Performance Evaluation Methods of Computer Vision Systems for Meal Assessment

Marios Anthimopoulos[1,2], Joachim Dehais[1], Stavroula Mougiakakou[1,3]

[1] ARTORG Center for Biomedical Engineering Research, University of Bern, Switzerland
[2] Department of Emergency Medicine, Bern University Hospital, Switzerland
[3] Department of Endocrinology, Diabetes and Clinical Nutrition, Bern University Hospital, Switzerland

DIABETES TECHNOLOGY RESEARCH GROUP | UNIVERSITÄT BERN | ARTORG CENTER BIOMEDICAL ENGINEERING RESEARCH

Introduction

Diet management and food intake assessment are key factors to treat and prevent diet and lifestyle related chronic diseases. To this end, mobile solutions have been developed that use images of a meal and calculate its nutritional information based on computer vision (Fig. 1). However, most of these systems use different datasets and protocols for evaluation. In this study, we present an overview of the available evaluation methods and datasets for the major computer vision stages of meal assessment systems.

Methods

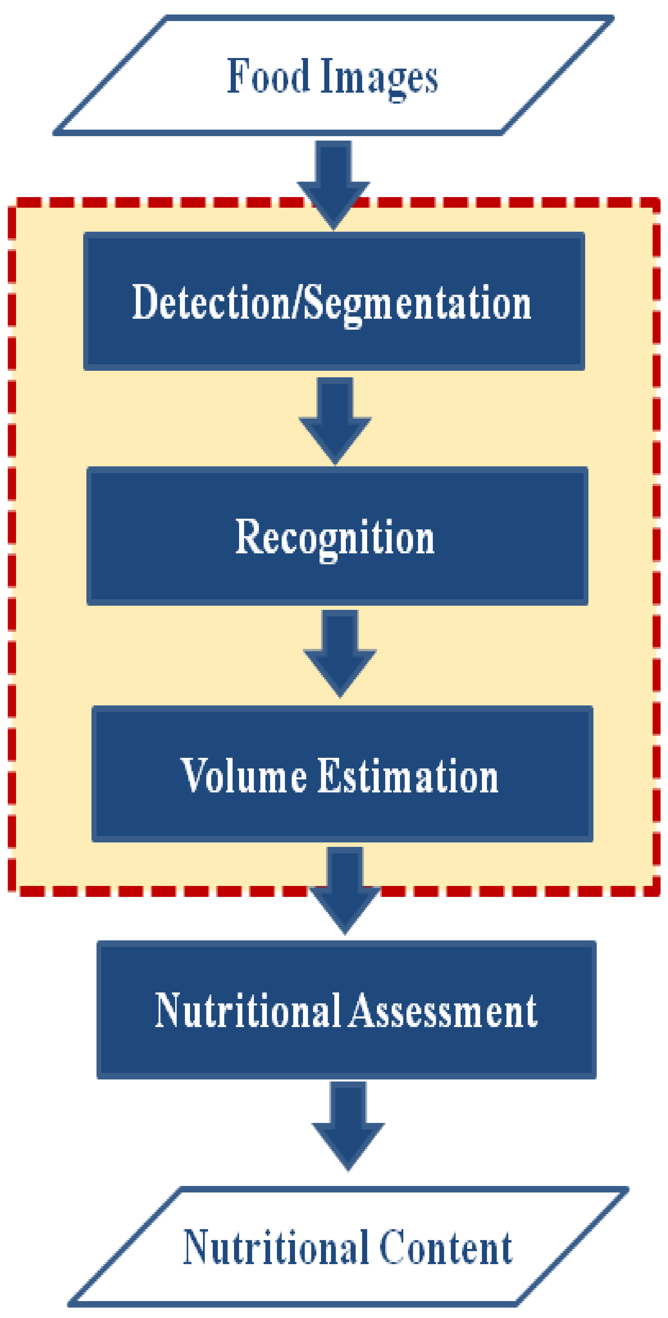

Figure 1 – Main stages of a computer vision based meal assessment system

Detection: Segmenting food from non-food regions is equivalent to food detection. Information retrieval (IR) measures have been used:

False Detection Rate $= \frac{N_{FP}+N_{FN}}{N_{FP}+N_{FN}+N_{TP}+N_{TN}}$, Recall $= \frac{N_{TP}}{N_{TP}+N_{FN}}$, Precision $= \frac{N_{TP}}{N_{TP}+N_{FP}}$

N_{TP}: #pixels correctly retrieved as food, N_{FP}: #pixels incorrectly retrieved as food, N_{TN}: correctly retrieved as non-food and, N_{FN}: incorrectly retrieved as non-food.

Segmentation: Segmenting multiple foods and background

1. IR measures on the retrieved boundary/non-boundary pixels
Drawback: Similarity of boundary pixels does not directly link to the generated segments.

2. Indirect evaluation through recognition
Drawbacks: (i) bad results may be due to unsuccessful recognition on correct segments, (ii) one-to-one match between the recognized and the reference labels is not always possible.

3. Region overlap metrics between result $R = \{R_i\}_{i=1}^{m}$ and ground truth ($T = \{T_i\}_{i=1}^{n}$):
$I(T \Rightarrow S) = \frac{\sum_i \mathrm{Max}_j(|S_i \cap T_j|)}{\sum_i |S_i|}$ and $I(S \Rightarrow T) = \frac{\sum_i \mathrm{Max}_j(|T_i \cap S_j|)}{\sum_i |T_i|}$ combined in some mean

Recognition: Food image/object classification → Typical IR problem

Evaluation frameworks:
- Train-validation-test scheme: does not exploit the entire dataset
- K-fold Cross-validation: computationally expensive

Evaluation metrics:
- Accuracy: ignores imbalances among classes and precision/recall
- Average per-class accuracy: ignores imbalances among precision/recall
- Average f-score over classes: balanced over classes and precision/recall

Table 1 – Food recognition datasets

Database	Total Size	Classes	Annotation
PFID	2424	101	Recognition map
Food-101	101000	101	labels
UNICT-FD889	3583	889	labels
UEC FOOD 100	9060	100	Bounding box and labels
UEC FOOD 256	31651	256	Bounding box and labels
Geolocalized	117504	-	Location and labels

Volume estimation: The comparison between measured (V) and true value (V^{true}) can be signed ($Diff = V - V^{true}$), absolute ($AbsDiff = |Diff|$), relative ($RelDiff = \frac{Diff}{V^{true}}$), or both absolute and relative ($RelAbsDiff = |RelDiff|$)
Statistics over different measured items are used as overall metrics:

Trueness $= \overline{Diff}$: Measures consistent over- or underestimations → Easy to bring to zero

Precision $= \sqrt{\overline{(Diff - Trueness)^2}}$: Measures the statistical variability of the errors → Has to be combined with trueness

Error $= \overline{AbsDiff}$: Measures the expected absolute error → Can be used alone

Conclusions

Designing a proper performance evaluation framework is a crucial task that formally defines the problem, and enables the comparison among solutions, as well as the optimization of each system. Establishing public datasets and standardized evaluation protocols in the new field of automatic meal assessment will strengthen the community and accelerate its progress.

Acknowledgements

The work was funded in part by the Bern University Hospital "Inselspital" and the European Union Seventh Framework Programme (FP7-PEOPLE-2011-IAPP) under grant agreement n° 286408 [www.gocarb.eu].

INSELSPITAL
UNIVERSITÄTSSPITAL BERN
HOPITAL UNIVERSITAIRE DE BERNE
BERN UNIVERSITY HOSPITAL