RWMF: A Real-World Multimodal Foodlog Database

Pengfei Zhou†, Cong Bai*†, Kaining Ying†, Jie Xia‡ and Lixin Huang†
†College of Computer Science and Technology
Zhejiang University of Technology, Hangzhou, China
Email: {pengfeizhou, congbai}@zjut.edu.cn
‡College of Information Engineering
Zhejiang University of Technology, Hangzhou, China
Email: jiexia@zjut.edu.cn

***Abstract*—With the increasing health concerns on diet, it's worthwhile to develop an intelligent assistant that can help users eat healthier. Such assistants can automatically give personal advice for the users' diet and generate health reports about eating on a regular basis. To boost the research on such diet assistant, we establish a real-world foodlog database using various methods such as filter, cluster and graph convolutional network. This database is built based on real-world lifelog and medical data, which is named as Real-World Multimodal Foodlog (RWMF). It contains 7500 multimodal pairs, and each pair consists of a food image paired with a line of personal biometrics data (such as Blood Glucose) and a textual food description of food composition paired with a line of food nutrition data. In this paper, we present the detailed procedures for setting up the database. We evaluate the performance of RWMF using different food classification and cross-modal retrieval approaches. We also test the performance of multimodal fusion on RWMF through ablation experiments. The experimental results show that the RWMF database is quite challenging and can be widely used to evaluate the performance of food analysis methods based on multimodal data.**

I. Introduction

Digital life is becoming a general trend of modern lifestyle. With the development of wearable devices, more and more personal data is generated in the interaction between digital devices and people. Individuals are now able to use these interactive data to direct their daily activities, such as eating and sleeping [1]. Such interactive data collected by various devices has been summarized as lifelogs [2], which generated in the interaction between individuals and smart devices. Lifelogs have been utilized to help personal lives in real-world scenarios [3]. And the research on lifelog has received increasing attention in recent years, with many applications related to human memory [4], health and wellness [5], and life assistant [6].

Among these implementations, foodlog is a food-oriented lifelog application related to personal health, which can be used for dietary management [7]. Foodlog records a user's food intake with the detail of diets (such as visual images, calories, time, location, etc.) [8]. As shown in Fig.1, the modalities of foodlog are abundant, which include food images and related biometrics data captured in people's daily life. So, the foodlog is an important resource to help develop the personal diet assitant. At present, the research on multimodal foodlog is still developing [9].

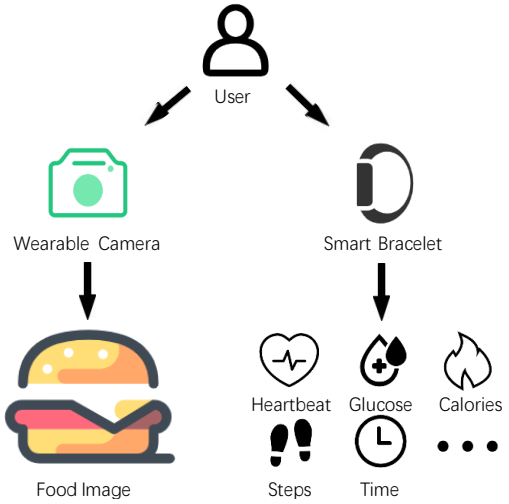

Fig. 1. Example of real-world foodlog forms.

Due to the fact that the modalities of foodlog data are various, a lot of researchers turn to cross-modal retrieval (CMR) to make better use of these data. The CMR is able to retrieve results bridging the semantic gaps between modalities [10,11]. Moreover, it can improve the efficiency of retrieval by using complementary data of different modalities. In recent years, cross-modal retrieval is becoming one hotspot in the field of foodlog research, which can be used for recipe retrieval as well as food analysis. For example, Salvador et al. [12] use cross-modal embeddings to retrieve recipes with food images. Jiang et al. [13] recognize food images using the multi-view deep feature. Min et al. [14] retrieve recipes of food with ingredients based on cross-region food analysis. Pouladzadeh et al. [15] build a monitoring system that can help users calculate their daily food energy intake.

However, because the current foodlog database such as Food-101 [20] barely includes enough real-world personal data, it is still difficult to utilize the foodlog to solve the

Manuscript 440 submitted to 2020 25th International Conference
on Pattern Recognition. Received April 15, 2020.

specific personal issues. For example, diabetic patients may find it difficult to plan and monitor their daily meals [16]. They need an intelligent diet assistant instead of keeping the healthy matters in mind by themselves [17]. Such an assistant is supposed to track the characters of food and give the health advice to users to help patients keep the balance between dietary health and eased life [18]. For example, the general system established by Tusor et al. [19] uses knowledge representation of numerical and fuzzy data to help individuals make their own dietary decisions.

So, further explorations on diet assistant, especially those based on cross-modal retrieval techniques, are limited by the insufficiency of the database. The previous food database did not include enough personal data to serve the specific user. Moreover, the data captured in the experimental environment are not very suitable for evaluating real-world issues. It is due to insufficient accuracy as well as missed vital information. For example, the current Food-101 dataset only has images and labels, whereas biological data of food nutrition is definitely needed for food analysis. As far as we know, no food database with such multimodal data has been publicly available yet. Therefore, an effective dataset is expected to be established to boost the development of the diet assistant system in real-world environment.

In this paper, we propose a new database, Real-World Multimodal Foodlog (RWMF), which is built based on various authoritative meta-databases using Blur Filter, Coverage Filter, Concept Filter, Learning Vector Quantization and classifier based on Graph Convolutional Network[21]. It not only can provide multimodal data efficiently, but also can be used to build diet assistants that help people learn more about how diet is related to health. One benefit of RWMF is that it includes data that captured in the real-world environment, rather than in a controlled laboratory environment.

The current RWMF database contains 7500 multimodal pairs, each pair includes a food image paired with a line of personal biometrics data (e.g., Real-Time Glucose and Heartbeat) and a textual description of food paired with a line of food nutrition data (e.g., Glycemic Index and Carbohydrate Content). The image labels and the ground truth of health categories (overweight, healthy and other) are also released. The example of RWMF is illustrated in Fig. 3.

The contributions of this paper can be summarized as follows:

Firstly, we introduce the real-world data of other modalities, such as the textual description of food health notification and personal biometrics data into our dataset. These data can contribute to the multimodal foodlog analysis. To the best of our knowledge, it is the first food-related database that includes such real-world multimodal data.

Secondly, to deal with the difficulty of labeling a large database manually, we use the multi-label graph convolutional network (ML-GCN) [24] to help label the images effectively. In order to obtain a precise multi-label classifier, we extend the current large-scale Food-101 dataset by merging it with the categories in USDA to improve its usability and then train ML-GCN on it. At last, the filtered food images are obtained using the automatic annotations and manually re-check.

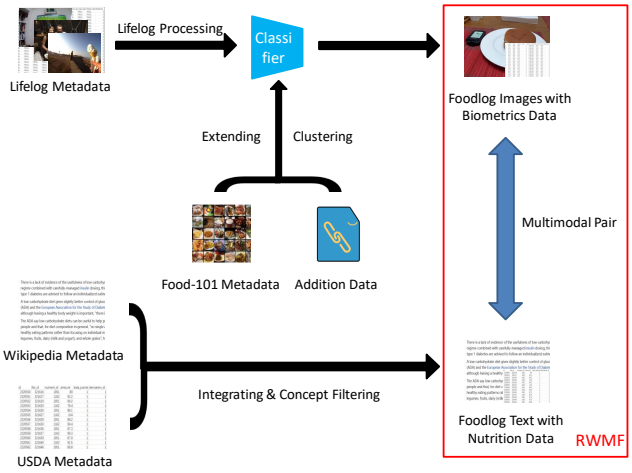

Fig. 2. The pipeline of establishing RWMF.

Thirdly, we evaluate the proposed RWMF database from different perspectives. The result indicates that our database is capable of evaluating the performance of cross-modal retrieval as well as food classification. The fusion of multimodal data can help improve the accuracy of image classification and cross-modal retrieval. In the meantime, this database is still challenging and waiting for further exploration.

The rest of the paper is organized as follows. In Section 2, we introduce the framework of setting up database. And in Section 3 we give the details about processing procedures. Then we present experiments of evaluation in Section 4. Eventually, this paper closes with the conclusion and perspective in Section 5.

II. Framework

RWMF is the new dataset that we build based on various metadata sets, which are lifelog [2], Wikipedia, USDA[1], and Food-101 [20]. We establish dataset by data mining, filtering, and clustering to obtain a diet dataset for better evaluating cross-modal retrieval on real-world foodlog scenorio. And this database can be valuable for futher research on food analysis and personal diet assistant.

The framework is presented in Fig. 2. We first establish the image set by processing the lifelog metadata set. Then we extend the original Food-101 dataset to train an optimization GCN-based food classifier on Food-101-extended. We label our food images by using the classifier that trained on Food-101-extended. And we obtain the food description from Wikipedia using a concept detector and extract food nutrition data from USDA. Finally, the database is formed into multimodal pairs by matching the images with texts using tagged labels. It includes 7500 multimodal pairs, and each pair contains food images, personal biometrics data, textual food descriptions, and food nutrition data.

[1]https://ndb.nal.usda.gov/ndb/search/list

III. Procedures

A. Foodlog Image Set

We carefully select 200,000 pairs of lifelog data in lifelog dataset [2] as the metadata for the foodlog image set. These data are captured in users' daily lives, which include real-life images taken by wearable cameras and the real-time biometrics data tracked by smart bracelets. Each lifelog image is paired with a line of personal biometrics data, including Heartbeat, Calories Consumption and Glucose. As many of these data are related to food and diet, we use the lifelog data as the basic source of food images to constitute the real-world foodlog image datasets that are not only related to life and diet but also personal health information.

However, the lifelog data are collected all day around by users with no limitation given [2], so the quality of these data cannot be guaranteed. Thus, the metadata needs to be processed by the following filters so that high-quality food images can be obtained from the metadata set.

1) Blur Filter: Firstly, there are many futile images such as blur images. The existence of these data would produce an effect on the effectiveness of the whole pipeline. Thus, we use two filters to eliminate the blur images. The first filter is a Laplacian filter with a 3x3 kernel that calculates the blur as the variance of the convolution result. In a normal image, the boundary is clear so that the variance is large. In contrast, the boundary information contained in the blurred image is very sparse, so the variance will be smaller. Therefore, we filter the images that have the variance lower than 30. Then a Fast Fourier Transform is applied to images. Once this step is completed, the average value in the transformed image is obtained and then scaled according to the size of the images to compensate for the tearing effect. The average value is then used as a threshold between the image with the larger value representing the focused image and the lower value representing the blurred image.

2) Coverage Filter: To detect if an image is covered by something or facing the ceiling or wall, we convert images into the binary matrices and use maximum connected area detectors to calculate the proportion of subjects. Then we remove the images that have a subject's size over 90% of the whole area in the binary matrices.

After passing the Blur filter and Coverage Filter, the remaining 120,000 images are clear enough. Then we use the concept filter to extract the food images with paired biometrics data in lifelog metadata set and get pairs of images and biometrics data that are related to dietary scenes.

3) Concept Filter: We first use a pre-trained ResNet-101 [23] to extract the concepts of images, then we maintain images related to food topics and remove the rest. Then we select samples from Food-101-Extended and MSCOCO [22] to arrange a two-classes training set which include the food images and others. After that, we fine-tune a two-class classifier on this two-class training set. And we use this fine-tuned classifier to detect whether an image is a food or not. If not, we abandon it. In the above two procedures, all of the biometrics data that matched the specific images are packed together to form the basic biometrics data set. And we finally get 5000 food images paired with the personal biometrics data.

B. Personal Biometrics Data Set

Because the metadata has many drawbacks such as missing of vital instant blood glucose data in some particular moments, it can interfere with the performance of the whole multimodal database. We need to process and rearrange the additional personal biometrics data entirely.

Since the personal biometrics data that paired with images are captured every second, we fulfill the blank data following the principle of continuity. To be more specific, for a few blanks in the biometrics data, such as the Null data in glucose attributes, we fill these gaps with the data generated from context. For example, a glucose level is measured at 7 mmol/L at 15:20:30 and 8 mmol/L at 15:20:32. Then we can determine that the missing glucose level at 15:20:31 is 7.5 mmol/L based on the deduction that the blood glucose cannot increase abruptly in such a short time.

In addition, we create a new attribute called average glucose in the biometrics data set. Since the original biometrics data contains two different kinds of glucose data, one is historical glucose level and the other is instant glucose level. We find it hard to match the historical glucose corresponding to previous moments. The reason is that the historical data referred to different time intervals. So, we make the average of instant glucose and historical glucose in each line as the buffer for the biometrics data set. This new set of glucose is created to reduce the noise caused by unmatched two sets of glucose and these three sets of glucose are used together in the personal biometrics data set.

C. Foodlog Labels

Due to the huge amount of food images, it is difficult to label all those images manually. We are supposed to label our dataset using an effective classifier that trained on an original food dataset. Food-101 [20] is chosen because of its huge amounts of images and the abundant categories of food. It includes 101 types of food and each type has about 1000 food images.

However, it still has several drawbacks. Firstly, although Food-101 has 101,000 images, it is not big enough to cover all the features of the food item, the coverage of its categories is too narrow to detect all the images that show in lifelog. This means that the categories of Food-101 must be extended. Secondly, the amount of food images in each class is not enough to train a precise classifier. If we want to train a classifier that could recognize the real-world food images as precise as possible, the amount of its images needs to be extended. Lastly, the images in this database contain not only food images but also noises such as the images of chairs or sticks, as this dataset is downloaded from a social network site and classified roughly by random forests. The dataset has not been cleaned carefully and there are also many images

Manuscript 440 submitted to 2020 25th International Conference
on Pattern Recognition. Received April 15, 2020.

divided into wrong classes. These noises must be eliminated to improve the performance of training.

To solve the first problem, we extend the Food-101 with multi-labels. Then, we replenish the images with google API to deal with the second disadvantage. After the replenishment, the number of images becomes 180,000. Then we implement a practical cluster method, Learning Vector Quantization (LVQ), to clean the data and filter the noises by clustering to solve the last problem. The details of extending and clustering are explained in the following.

1) Food-101 Extending: To extend the original Food-101 dataset. We merge it with the data crawled from the internet at first. And then, we replace 101 labels with more classes by analyzing the category relationship of food. For example, an image of filet steak can be also labeled by several other classes such as animal protein and meat. All of these categories are selected from the USDA database to ensure the authority of classification. We choose the classes in USDA because it is also a metadata set we intend to employ to get the nutrition data of food. The extended dataset, which is named Food-101-Extended, includes 150 classes of the 9 high-level classes (such as grain, meat, non-food and etc.), 40 middle-level classes (such as bakery products, fresh fish and etc.) and 101 detailed classes (such as garlic bread, sashimi and etc.).

2) Wiping Off Noise: In our framework, we choose a cluster to get rid of noises in the metadata. The LVQ cluster is used to clean the data and filter the noises. We first extract the feature vectors by pre-trained ResNet-101 and use the cluster to calculate the distance between each vector and the class sample point. Eventually, we rearrange labels and remove the images whose feature distance is much far from the sample point. By implementing clusters and removing the irrelevant images based on the Euclidean distance, the images are filtered into about 110,000 in Food-101-Extended.

3) Labeling: Finally, we train a multi-label classifier using Multi-Label Graph Convolutional Network (ML-GCN) proposed in [24] on Food-101-Extended and use it to label our foodlog images set. The ML-GCN trained on Food-101-Extended, which reaches 95% accuracy on the test set, has the ability to recognize the food and give descriptive labels to food images. At last, each image has a various number of labels. For example, an image of filet steak could be labeled as steak, animal protein, and meat. And a non-food image can only correspond to one label (other). We recheck the labels again with the respect to confidence scores of classification and obtain a labeled foodlog image set. Such labels can help us match the multimodal pairs eventually.

D. Foodlog Text Set

We choose a wide variety of paragraphs in Wikipedia and use them as the metadata set of text modality in our database.

Since the data must be related to food and personal health, a filter operation should be done. a concept detector is trained based on the pre-trained BERT [25]. Then it is used to extract the concept of paragraphs and detect whether it is related to food and health. And we maintain the useful texts which

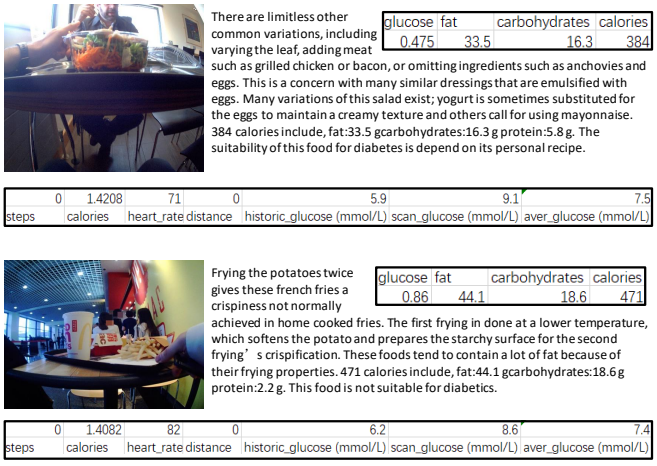

Fig. 3. The examples of RWMF. The bar on the bottom presents the personal biometrics data that paired with the image. And the bar on the top right shows the food nutrition data paired with the text.

are related to appointed topics and remove the other texts. We double-check the texts manually to ensure the quality of data. Afterward, we get a number of paragraphs that are related to dietary health. We label these texts with the chosen classes in USDA and Food-101 manually. Thus, this corpus includes food descriptions and multi-category labels. At last, we integrate the texts with the food nutrition data to get the sets of two description modalities, food description texts and food nutrition data.

E. Foodlog Nutrition Data Set

We preprocess the public dataset USDA to integrate the useful data into our database. Food nutrition data, such as glucose, fat, carbohydrate, and calories are obtained in USDA because these indicators are proximate factors related to dietary health. The measuring unit of glucose is Glycemic Index. The units of fat, carbohydrate and protein are percentage (fat + carbohydrate + protein $\approx$ 1), the unit of calories is kilocalorie/100g. We pair these indicators with 150 food classes as we mentioned before. We further analyze several authoritative healthcare websites as well as get suggestions from the medical experts to acquire the key information of diet nutrition. Then we classify every label into overweight, healthy or other utilizing health information mentioned above. The overweight category includes the food that is unhealthy and must be forbidden from those who want to lose their weight and cannot consume high-energy food. The healthy category contains the food that is healthy for people to eat under normal circumstances. The other category refers to the food that is not easily judged, and the health factors of these foods are basically based on their optional ingredients. Finally, the overweight category contains 90 unhealthy food classes, the healthy category contains 38 healthy food classes and the left 23 classes belong to the other.

Manuscript 440 submitted to 2020 25th International Conference
on Pattern Recognition. Received April 15, 2020.

F. Multimodal Pairing

After all the above procedures, the five kinds of sets are prepared completely, including foodlog image set paired with personal biometrics data set, foodlog text set paired with food nutrition data set, and health categories information. So, we can establish RWMF by using these five sets of data finally.

We affiliate the two description sets with two image sets by matching the picture to text to shape the final form of the RWMF. We utilize the labels of images and texts to help match the food images with food descriptions. A foodlog image and a foodlog text that share the same label can be paired together. In this case, one image can refer to several different descriptions and one text can correspond to various images as well. And we double-check those pairs manually. We call the matched data multimodal pairs because each pair has four different kinds of data totally.

After the procedures of processing, the RWMF has 7500 multimodal pairs. Each multimodal pair contains a food image paired with a line of personal biometrics data and a textual description of food paired with a line of food nutrition data, as the example presented in Fig. 3. Furthermore, we divide the whole RWMF into 3500 overweight food images, 2500 healthy food images, and 1500 other food images via the class labels.

IV. Evaluation

The main purposes of evaluation on the RWMF database are two perspectives. First, we estimate the difficulty level of food classification on RWMF. Second, we want to assess the performance of RWMF database by performing cross-modal retrieval methods.

A. Evaluation Metrics

Accuracy (Acc), average per-class precision (CP), average per-class recall (CR), average per-class F1 measure (CF1) are used as the evaluation metrics of food detection. And the mean average precision (mAP) is utilized in the retrieval evaluation. The F1 measure is an overall evaluation of precision and recall in each category, which is defined as:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{1}$$

We use the average of mAP in image query and text query as the final result. The mAP is the average of the AP of all categories where the AP is defined as:

$$AP = \frac{1}{L} \sum_{i=1}^{R} Pre(i) \cdot Rel(i) \tag{2}$$

where L is the number of relevant results in the retrieval set, and $Pre(i)$ is the percentage of relevant results in the top i retrieved item. And if the item in the prediction ranking i is relevant to query, then $Rel(i) = 1$, otherwise $Rel(i) = 0$. R represents the number of retrieved texts to be examined. If no otherwise specified, we refer to $R = All$ in this paper.

TABLE I
The Food Classification Evaluation Results Using Different Methods

Dataset	Methods	Acc	CP	CR	CF1
RWMF	SVM	83.97	0.75	0.71	0.73
	SVM-fusion	**85.21**	**0.78**	**0.74**	**0.76**
	VGG [26]	76.84	0.45	0.58	0.51
	ResNet-50 [23]	80.26	0.65	0.66	0.65
	Inception-v3 [27]	**82.13**	**0.70**	**0.73**	**0.71**
Food-101	ResNet-50 [23]	**90.64**	**0.85**	**0.86**	**0.86**

B. Food Classification Evaluation

It is expected that an effective classifier can be trained on RWMF. We are supposed to evaluate the performance of classifiers that trained on RWMF. In evaluation, we divided training set and test set on a scale of ten to one. For this evaluation task, we use the food images as input and use three health categories as output. We present the classification results in Table I. In addition, we present the image classification result on Food-101 to provide a reference.

We can see from the table that the CF1 of all methods on RWMF are lower than 0.8, and the CF1 for ResNet-50 on RWMF (0.65) is lower than that on Food-101 (0.86). That is to say, the comparison of results show that the RWMF is quite challenging for food classification, which is more challenging than Food-101. One explanation is that the RWMF contains real-life food images instead of images captured in laboratory environments, which increases the difficulty of classification. In addition, since this database has multimodal data, we conduct a control experiment using the classical SVM algorithm as the benchmark to better quantify the RWMF. We fuse the image feature vectors (extracted by pre-trained ResNet) with personal biometrics data and present the results in Table I, where SVM-fusion means fusing the image with biometrics data as input. After the fusion of biometrics data, the CF1 of classification rises to 0.76 from 0.73. Thus, we can conclude that with the help of biometrics data, it is easier for the model to classify food.

C. Cross-modal Retrieval Evaluation

We randomly choose 6500 multimodal pairs in RWMF as the training set and leave 1000 pairs as a test set to evaluate the performance of state-of-the-art cross-modal methods on RWMF. The following three baseline algorithms are evaluated, including the DCCA [11], MNiL [28], DSCMR [29]. In cross-modal retrieval evaluation, we use the images to retrieve corresponding texts and use the text to retrieve corresponding images. The mAP shown in Fig. 4 is the average performance of models in these two retrieval tasks.

Fig. 4 shows the experimental results of these three different state-of-the-art methods in terms of mAP. We can see that all the methods have their mAP lower than 50%, which indicates that foodlog cross-modal retrieval on RWMF is very challenging. On the other hand, the fusion of biometrics data and

Manuscript 440 submitted to 2020 25th International Conference
on Pattern Recognition. Received April 15, 2020.

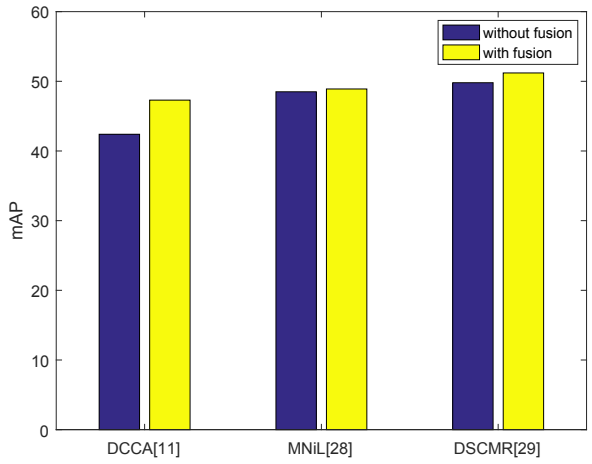

Fig. 4. Comparison of different cross-modal retrieval methods in terms of mAP. With fusion means cross-modal retrieval using the fusion of two kinds of biodata, vice versa.

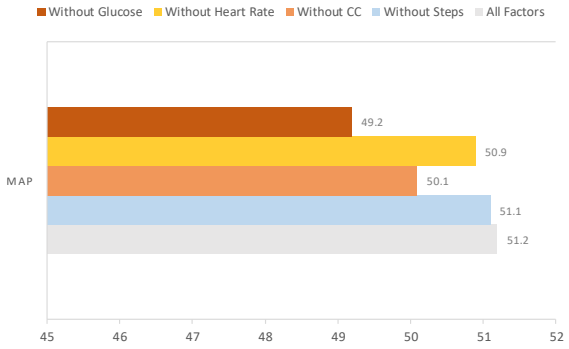

Fig. 5. The performance comparison of mAP when each kind of personal biometrics data is removed. The CC means the calories consumption.

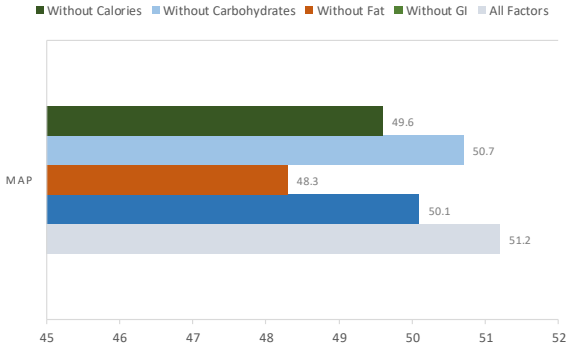

Fig. 6. The performance comparison of mAP when each kind of food nutrition data is removed. The GI means the Glycemic Index.

nutrition physiological data improves retrieval performance significantly.

Furthermore, we analyze the influence of personal biometrics data and food nutrition data by ablation experiments. We remove a certain kind of data from training set and test set to futher compare how these data affect the accuracy of foodlog cross-modal retrieval. We compare the results of mAP in these ablation experiments. The results are shown in Fig. 5 and Fig. 6, where each bar represents the retrieval performance when we remove this kind of data, and the bar of all factors represents the retrieval result when all factors are used in data fusion. Fig. 5 shows the results of mAP when each kind of personal biometrics data is removed. It shows that glucose has the greatest positive impact on retrieval performance, which means the instant glucose is the most helpful factor in personal foodlog retrieval. Fig. 6 shows the results of mAP when each kind of food nutrition data is removed. We can see that the retrieval performance drops when the GI or fat is removed from dataset. As the fat composition has the biggest influence on the retrieval information, fat content of food can be the most related factor to dietary health in RWMF, which can help the model decide whether a food is unhealthy and retrieve the most precise information for users.

Experimental results indicate that modeling the cross-modal retrieval mechanism in real-world scenes is a new issue with many unsolved problems. At present most of the state-of-the-art cross-modal retrieval models barely adapt to complex real-world environments such as the foodlog scenario. So, furthering explorations for multimodal models based on real-world data must be made for the development of cross-modal retrieval.

V. Conclusion

This paper introduces the Real-World Multimodal Foodlog (RWMF) database, which contains 7500 multimodal pairs of 150 food classes. To the best of our knowledge, this dataset is the first set of foodlog images with such multi-modal data. We also explain the basic facts of two evaluation experiments. Several classical and state-of-the-art algorithms have been evaluated on RWMF. Experimental results show that the RWMF is practical enough for evaluating image classification and cross-modal retrieval in a real-life scenario. In the meantime, it is still a quite challenging database. As the data used in this database is multimodal, including image, text and biometrics data, we provide researchers who are working on cross-modal retrieval and food-analyzing a suitable dataset to evaluate the performance of cross-modal retrieval methods. We are supposed to further extend this database with more multimodal pairs in the future. And we will add more medical details into RWMF that can be used to help with the diet of specific population groups, such as diabetic patients and allergy sufferers.

Acknowledgment

This work is supported by National Key Research and Development Program under Grant No. 2018YFB1305200, and Natural Science Foundation of China under Grant No. U1908210 and 61976192.

REFERENCES

[1] C. Gurrin, A. F. Smeaton, and A. R. Doherty, "Lifelogging: Personal big data," Found. Trends Inf. Retr. vol.8, no.1, pp.1–125, 2014.

[2] D. T. D. Nguyen, L. Zhou, R. Gupta, M. Riegler, and C. Gurrin. "Building a Disclosed Lifelog Dataset: Challenges, Principles and Processes," in CBMI Workshops, 2017.

[3] J. Gemmell, A. Aris and R. Lueder, "Telling Stories with Mylifebits," in ICME, pp. 1536-1539, 2005.

[4] Liting Zhou, Zaher Hinbarji, Duc-Tien Dang-Nguyen and Cathal Gurrin. "LIFER: An Interactive Lifelog Retrieval System," pp. 9-14. in ICMR, 2018.

[5] M. B. Amin, O. Banos, W. A. Khan, H. S. M. Bilal, J. Gong, D. Bui, S. H. Cho, S. Hussain, T. Ali, U. Akhtar, T. Chung, and S. Lee, "On Curating Multimodal Sensory Data for Health and Wellness Platforms," Sensors, vol.16, no.7, 2016.

[6] W. Min, S. Jiang, J. Sang, H. Wang, X. Liu, L. Herranz, "Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration," IEEE Trans. Multimedia, vol. 19, no. 5, pp. 1100-1113, May. 2017.

[7] W. Min, S. Jiang, L. Liu, Y. Rui and R. Jain. "A survey on food computing," ACM Computing Surveys (CSUR), vol. 52, no. 5, pp.1-36, Sept. 2019.

[8] W. Min, S. Jiang and R. Jain. "Food Recommendation: Framework, Existing Solutions and Challenges," IEEE Trans. Multimedia, Dec. 2019.

[9] K. Aizawa and M. Ogawa, "Foodlog: Multimedia tool for healthcare applications," IEEE MultiMedia, vol. 22, no. 2, pp. 4–8, 2015.

[10] N. Rasiwasia, J. Costa Periera, R. Lanckriet, et al. "A new approach to cross-modal multimedia retrieval," in ACM International Conference on Multimedia: 251-260, 2010.

[11] F. Yan, K. Mikolajczyk. "Deep correlation for matching images and text," in IEEE CVPR: 3441-3450. 2015.

[12] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, and A. Torralba, "Learning cross-modal embeddings for cooking recipes and food images," in IEEE CVPR, pp. 3020–3028, 2017.

[13] P. Pouladzadeh, S. Shirmohammadi, and R. Almaghrabi. "Measuring Calorie and Nutrition from Food Image," IEEE Trans. I&M, vol.63, no.8, pp. 1947–1956, 2014.

[14] S. Jiang, W. Min, L. Liu and Z. Luo, "Multi-Scale Multi-View Deep Feature Aggregation for Food Classification," IEEE Trans. Image Process., vol. 29, pp. 265-276, 2019.

[15] W. Min, B. K. Bao, S. Mei, Y. Zhu, Y. Rui and S. Jiang, "You are what you eat: Exploring rich recipe information for cross-region food analysis," IEEE Trans. Multimedia., vol. 20, no. 4, pp.950-964, Apr. 2017.

[16] A. Arwan, M. Sidiq, B. Priyambadha, H. Kristianto and R. Sarno, "Ontology and semantic matching for diabetic food recommendations," in ICITEE, pp. 170-175, 2013.

[17] K. Grifantini, "Knowing What You Eat: Researchers Are Looking for Ways to Help People Cope with Food Allergies," in IEEE Pulse, vol. 7, no. 5, pp. 31-34, Sept.-Oct. 2016.

[18] R. Xu, L. Herranz, S. Jiang, S. Wang, X. Song and R. Jain, "Geolocalized Modeling for Dish Recognition," in IEEE Trans. on Multimedia, vol. 17, no. 8, pp. 1187-1199, 2015.

[19] B. Tusor, G. Simon-Nagy, J. T. Tóth and A. R. Várkonyi-Kóczy, "Personalized dietary assistant — An intelligent space application," in Proc. IEEE INES, pp. 27-32, 2017.

[20] L. Bossard, M. Guillaumin, L. V. Gool. "Food-101 – Mining Discriminative Components with Random Forests," in ECCV, 2014.

[21] T. N. Kipf, M. Welling. "Semi-Supervised Classification with Graph Convolutional Networks," in ICLR, 2017.

[22] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. Zitnick. "Microsoft COCO: Common Objects in Context," in ECCV, 2014.

[23] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition," in IEEE CVPR, pages 770–778, 2016.

[24] Z. M. Chen, X. S. Wei, P. Wang, et al. "Multi-Label Image Recognition with Graph Convolutional Networks," in IEEE CVPR, 2019.

[25] J. Devlin, M. W. Chang, K. Lee, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018.

[26] K. Simonyan, A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition," Computer Science, 2014.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in IEEE CVPR, pp. 2818-2826, 2016.

[28] L. Zhang, B. Ma and G. Li. "Multi-Networks Joint Learning for Large-Scale Cross-Modal Retrieval," in ACM MM, 907-915. 2017.

[29] L. Zhen, et al. "Deep Supervised Cross-Modal Retrieval," in IEEE CVPR, 2019.