatin
- 1/2 cup milk
- 1/2 cup chocolate shavings
- 1/2 cup whipped cream
- 1/2 cup vanilla extract
- 1/2 cup almond extract
- 1/2 cup hazelnut extract
- 1/2 cup orange extract
- 1/2 cup lemon extract
- 1/2 cup lime extract
- 1/2 cup raspberry extract
- 1/2 cup strawberry extract
- 1/2 cup blueberry extract
- 1/2 cup blackberry extract
- 1/2 cup raspberry jam
- 1/2 cup strawberry jam
- 1/2 cup blueberry jam
- 1/2 cup blackberry jam
- 1/2 cup raspberry sauce
- 1/2 cup strawberry sauce
- 1/2 cup blueberry sauce
- 1/2 cup blackberry sauce
- 1/2 cup raspberry glaze
- 1/2 cup strawberry glaze
- 1/2 cup blueberry glaze
- 1/2 cup blackberry glaze
- 1/2 cup raspberry frosting
- 1/2 cup strawberry frosting
- 1/2 cup blueberry frosting
- 1/2 cup blackberry frosting
- 1/2 cup raspberry filling
- 1/2 cup strawberry filling
- 1/2 cup blueberry filling
- 1/2 cup blackberry filling
- 1/2 cup raspberry compote
- 1/2 cup strawberry compote
- 1/2 cup blueberry compote
- 1/2 cup blackberry compote
- 1/2 cup raspberry sauce
- 1/2 cup strawberry sauce
- 1/2 cup blueberry sauce
- 1/2 cup blackberry sauce
- 1/2 cup raspberry glaze
- 1/2 cup strawberry glaze
- 1/2 cup blueberry glaze
- 1/2 cup blackberry glaze
- 1/2 cup raspberry frosting
- 1/2 cup strawberry frosting
- 1/2 cup blueberry frosting
- 1/2 cup blackberry frosting
- 1/2 cup raspberry filling
- 1/2 cup strawberry filling
- 1/2 cup blueberry filling
- 1/2 cup blackberry filling
- 1/2 cup raspberry compote
- 1/2 cup strawberry compote
- 1/2 cup blueberry compote
- 1/2 cup blackberry compote

**Instructions**

1. In a large bowl, whisk together the mascarpone cheese, heavy cream, sugar, and coffee. Add the cocoa powder and gelatin, and whisk until smooth.
2. In a separate bowl, whisk together the milk and chocolate shavings. Pour this mixture into the mascarpone mixture and whisk until well combined.
3. In a third bowl, whisk together the vanilla extract, almond extract, hazelnut extract, orange extract, lemon extract, lime extract, raspberry extract, strawberry extract, blueberry extract, blackberry extract, raspberry jam, strawberry jam, blueberry jam, blackberry jam, raspberry sauce, strawberry sauce, blueberry sauce, blackberry sauce, raspberry glaze, strawberry glaze, blueberry glaze, blackberry glaze, raspberry frosting, strawberry frosting, blueberry frosting, blackberry frosting, raspberry filling, strawberry filling, blueberry filling, blackberry filling, raspberry compote, strawberry compote, blueberry compote, blackberry compote, raspberry sauce, strawberry sauce, blueberry sauce, blackberry sauce, raspberry glaze, strawberry glaze, blueberry glaze, blackberry glaze, raspberry frosting, strawberry frosting, blueberry frosting, blackberry frosting, raspberry filling, strawberry filling, blueberry filling, blackberry filling, raspberry compote, strawberry compote, blueberry compote, blackberry compote.

# Problem

## Core Problem

Food category classification: key technology for many food-related applications such as:

- 1 monitoring healthy diet<sup>a b</sup>
- 2 recording eating activities<sup>c</sup>
- 3 food recommendation system<sup>d</sup>

---

<sup>a</sup> Aizawa, K., Maruyama, Y., Li, H., and Morikawa, C. (2013). Food balance estimation by using personal dietary tendencies in a multimedia food log. *IEEE Transactions on Multimedia*,

<sup>b</sup> Khanna, N. and et al. (2010). An overview of the technology assisted dietary assessment project at purdue university.

<sup>c</sup> Aizawa, K. and et al. (2014). Comparative Study of the Routine Daily Usability of FoodLog: A Smartphone-based Food Recording Tool Assisted by Image Retrieval. *Journal of diabetes science and technology*.

<sup>d</sup> Takuma Maruyama Yoshiyuki Kawano Keiji Yanai, Real-time Mobile Recipe Recommendation System Using Food Ingredient Recognition, IMMPD'12

# Motivation

## Motivation 1

- 1 Needs of large scale dataset:
  - Pittsburgh Food Image Dataset (PFID): 4556 images
  - UNICT-FD889 dataset: 889 images
  - UEC-Food100 (without extension): 100 categories, 100 images / category
- 2 Needs of multi-modal food dataset

# Motivation

## Motivation 1

- 1 Needs of large scale dataset:
  - Pittsburgh Food Image Dataset (PFID): 4556 images
  - UNICT-FD889 dataset: 889 images
  - UEC-Food100 (without extension): 100 categories, 100 images / category
- 2 Needs of **multi-modal** food dataset

# Motivation

## Motivation 1: Large scale multi-modal dataset

Training a generic multi-modal supervised automated recipe recognition system needs:

- 1 Large number of food categories
- 2 Large number of food examples
- 3 Food image + text

## Motivation 2: Twin datasets

Same categories, different data sources:

- ETHZ Food-101 [ECCV 2014]: Collected from only gourmet sites
- UPMC Food-101: Collected from Google Images Search engine



# Motivation

## Motivation 1: Large scale multi-modal dataset

Training a generic multi-modal supervised automated recipe recognition system needs:

- 1 Large number of food categories
- 2 Large number of food examples
- 3 Food image + text

## Motivation 2: Twin datasets

Same categories, different data sources:

- ETHZ Food-101 [ECCV 2014]: Collected from only gourmet sites
- UPMC Food-101: Collected from Google Images Search engine

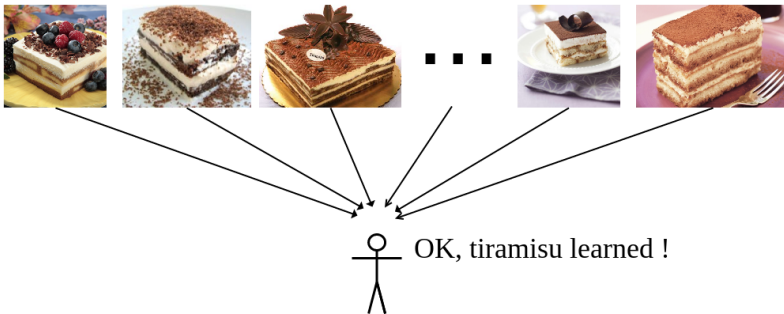
# Outline

- 1 Context
- 2 New Dataset: UPMC Food-101**
- 3 Experiments
- 4 Conclusions & Perspectives

# Properties

- 1 Large scale: large number ( $\sim 101,000$ ) of training images

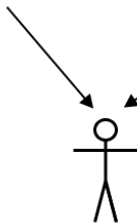
$\sim 1000$  tiramisus for learning



# Properties

- 1 Large scale
- 2 Multimodal dataset: visual information + textual information

**Tiramisu** (from [Italian](#), spelled [tiramisù](#) [[tʃamiˈsu](#)], meaning "pick me up" or "lift me up") is a popular coffee-flavored [Italian](#) dessert. It is made of [ladyfingers](#) ([Italian: Savoiardi](#), [[savoˈardi](#)]) dipped in coffee, layered with a whipped mixture of eggs, sugar, and [mascarpone cheese](#), flavoured with [cocoa](#). The recipe has been adapted into many varieties of [cakes](#) and other [desserts](#).<sup>[1]</sup> Its origins are often disputed between Italian regions such as [Veneto](#), [Friuli Venezia Giulia](#), [Piedmont](#), and others.



I can read and see !

# Properties

- 1 Large scale
- 2 Multimodal dataset
- 3 Web-based source: Building UPMC Food-101 from web



# Dataset acquisition protocol

## Dataset acquisition protocol

- 1 Same category names as ETHZ Food-101
- 2 Suffix *recipe* after each category name
- 3 First 1000 results of Google Image search engine
- 4 Data cleaning

# Dataset acquisition protocol

## Dataset acquisition protocol

- 1 Same category names as ETHZ Food-101
- 2 Suffix *recipe* after each category name
- 3 First 1000 results of Google Image search engine
- 4 Data cleaning

# Dataset acquisition protocol

## Dataset acquisition protocol

- 1 Same category names as ETHZ Food-101
- 2 Suffix *recipe* after each category name
- 3 First 1000 results of Google Image search engine
- 4 Data cleaning



# Dataset acquisition protocol

## Dataset acquisition protocol

- 1 Same category names as ETHZ Food-101
- 2 Suffix *recipe* after each category name
- 3 First 1000 results of Google Image search engine
- 4 Data cleaning

# 100 category examples of UPMC Food-101 dataset



## Class wise performance: Best 5

					Spaghetti Carbonara 73 %
					Deviled Eggs 65 %
					Grilled cheese sandwich 59 %
					Prime Rib 57 %
					Chicken Wings 52 %

## Class wise performance: Worst 5



## Difficulties of dataset

- Food deformation
- Noisy image



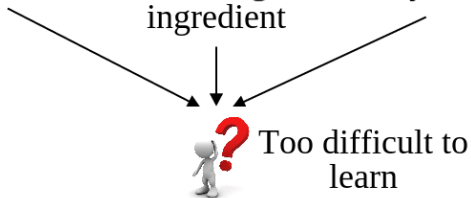
Hamburger



Hamburger  
ingredient



Noisy image



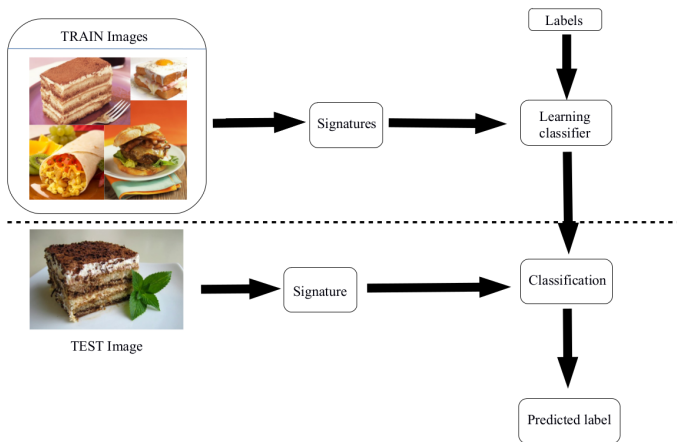
# Outline

- 1 Context
- 2 New Dataset: UPMC Food-101
- 3 Experiments**
- 4 Conclusions & Perspectives

# Supervised food category classification

## Goal

Predict food category by using the data (image / text)



# Visual features (1)

## Traditional visual features: Dense-SIFT + BoW Presentation

- 1 Dense-SIFT: Bag-of-Words histogram with a spatial pyramid
- 2 BossaNova<sup>a</sup>: Pooling by considering distance between a word and a given center of a cluster

<sup>a</sup>Avila, S., and Thome, N., Cord M., Valle E., Araujo A. Pooling in image representation: The visual codeword point of view. CVIU 2013

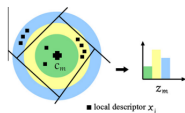


Figure: BossaNova pooling

BoW	BossaNova
23.96%	28.59%

Table: Avg. accuracy of BoW and BossaNova



## Visual features (2)

### Deep visual features

- 1 Deep CNN<sup>a</sup>: *fast network*, 9 layers CNN, 4096 dimensions.
- 2 Very Deep CNN<sup>b</sup>: 19 layers CNN, 4096 dimensions.

<sup>a</sup>Sermanet, P., Eigen, D., Zhang, X., and LeCun, Y. . Overfeat: Integrated recognition, localization and detection using convolutional networks, ICLR 2014.

<sup>b</sup>Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition, ICLR 2015.

Deep	Very Deep
33.91%	40.21%

**Table:** Avg. accuracy of deep and very deep features.

# Textual features

## Textual Features

- 1 TF-IDF : Term frequency - Inverse Document Frequency
- 2 Word2vec<sup>a</sup> : Embedded vector representations of words

<sup>a</sup>Mikolov, T., Sutskever, I., and Chen, K.. Distributed representations of words and phrases and their compositionality, NIPS 2013.

TF-IDF	word2vec
85.10%	67.21%

**Table:** Avg. accuracy of TF-IDF and word2vec

# Features and Performances

## Late fusion

Fusion score:

$$s_f = \alpha s_i + (1 - \alpha) s_t,$$

$\alpha$ : Fusion parameter, in the range  $[0, 1]$

$s_i$ : Classification score of the image classifier

$s_t$ : Classification score of the text classifier

Visual				Textual		Fusion
BoW	BossaNova	Deep	Very Deep	TF-IDF	word2vec	TF-IDF + Very Deep
23.96%	28.59%	33.91%	40.21%	82.06%	67.21%	85.10%

**Table:** Classification results (avg. accuracy %) on UPMC Food-101 for visual, textual features and fusion.

# Comparing UPMC Food-101 and ETHZ Food-101

train / test	UPMC	ETHZ
UPMC (600 examples)	40.56%	25.63%
ETHZ (600 examples)	25.28%	42.54%
UPMC (all examples)	-	24.06%
ETHZ (all examples)	24.92%	-

**Table:** Avg. accuracy of transfer learning (very deep features) between UPMC Food-101 and ETHZ Food-101.

## Analysis

- 1 Systematic loss:  $\sim 15\%$  when training on one dataset and testing on the other dataset.
- 2 Increasing training images does not achieve better results.

Different data sources

# Comparing UPMC Food-101 and ETHZ Food-101

train / test	UPMC	ETHZ
UPMC (600 examples)	40.56%	25.63%
ETHZ (600 examples)	25.28%	42.54%
UPMC (all examples)	-	24.06%
ETHZ (all examples)	24.92%	-

**Table:** Avg. accuracy of transfer learning (very deep features) between UPMC Food-101 and ETHZ Food-101.

## Analysis

- 1 Systematic loss:  $\sim 15\%$  when training on one dataset and testing on the other dataset.
- 2 Increasing training images does not achieve better results.

Different data sources

# Comparing UPMC Food-101 and ETHZ Food-101

train / test	UPMC	ETHZ
UPMC (600 examples)	40.56%	25.63%
ETHZ (600 examples)	25.28%	42.54%
UPMC (all examples)	-	24.06%
ETHZ (all examples)	24.92%	-

**Table:** Avg. accuracy of transfer learning (very deep features) between UPMC Food-101 and ETHZ Food-101.

## Analysis

- 1 Systematic loss:  $\sim 15\%$  when training on one dataset and testing on the other dataset.
- 2 Increasing training images does not achieve better results.

Different data sources

# Comparing UPMC Food-101 and ETHZ Food-101

train / test	UPMC	ETHZ
UPMC (600 examples)	40.56%	25.63%
ETHZ (600 examples)	25.28%	42.54%
UPMC (all examples)	-	24.06%
ETHZ (all examples)	24.92%	-

**Table:** Avg. accuracy of transfer learning (very deep features) between UPMC Food-101 and ETHZ Food-101.

## Analysis

- 1 Systematic loss:  $\sim 15\%$  when training on one dataset and testing on the other dataset.
- 2 Increasing training images does not achieve better results.

## Different data sources

# Word2vec: a powerful semantic tool

## Word2vec evaluation

rice	japan	rice japan
calros 0.59	osaka 0.70	<b>koshihikari 0.64</b>
basmati 0.59	tokyo 0.62	awabi 0.61
vermicelli 0.58	kyoto 0.62	japanes 0.61
stirfri 0.58	chugoku 0.61	nishiki 0.59
veget 0.58	gunma 0.60	chahan 0.57

**Table:** Short phrase **rice japan**, represented as the average of **rice** and **japan**, is closest to **koshihikari**, which is neither among the 101 categories, nor among the neighbors of **rice** or **japan**.




# Outline


- 1 Context
- 2 New Dataset: UPMC Food-101
- 3 Experiments
- 4 Conclusions & Perspectives**

# Recognizer based on UPMC Food-101


Demo (ongoing work)

Query:  New search

**pizza - PREDICTION SCORE: 4.2368**



**spaghetti\_carbonara - PREDICTION SCORE: -1.9803**



**guacamole - PREDICTION SCORE: -1.998**





Figure: Results for a pizza image

# Recognizer based on UPMC Food-101

Demo (ongoing work)

Query:  New search

**pizza - PREDICTION SCORE: 4.2368**

**spaghetti\_carbonara - PREDICTION SCORE: -1.9803**

**guacamole - PREDICTION SCORE: -1.998**

**Recipe card details:**

**Ingredients:**

- 1/2 cup (120 ml) milk
- 1/2 cup (120 ml) heavy cream
- 1/2 cup (120 ml) parmesan cheese
- 1/2 cup (120 ml) mozzarella cheese
- 1/2 cup (120 ml) ricotta cheese
- 1/2 cup (120 ml) butter
- 1/2 cup (120 ml) olive oil
- 1/2 cup (120 ml) salt
- 1/2 cup (120 ml) pepper
- 1/2 cup (120 ml) garlic
- 1/2 cup (120 ml) onion
- 1/2 cup (120 ml) mushrooms
- 1/2 cup (120 ml) tomatoes
- 1/2 cup (120 ml) basil
- 1/2 cup (120 ml) oregano
- 1/2 cup (120 ml) thyme
- 1/2 cup (120 ml) rosemary
- 1/2 cup (120 ml) sage
- 1/2 cup (120 ml) dill
- 1/2 cup (120 ml) chives
- 1/2 cup (120 ml) parsley
- 1/2 cup (120 ml) cilantro
- 1/2 cup (120 ml) mint
- 1/2 cup (120 ml) basil
- 1/2 cup (120 ml) oregano
- 1/2 cup (120 ml) thyme
- 1/2 cup (120 ml) rosemary
- 1/2 cup (120 ml) sage
- 1/2 cup (120 ml) dill
- 1/2 cup (120 ml) chives
- 1/2 cup (120 ml) parsley
- 1/2 cup (120 ml) cilantro
- 1/2 cup (120 ml) mint

**Instructions:**

1. Preheat oven to 375°F (190°C).
2. In a large bowl, combine the milk, heavy cream, parmesan, mozzarella, and ricotta. Stir until well combined.
3. In a separate bowl, combine the butter, olive oil, salt, pepper, garlic, onion, mushrooms, tomatoes, basil, oregano, thyme, rosemary, sage, dill, chives, parsley, cilantro, and mint.
4. Pour the sauce over the pizza and bake for 15-20 minutes.
5. Serve hot.

Figure: Results for a pizza image

# Conclusion & Perspective

## Conclusions

- 1 UPMC Food-101: a large scale multimodal food recipe dataset
- 2 Detailed classification experiments
- 3 Semantic vectorial text representation tool word2vec
- 4 Recipe retrieval system prototype

## Perspectives

- Active learning to improve the retrieval system based on user interaction
- Levaraging twin datasets to achieve better results

# Conclusion & Perspective

## Conclusions

- 1 UPMC Food-101: a large scale multimodal food recipe dataset
- 2 Detailed classification experiments
- 3 Semantic vectorial text representation tool word2vec
- 4 Recipe retrieval system prototype

## Perspectives

- Active learning to improve the retrieval system based on user interaction
- Leveraging twin datasets to achieve better results

# Conclusion & Perspective

## Conclusions

- 1 UPMC Food-101: a large scale multimodal food recipe dataset
- 2 Detailed classification experiments
- 3 Semantic vectorial text representation tool word2vec
- 4 Recipe retrieval system prototype

## Perspectives

- Active learning to improve the retrieval system based on user interaction
- Leveraging twin datasets to achieve better results

# Conclusion & Perspective

## Conclusions

- 1 UPMC Food-101: a large scale multimodal food recipe dataset
- 2 Detailed classification experiments
- 3 Semantic vectorial text representation tool word2vec
- 4 Recipe retrieval system prototype

## Perspectives

- Active learning to improve the retrieval system based on user interaction
- Leveraging twin datasets to achieve better results

# Conclusion & Perspective

## Conclusions

- 1 UPMC Food-101: a large scale multimodal food recipe dataset
- 2 Detailed classification experiments
- 3 Semantic vectorial text representation tool word2vec
- 4 Recipe retrieval system prototype

## Perspectives

- Active learning to improve the retrieval system based on user interaction
- Leveraging twin datasets to achieve better results



# Conclusion & Perspective

## Conclusions

- 1 UPMC Food-101: a large scale multimodal food recipe dataset
- 2 Detailed classification experiments
- 3 Semantic vectorial text representation tool word2vec
- 4 Recipe retrieval system prototype

## Perspectives

- Active learning to improve the retrieval system based on user interaction
- Leveraging twin datasets to achieve better results

# Thank you for your attention!

## Questions?

<u>Xin Wang</u> <sup>(1)</sup>	xin.wang@lip6.fr
Devinder Kumar <sup>(1)</sup>	devinder.kumar@uwaterloo.ca
Nicolas Thome <sup>(1)</sup>	nicolas.thome@lip6.fr
Matthieu Cord <sup>(1)</sup>	matthieu.cord@lip6.fr
Frédéric Precioso <sup>(2)</sup>	frederic.precioso@polytech.unice.fr

(1) Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6

(2) Universités Nice Sophia Antipolis, UMR 7271, I3S, France

## Dataset available on demand



This work is supported by The French National Research Agency