Recipe Recognition with Large Multimodal Food Dataset

Xin WANG (1), Devinder Kumar (1), Nicolas Thome (1), Matthieu Cord (1), Frédéric Precioso (2)

(1) Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6
(2) Universités Nice Sophia Antipolis, UMR 7271, I3S, France

CEA workshop, ICME 2015

Outline

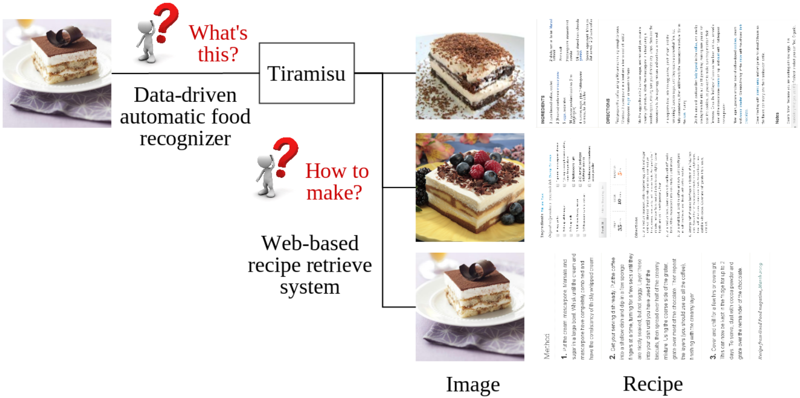
1. Context
2. New Dataset: UPMC Food-101
3. Experiments
4. Conclusions & Perspectives

Outline

Problem

Problem

Recipe recognition: food description $\Rightarrow$ food recipe



Problem

Core Problem

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

1. monitoring healthy diet[a] [b]
2. recording eating activities[c]
3. food recommendation system[d]

---

[a] 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,

[b] Khanna, N. and et al. (2010). An overview of the technology assisted dietary assessment project at purdue university.

[c] 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.

[d] 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

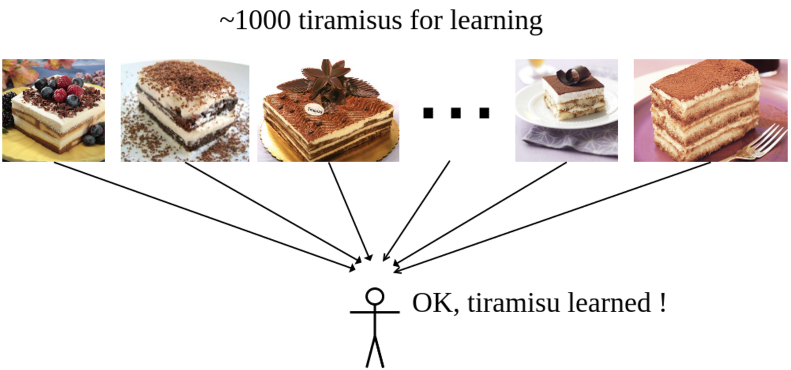
Same categories, different data sources:

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

Outline

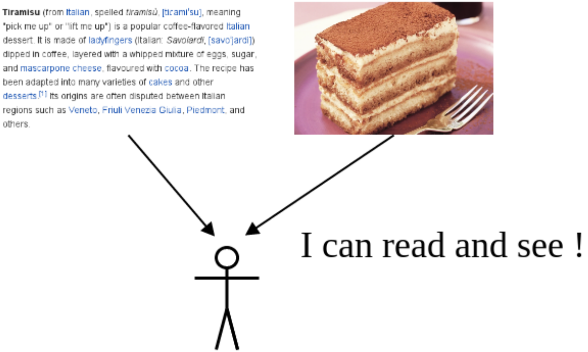
Properties

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



Properties

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



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

Ravioli
8.5 %

Gnocchi
~12.5 %

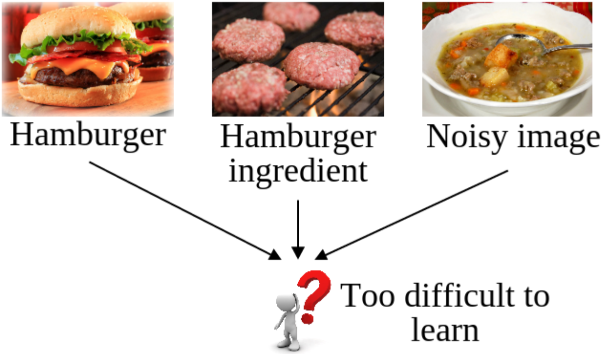
Hamburger
~12.5 %

Ice-cream
~12.5 %

Dumplings
13 %

Difficulties of dataset

- Food deformation
- Noisy image

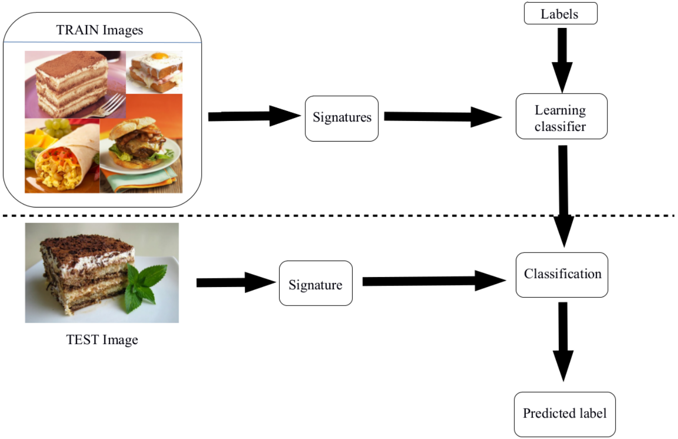


Outline

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[a]: Pooling by considering distance between a word and a given center of a cluster

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

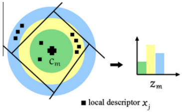


Figure: BossaNova pooling

BoW	BossaNova
23.96%	28.59%

Table: Avg. accuracy of BoW and BossaNova

Visual features (2)

Deep visual features

1. Deep CNN[a]: *fast network*, 9 layers CNN, 4096 dimensions.
2. Very Deep CNN[b]: 19 layers CNN, 4096 dimensions.

[a] Sermanet, P., Eigen, D., Zhang, X., and LeCun, Y. . Overfeat: Integrated recognition, localization and detection using convolutional networks, ICLR 2014.

[b] 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[a] : Embedded vector representations of words

[a] 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:

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

α: 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	**koshihikari 0.64**
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

Recognizer based on UPMC Food-101

Demo (ongoing work)

Figure: Results for a pizza image

Recognizer based on UPMC Food-101

Demo (ongoing work)

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 retrival 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 retrival 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 retrival 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 retrival 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 retrival 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 retrival system based on user interaction
- Levaraging twin datasets to achieve better results

Thank you for your attention! Questions?

Xin Wang(1)	xin.wang@lip6.fr
Devinder Kumar (1)	devinder.kumar@uwaterloo.ca
Nicolas Thome(1)	nicolas.thome@lip6.fr
Matthieu Cord(1)	matthieu.cord@lip6.fr
Frédéric Precioso (2)	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